JOURNAL OF TELECOMMUNICATIONS AND INFORMATION TECHNOLOGY

4/2010

**Robust and Accurate Iris Segmentation Algorithm for Color and Noisy Eye Images**
*P. Strzelczyk* — Paper — 5

**Iris Recognition System Based on Zak-Gabor Wavelet Packets**
*A. Czajka and A. Pacut* — Paper — 10

**Image Preprocessing for Illumination Invariant Face Verification**
*M. Leszczyński* — Paper — 19

**Face Tracking and Recognition with the Use of Particle-Filtered Local Features**
*Ł. A. Stasiak and A. Pacut* — Paper — 26

**Rule Based Speech Signal Segmentation**
*M. Greibus and L. Telksnys* — Invited paper — 37

**Speech Segmentation Algorithm Based on an Analysis of the Normalized Power Spectral Density**
*D. Pekar and S. Tsikhanenka* — Invited paper — 44

**Relaxing the WDO Assumption in Blind Extraction of Speakers from Speech Mixtures**
*W. Kasprzak, N. Ding, and N. Hamada* — Invited paper — 50

**Hidden Signature for DTW Signature Verification in Authorizing Payment Transactions**
*J. Putz-Leszczyńska and M. Kudelski* — Paper — 59

**Simulation Model of Biometric Authentication Using Multiagent Approach**
*A. Kapczyński and T. Owczarek* — Paper — 68

(*Contents Continued on Back Cover*)

# *Editorial Board*


Editor-in Chief: ..... ***Paweł Szczepański***

Associate Editors: ..... ***Krzysztof Borzycki***
***Marek Jaworski***

Managing Editor: ..... ***Maria Łopuszniak***

Technical Editor: ..... ***Ewa Kapuściarek***


# *Editorial Advisory Board*


Chairman: ..... ***Andrzej Jajszczyk***
***Marek Amanowicz***
***Daniel Bem***
***Wojciech Burakowski***
***Andrzej Dąbrowski***
***Andrzej Hildebrandt***
***Witold Hołubowicz***
***Andrzej Jakubowski***
***Alina Karwowska-Lamparska***
***Marian Kowalewski***
***Andrzej Kowalski***
***Józef Lubacz***
***Tadeusz Łuba***
***Krzysztof Malinowski***
***Marian Marciniak***
***Józef Modelski***
***Ewa Orłowska***
***Andrzej Pach***
***Zdzisław Papir***
***Michał Pióro***
***Janusz Stokłosa***
***Wiesław Traczyk***
***Andrzej P. Wierzbicki***
***Tadeusz Więckowski***
***Józef Woźniak***
***Tadeusz A. Wysocki***
***Jan Zabrodzki***
***Andrzej Zieliński***



ISSN 1509-4553 on-line: ISSN 1899-8852


© Copyright by National Institute of Telecommunications
Warsaw 2010



Circulation: 300 copies

Sowa - Druk na życzenie, www.sowadruk.pl, tel. 022 431-81-40

JOURNAL OF TELECOMMUNICATIONS AND INFORMATION TECHNOLOGY

# *Preface*

This issue of the *Journal of Telecommunications and Information Technology* contains papers devoted to various aspects of biometrical authentication systems, and in particular, to iris, face, speech, and on-line signature biometrics, multi-biometrics, and biometric template formation. The papers in this issue deal also with certain issues of crosstalk propagation in silicon substrates and with some mobile networks problems.

The first two papers of this issue deal with the iris biometrics. Przemysław Strzelczyk in his paper on *Robust and Accurate Iris Segmentation Algorithm for Color and Noisy Eye Images* addresses the segmentation problem of iris images captured in uncontrolled environments. His novel coarse-to-fine segmentation algorithm first uses modified Hough transform to roughly localize possible iris and pupil boundaries, initially approximating them by circles. Voting mechanisms are then applied to select the candidate iris regions, and the detailed iris boundary is approximated by a spline curve that minimize a boundary energy function. This algorithm is suitable for both monochrome and color eye images.

The paper on *Iris Recognition System Based on Zak-Gabor Wavelet Packets*, by Adam Czajka and Andrzej Pacut proposes a new iris coding method based on Zak-Gabor wavelet packet transform. The essential component of the proposed methodology is an effective adaptation of the transformation parameters that makes the coding sensitive to the frequencies characterizing the individual eye. The between-to-within class ratio of weakly correlated Zak-Gabor transformation coefficients is calculated, which allows for selection of the frequencies most suitable for iris recognition. The method provides a possibility to embed the biometric replay attack prevention methodology into the coding. The entire system includes a dedicated hardware, and its properties were evaluated using the database of 720 images. The methodology was also tested in remote access scenario, biometric smart cart development, and an iris recognition system with aliveness detection.

The next two papers discuss the face biometrics. In his *Image Preprocessing for Illumination Invariant Face Verification*, Mariusz Leszczyński discusses the crucial problem of the influence of face illumination conditions. Fourteen normalization algorithms based on histogram normalization, illumination properties and the human perception theory were compared using three verification methods. The best results were achieved using human perception related MSQ algorithm, with 38% less verification errors compared to the same DLDA discriminant method using un-preprocessed images. Very promising seems to be a combination of both histogram approach and photometric normalization.

In *Face Tracking and Recognition with the Use of Particle-Filtered Local Features* Łukasz A. Stasiak and Andrzej Pacut propose a particle filtering framework for parallel face tracking and recognition from video sequences. Randomized, particle filtering-driven local face features form a base for a cumulation mechanism of classification decisions. Application of the proposed mechanisms makes the framework resistant to brief visual distortions, such as occlusions, head rotations or face expressions. The framework can be implemented to operate in real time on a PC, and is tunable to various application requirements (security level, hardware constraints), achieving a high performance even on low resolution video frames.

Three next papers are devoted to speech analysis and speech biometrics. An invited paper of Mindaugas Greibus and Laimutis Telksnys on *Rule Based Speech Signal Segmentation* presents a solutions to an automated speech segmentation problem. Segmentation algorithms based on energetic threshold may fail in noisy environments. The Authors show that a rule based postprocessing of segments can give more stable result. Offline, online and extrema types of rules were investigated. The proposed extrema-type segmentation algorithms is enhanced by a rule base to extract higher energy level segments from noise. The algorithm performs about 8% better than threshold or dynamic threshold algorithms, yet its numerical complexity is low.

In their invited paper on *Speech Segmentation Algorithm Based on an Analysis of the Normalized Power Spectral Density*, Dzmitry Pekar and Siarhei Tsikhanenka present a new approach to speaker independent phoneme detection, which does not require any prior information about the signal, acoustic models of phonemes, or speaker's individual characteristics. In their approach, the distance between normalized power spectral densities is measured in adjacent, short-time segments and verified using velocity of changes in short-time signal energy. The proposed algorithm enables to reveal a phoneme structure of pronounced speech with about 90% probability.

The invited paper on *Relaxing the WDO Assumption in Blind Extraction of Speakers from Speech Mixtures* is presented by Włodzimierz Kasprzak, Ning Ding, and Nozomu Hamada. The time-frequency masking approach in blind speech extraction includes the spectrogram masking used to reconstruct the sources. Usually the binary mask is generated under strong W-disjoint orthogonality (WDO) assumption which is often violated in practice, thus leading to weak quality of reconstructed sources. The Authors propose WDO to be relaxed by allowing some frequency bins to be shared by both sources. Along with a detection of instantaneous fundamental frequencies, a harmonic structure of speech is explored to support creation of the mask. The proposed method is proved to be effective and reliable in both simulations and in experiments with genuinely acquired mixtures.

The signature biometrics is discussed by Joanna Putz-Leszczyńska and Michał Kudelski in their paper on *Hidden Signature for DTW Signature Verification in Authorizing Payment Transactions*. The authors propose to form the on-line signature template using an approach that combines the dynamic time warping (DTW) technique with the least squares estimation. The verification algorithm employs a standardized error signals between the signature in question and the model. The method can also be used in mobile or embedded systems. The Authors describe a real-world implementation of the algorithm, namely the complete system designed for biometric payment transactions authorization. The authorization is performed directly at a point of sale.

In the paper on *Simulation Models of Biometric Authentication Systems Using Multi-Agent Approach*, Adrian Kapczyński and Tomasz Owczarek present an application of multiagent approach to biometric authentication systems. After presenting the state of the art, four simulation models of biometric authentication environments are examined. Those models are appropriately parameterized and explored under various conditions with the use of programmable modeling environment for simulating natural and social phenomena.

The next two papers discuss the problem of creation of biometric templates. Since the variability of biometric data obtained for the same person is an intrinsic property of every biometric system, the problem of finding "a best representative" of the enrollment set, called the template, is present since the beginning of biometrics. This problem is solved differently for different biometric types, yet usually the template somehow averages the collected data samples. Andrzej Pacut in his paper on *Probabilistic Issues in Biometric Template Design* discuss various basic possible meanings of the term "the best representative", basing his derivations on minimization of various average dissimilarity indexes in $L_1$ or $L_2$ spaces. He solves the underlying minimization problems and discuss some properties of the resulting

templates. In particular, for i.i.d. samples of independent component vectors, a probability is derived how likely it is for a vector of medians to belong to the sample.

Marcin Chochowski in his paper on *The Template Selection in Biometric Systems Based on Binary Iris Codes* considers the general problem of a small enrolment sample size as compared to sample vector lengths, using as an important example the iris biometrics. He shows by simulations that the averaging is justified also in the case of iris template creation. This is an important fact, which can significantly improve the performance of biometric template protection methods. In particular, binary iris coding algorithms using the majority code as the template lead to better results.

The next paper on *Impact of Crosstalk into High Resistivity Silicon Substrate on the RF Performance of SOI MOSFET*, Khaled Ben Ali, César Roda Neve, Ali Gharsallah, and Jean-Pierre Raskin discuss crosstalk propagation through silicon substrate, which is a serious limiting factor of the performance of the RF devices and circuits. They analyze experimentally the substrate crosstalk in high resistivity silicon substrate, and discuss its impact on the RF behavior of silicon-on-insulator (SOI) MOS transistors. The authors demonstrate an efficiency in the reduction of the sideband noise tones of the solution that employs a trap-rich polysilicon layer located underneath the buried oxide (BOX) of the high resistivity (HR) SOI wafer. They also propose an equivalent circuit to model and analyze the generation of these sideband noise tones.

Ayodeji James Bamisaye and Michael O. Kolawole in their paper on *Capacity and Quality Optimization of CDMA Networks* discuss coverage and capacity issues of the process of planning cellular mobile networks of third generation (3G) that aims to allow the maximum number of users sending and receiving signal of adequate strength. The paper describes conceptual expressions required for network coverage and capacity optimization analysis, examines service quality issues, and presents some practical solutions to CDMA networks.

The paper of Sunilkumar S. Manvi, Lokesh B. Bhajantri, and Vittalkumar K. Vagga on *Routing Misbehavior Detection in MANETs Using 2ACK* proposes a routing misbehavior detection in mobile ad-hoc networks (MANETs) using 2ACK scheme. The MANETs routing protocols are designed on the assumption that all participating nodes are fully cooperative. This may not be fulfilled due to the open structure and scarcely available battery-based energy. As there is no retransmission of packets once it is sent in MANETs, care must be taken not to loose packets. The authors analyze and evaluate the 2ACK technique to detect and mitigate the effect of a routing misbehavior and propose to embed certain security aspects to the 2ACK scheme to check confidentiality of the message. They use a verification of the original hash code with the hash code generated at the destination. If 2ACK is not received within the wait time or the hash code of the message is changed then the node to next hop link of sender is declared as the misbehaving link. The performance of the proposed setup is checked by simulations.

Andrzej Pacut
Guest Editor

*Institute of Control and Computation Engineering*
*Warsaw University of Technology*

*Biometric Laboratories*
*Research and Academic Computer Network NASK*

Paper

# Robust and Accurate Iris Segmentation Algorithm for Color and Noisy Eye Images

Przemysław Strzelczyk

**Abstract—Efficient and robust segmentation of iris images captured in the uncontrolled environments is one of the challenges of non-cooperative iris recognition systems. We address this problem by proposing a novel iris segmentation algorithm, which is suitable both for monochrome and color eye images. The method presented use modified Hough transform to roughly localize the possible iris and pupil boundaries, approximating them by circles. A voting mechanisms is applied to select a candidate iris regions. The detailed iris boundary is approximated by the spline curve. Its shape is determined by minimizing introduced boundary energy function. The described algorithm was submitted to the NICE.I iris image segmentation contest, when it was ranked 11th and 10th out of total 97.**

***Keywords—biometrics, image segmentation, iris recognition.***

## 1. Introduction

One of the main problems with the iris recognition technology is a difficulty of acquiring good quality iris images in uncontrolled environments. In out-of-laboratory scenarios the probability of capturing undisturbed eye images not affected by noise, blurring and different kinds of occlusions is very low. Therefore the overall performance of iris-based biometric systems is highly influenced by the efficiency of segmentation algorithm. This paper introduces a new iris segmentation method that can be used to extract iris regions of both color and monochrome images. The simplified Hough transform for circle-like object is introduced which significantly reduces the computation time in the coarse localization of the pupil and the iris. The active contours based on splines are used to accurately estimate the iris boundaries and separate the iris from the eyelids and the eyelashes. A discrete optimization technique based on dynamical programming is used to robustly fit the active contours. The proposed method has been tested using color images from UBIRIS.v2 database from NICE.I (Noisy Iris Challenge Evaluation – Part I) competition [1], [2].

## 2. Related Work

The segmentation of the iris image in color and in the infrared light are slightly different. Iris recognition systems usually use monochrome images captured in infrared light, because for that light wavelengths the performance of the system is the highest. However color images contain more information which is usable for localization and segmentation. The NICE.I competition participants introduced segmentation algorithms which work on color images captured in visible light [3]–[10]. Tieniu Tan *et al.* in his winning algorithm use integrodifferential constellation method [3] which accelerates the integrodifferential operator used by Daugman [11], [12]. They also introduce the curvature model for eyelids detection and histogram based reflection removal. Luengo-Oroz, Faure and Angulo applied mathematical morphology for polar image filtering and approximated the iris boundary by looking for the shortest paths in generalized gray-level distances [4]. Labati and Scotti start with integrodifferential approach to coarsely localize the iris boundaries, then search for continuous boundary segments in polar image and interpolate the missing fragments using low-pass filtering [10]. They use the Gabor filters to detect and remove eyelashes. Dae Sik Jeong *et al.* detect pupil and iris simultaneously using two circular edge detectors [9]. If the number of specular reflection on the image is too low they use AdaBoost eye detection technique to check if it is possible that the eye is closed. Then they use adaptive convolution kernels to detect separable eyelashes and parabolic Hough transform to detect eyelids. Finally they apply color-bases iris area detection. Peihua Li *et al.* introduced limbic boundary localization algorithm combining Hough transform, K-means clustering and skin color modeling [8]. They detect eyelids using parabolic integro-differential operator combined with RANSAC-like techniques. Other autors usually use integrodifferential operator for iris and pupil localization, the parabolic Hough transform for eyelids removal and the local histogram thresholding for the eyelashes detection.

## 3. Color of an Eye

The light rays which travel through the pupil usually reflects many times before they come out of the eye interior. Most of the light is absorbed, so only a small amount of it is visible on the eye image making pupil a very dark object. However the fundus, which is the interior surface of the eye, and includes: retina, optic disc, macula, fovea and posterior pole, has a red tint for the human. Therefore under very high illumination, pupil may be observable as red, which we can experience in the red-eye effect. The human iris color ranges from brown to green, blue, gray and hazel. It depends on many factors as texture, pigmentation, fibrous tissue and blood vessels. The melanin, which is

the main iris pigment, is responsible for the colors which vary from yellowish-brown through dark brown to black depending on the concentration and the location of pigments. Other colors are observed due to selective reflection and absorption of biological molecules which are a part of different iris components (e.g., red color of hemoglobin in blood, colors produced by collagen in tissues, yellowish color of lipochrome). Some colors of iris may also be a result of the light diffraction and scattering (e.g., blue color is an effect of process called Tyndall scattering which has place in turbid layer of the iris). The sclera is the lightest part of the human eye. It contains mainly collagen and elastic fiber and it has usually white color, due to lack of pigment. A very light blue tint may be observable in children, for which the sclera is very thin, and a light yellow tint may be visible in elderly people due to fatty deposits or illness.

Due to the color richness of irises it is difficult to define their characteristic color subspace and distinguish the iris area based only on its color. However information about the color can be used to find the boundary between the iris and the sclera and between the iris and the pupil. If the reflections are removed the pupil area is the darkest part of the image regardless of the color component used. On the other hand, the iris area contains at least one color component which has a significant higher values than for the pupil area. Therefore in order to distinguish the pupil and the iris we generate a monochrome image with intensity values of each pixel equal to maximum intensity values of red (R), green (G) and blue (B) color components of the source eye image $I_{pupil\text{-}iris} = \max(I_R, I_G, I_B)$. The sclera is the lightest part of the eye image. Usually all the color components have very high intensity values in the sclera region. Therefore in offer to find the outer iris boundary we generate second monochrome image with intensity values equal to the minimum of the original image color components $I_{sclera\text{-}iris} = \min(I_R, I_G, I_B)$.

# 4. Segmentation algorithm

We noticed that the differences in intensity between the sclera and the iris are more significant that the differences between the iris and the pupil. Therefore our algorithm starts with the search for the outer iris boundary, and later searches its inner boundary. The pupil location and the size can be easily approximated when the iris size and the iris location is known.

In the first step of the segmentation we remove the highlights using histogram thresholding technique. Afterwards a directional image is computed, which for each image pixel approximate a local gradient of the intensity. Two filters based on directional derivarates of Gaussian are used to estimate the gradient components: horizontal $g_x(x,y)$ and vertical $g_y(x,y)$. The Gaussian-based filters are used in order to reduce the noise and high-frequency disturbances. The resulting array stores the gradient directions $\Theta(x,y) = \arctan(\frac{g_y(x,y)}{g_x(x,y)})$ and the gradient absolute values $\overline{g}(x,y) = \sqrt{g_x(x,y)^2 + g_y(x,y)^2}$. Afterwards a subset $S$ of all points is created, for which the absolute gradient magnitude is greater than a specific threshold. These points are candidate circle perimeter points. Similarly to the standard Hough transform an accumulator array $A(x,y,r)$ is prepared. Its values are initially set to zero. The algorithm updates the accumulator array for each candidate boundary point and for each radius with the absolute values of the gradient in the candidate boundary points as follows:

$$\begin{aligned} \widetilde{x_i} &\leftarrow x_i - r\cos\Theta(x_i,y_i), \\ \widetilde{y_i} &\leftarrow y_i - r\sin\Theta(x_i,y_i), \\ A(\widetilde{x_i},\widetilde{y_i},r) &\leftarrow A(\widetilde{x_i},\widetilde{y_i},r) + \alpha\overline{g}(x_i,y_i). \end{aligned}$$

The array value $A(x,y,r)$ is increased if and only if there is a probability that a candidate point from $S$ lay on the perimeter of the circle in the center in $(x,y)$ and radius $r$. To compensate the inaccuracy in the gradient magnitude and direction estimation, the accumulator array is convolved with 3-dimensional Gaussian filter. The local maxima in the smoothed accumulator array are the candidate circle centers. Next we use an iterated procedure to select $k$ local maxima which corresponds to the $k$ best candidates. In each iteration the global maximum is found by full search. The global maximum coordinates together with corresponding accumulator array value $A(x_j,y_j,r_j)$ are put on the centers candidate list. After that all the values of the array in the stored maximum neighborhood are set to zero to suppress any false center, and the algorithm proceed to the next iteration. When the process is finished a voting mechanism decide if the candidate passes to the next phase. The mechanism compares the accumulator array values of each two consecutive candidates. If the relative differences of the array values between the $i$th and ($i$+1)th maximum is greater that the predefined threshold, all the maxima from $i+1$ to $k$ are discarded. The range of the $x$, $y$ and $r$ parameters depends on the image size and the image capture conditions. Usually the circle center $(x,y)$ is within the image range but not too close to the image border. The iris radius is not greater than $1/3$ of the image height and the iris to pupil size ratio is usually between 1.5 and 3.

After the coarse approximation of iris and pupil center is finished the detailed boundary is modeled with splines. The spline is defined in the polar coordinates system anchored in the circle center. We use the 1st degree periodic spline based on the $K$ control points called later knots. The number of knots $K$ depends on the circle radius $R$ and spacing constant $c$ as follows: $K(r) = 2^{\lceil \log_2 \pi R \rceil - c}$. The spline is defined as a piecewise function, which returns the distance from the circle center based on the angle $\alpha$ and the parameters vectors $\widehat{\alpha}$ and $\widehat{r}$:

$$R(\alpha,\widehat{\alpha},\widehat{r}) = R^*(\alpha,\widehat{r}_k,\widehat{r}_{k+1},\widehat{\alpha}_k,\widehat{\alpha}_{k+1}) \text{ for } \alpha \in <\widehat{\alpha}_k,\widehat{\alpha}_{k+1}),$$

where

$$R^*(\alpha,\widehat{r}_k,\widehat{r}_{k+1},\widehat{\alpha}_k,\widehat{\alpha}_{k+1}) = \frac{1}{\Delta\alpha}\left(\widehat{r}_k(\widehat{\alpha}_{k+1}-\alpha) + \widehat{r}_{k+1}(\alpha-\widehat{\alpha}_k)\right)$$

$$\widehat{r}_0 \equiv \widehat{r}_K$$

The knots angular coordinates are fixed and uniformly distributed: $\widehat{\alpha}_k = k\Delta\alpha$, for $k = 0 \ldots K$, where $\Delta\alpha = \frac{2\pi}{K}$. The second coordinate of each knot $\widehat{r}_k$ is determined by the minimization of the introduced boundary energy function $E$. This energy function takes into account the intensity changes in the direction of radius calculated between the knots, the deviation of the resulting shape from the circle and the length of the boundary. The boundary is defined piecewise so the energy function $E$ can be decomposed into sum of partial energies $E_k$.

$$E(\widehat{\alpha},\widehat{r}) = \sum_{K}^{k=0} E_k(\widehat{\alpha},\widehat{r}),$$

where

$$E_k(\widehat{\alpha},\widehat{r}) = \lambda_1 \frac{\int_{\alpha_{k+1}}^{\alpha_k} -\frac{\partial(I^\diamond * G^\diamond)}{\partial r}(R(\alpha,\widehat{\alpha},\widehat{r}),\alpha)\, d\alpha}{N(\widehat{\alpha},\widehat{r},k)} + \lambda_2 \left| R - \frac{1}{\Delta\alpha}\int_{\alpha_{k+1}}^{\alpha_k} R(\alpha,\widehat{\alpha},\widehat{r})\, d\alpha \right|.$$

The first term of the energy function calculates the blurred directional derivarate of the image intensity along the radius. The calculation is made for the polar image representation. The partial derivarate of the image $I^\diamond(r,\alpha)$ is convolved with the polar Gaussian filter $G^\diamond(r,\alpha)$. The resulting image emphasizes the iris or pupil boundary. The boundary points has higher intensity values in the resulting image and therefore the integral in the first term obtains lowest values for the curves which best fits the pupil and the iris boundaries. The normalizing function $N(\widehat{\alpha},\widehat{r},k)$ is used to compensate different path segment lengths. The second term of the equation introduces the weighted penalty for the deviation of the ring curve from the circle with radius $R$.

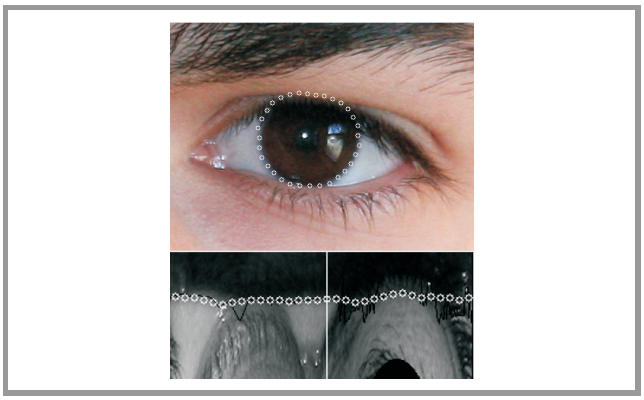


***Fig. 1.*** An eye image with the iris outer boundary approximated by the splines in the cartesian (top) and the polar (bottom) coordinate systems. Small circles indicate the spline knots positions. The white vertical line in the polar image shows the starting point of the spline shape optimization.

The knots second coordinates $r_k$ accept only discrete values, and should not change significantly between two neighboring knots. The differences from one knot to another should be small as the iris boundary is rather smooth. We could add the penalty for roughness to the energy function but we preferred to limit these differences as follows: $|r_k - r_{k+1}| \in \{0,1,\ldots,L-1,L\}$. In that way we obtained a discrete and limited space of the parameters to be optimized $r_k$. As the optimization of the energy function can be easily decomposed the dynamic programming can be used to find the optimal solution. The only problem is that the optimization horizon is periodic, and we need to fix one knot, or in other words we need to find the proper starting point. In our algorithm we used a simple heuristic for that purpose, which finds the stable boundary fragment and selects a point in the middle of it. Figure 1 shows the outer iris boundary approximated by the splines both in the original and the polar image. The small white circles indicate the knots positions.

Usually algorithm localizes more than one possible irise or pupil. To select the proper one we analyze the image intensity values near the shape boundary. A single score based on the t-test is used to select the best candidate. If we denote $X_o$ as the intensity values of the pixels near the boundary but outside the shape and $X_i$ as the intensity values of pixels near the boundary but inside the shape the score can be described as follows:

$$s = \frac{|\overline{X}_o - \overline{X}_i|}{\sqrt{0.5(\mathrm{Var}(X_o) + \mathrm{Var}(X_i))}}.$$

The two voting mechanisms, first based on accumulator array value and the second based on the $s$ successfully eliminate the non-iris and non-pupil image objects.

## 5. Tests and results

The presented algorithm was evaluated in *The Noisy Iris Challenge Evaluation, Part I*, organized by the Soft Computing and Image Analysis Group (SOCIA Lab) of the University of Beira Interior [1], [2]. The aim of the contest was to localize the undisturbed parts of the iris in the color iris images. The two parts of the second version of the UBIRIS database were used as a training set and an evaluation set . The UBIRIS.v2 consists of color eye images with the high level of noise to simulate the less controlled image acquisition conditions. The images include noise factors such as: poor focus, specular reflections, off-angle shots, no-iris images, rotation, motion blur, closed eyes, occlusions by eyelashes, eyelids, glasses and contact lenses. Some sample images from database are shown in Fig. 2

The database contain 500 close-up iris images and corresponding binary masks indicating the position of unoccluded iris area. The binary masks were prepared manually by the SOCIA Lab. The images were captured in visible light using digital camera from the distance of 3 to 7 m with resolution $400 \times 300$ pixels. A sample image with a binary mask from the training set is shown in Fig. 3.

The segmentation algorithms were evaluated using two performance indicators. The first one is the classification error



***Fig. 2.*** Sample images from UBIRIS.v2 database.

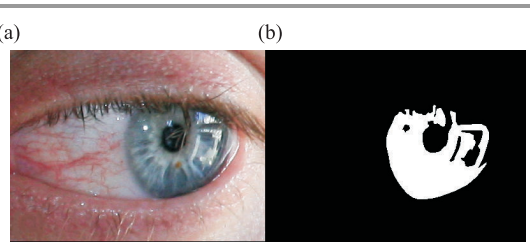



***Fig. 3.*** A sample image from UBIRIS.v2 database (a) with corresponding binary mask indicating iris area (b).

rate $E^*$, which is measured as the average proportion of the corresponding disagreeing pixels between the manually segmented images $M_i(x,y)$ and images segmented by the algorithm $A_i(x,y)$. It is calculated as the average normalized Hamming distance between the binary segmentation masks for all images as follows:

$$E^* = \frac{1}{N}\sum_i E_i^*,$$

$$E_i^* = \frac{1}{WH}\sum_{x,y}\left(M_i(x,y) \oplus A_i(x,y)\right),$$

where $H$ is the image height, $W$ is the image width and $\oplus$ is the binary ExOR operation.

The second performance indicator measures the average type-I and type-II error rates. It tends to compensate the disproportion between the a priori probabilities of the "iris" and "non-iris" pixels in the image. For a single image the second measure is defined as the average value of the the false-positives $FP_i$ and the false-negatives $FN_i$ as follows:

$$E^{\#} = \frac{1}{2N}\sum_i (FP_i + FN_i),$$

$$FP_i = 1 - \frac{\sum_{x,y}\left(M_i(x,y) \wedge A_i(x,y)\right)}{\sum_{x,y} M_i(x,y)},$$

$$FN_i = 1 - \frac{\sum_{x,y}\left(\neg M_i(x,y) \wedge \neg A_i(x,y)\right)}{\sum_{x,y} \neg M_i(x,y)},$$

where $\wedge$ is a binary AND operation (logical conjunction) and "$\neg$" is binary negation.

When the $E^*$ error rate was used our algorithm was ranked 11th out of 97 participants with score 3.41%. The best result obtained in the competition was 1.30% and the mean and the median of the first 10 best algorithms were 2.53% and 2.85% accordingly. When the $E^{\#}$ error rate was used our algorithm was ranked 10th with the score 11.70%. The lowest error rate in this classification was 5.50%, and the mean and the median of the first 10 algorithms were 8.28% and 7.30%. The scores of the first 24 participants are shown on Figs. 4 and 5. The algorithm was implemented

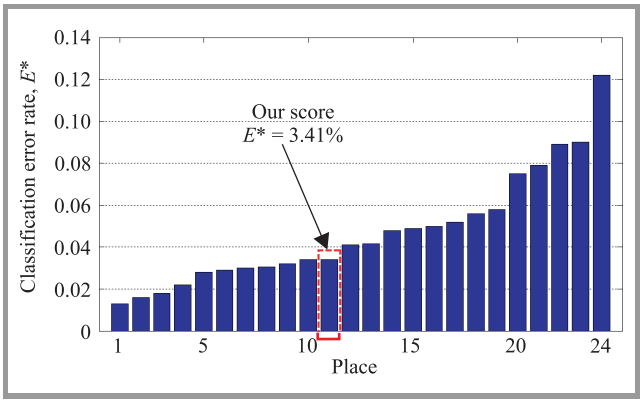



***Fig. 4.*** Classification error rates of the first 24 participants.

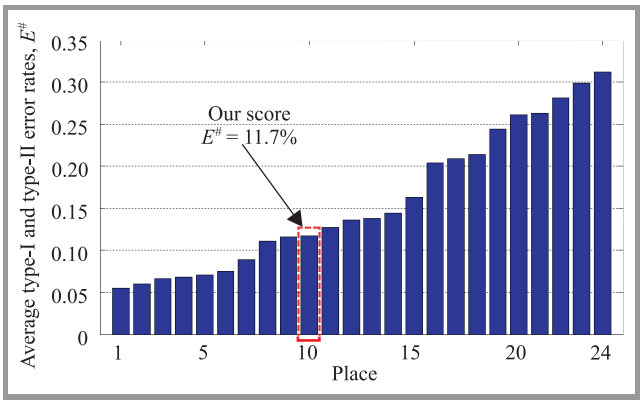



***Fig. 5.*** Average type-I and type-II error rates of the first 24 participants.

in C++ and it is worth to mention that although it was not optimized the segmentation time of the single eye image never exceeded 1 s for the images from UBIRIS database. In our opinion, the results were satisfactory and promising. The main advantages of the method were: the two-phase iris segmentation scheme with coarse localization based on Hough transform and fine segmentation based on spline based active contours. The voting system enabled proper selection of the genuine iris region and reflection removal in preprocessing reduced the errors in the coarse and detailed iris localization. Unfortunately the method has some drawbacks. The submitted version of the algorithm was not fully optimized for the search of pupil, and for some im-

ages the location, the shape and the size of the pupil was guessed based only on the training set and the iris outer boundary. The algorithm had also problems with images, where the iris was not visible. Moreover, it was found later the explicit eyelid, eyelashes and shadow detection models were very important in the NICE.I competition and all the leading algorithms had them implemented. Masking the areas occluded by eyelids and eyelashes significantly reduced the classification error. The conclusion of our tests is that the noisy iris image segmentation still remains an open problem, which deserves more efforts.

## 6. Conclusions

In this paper, we have presented an efficient and robust algorithm for the segmentation of the iris image captured in uncontrollable conditions and less-cooperative context. The algorithm was prepared in response to the NICE.I competition and works on color images, but after some modifications it can be applied to the state-of-art iris recognition devices. The performance evaluation results obtained in the competition shown that the presented approach is promising, but some improvements are required. There is a need to better model the eyelids and eyelashes areas and to introduce more context-based approach. The main advantage of the presented algorithm is the coarse-to-fine iris localization method which roughly approximates the iris shape with circle and later accurately tracks its boundaries with spline-based active contours.

## Acknowledgements

This paper has been financed by the Ministry of Science and Higher Education grant OR00 0026 07 "*A platform of secure biometrics implementations in personal verification and identification*".

## References

[1] H. Proença and L. A. Alexandre, "The NICE.I: noisy iris challenge evaluation, part I", in *Proc. Int. Conf. Biometr. Theory, Appl. Sys.*, Washington, USA, 2007, pp. 1–4.

[2] H. Proença and L. A. Alexandre, "UBIRIS: a noisy iris image database", *Lect. Notes Comput. Sci.*, vol. 3617, pp. 970–977, 2005 [Online]. Available: http://iris.di.ubi.pt

[3] T. Tan, Z. He, and Z. Sun, "Efficient and robust segmentation of noisy iris images for non-cooperative iris recognition", *Image Vision Comput.*, vol. 28, pp. 223–230, 2010.

[4] M. A. Luengo-Oroz, E. Faure, and J. Angulo, "Robust iris segmentation on uncalibrated noisy images using mathematical morphology", *Image Vision Comput.*, vol. 28, pp. 278–284, 2010.

[5] Y. Chen, M. Adjouadi, C. Han, J. Wang, A. Barreto, N. Rishe, and J. Andrian, "A highly accurate and computationally efficient approach for unconstrained iris segmentation", *Image and Vision Comput.*, vol. 28, pp. 261–269, 2010.

[6] W. Sankowski, K. Grabowski, M. Napieralska, M. Zubert, and Andrzej Napieralski, "Reliable algorithm for iris segmentation in eye image", *Image Vision Comput.*, vol. 28, pp. 231–237, 2010.

[7] P. de Almeida, "A knowledge-based approach to the iris segmentation problem", *Image Vision Comput.*, vol. 28, pp. 238–245, 2010.

[8] P. Li, X. Liu, L. Xiao, and Q. Song, "Robust and accurate iris segmentation in very noisy iris images", *Image Vision Comput.*, vol. 28, pp. 246–253, 2010.

[9] D. Sik Jeong, J. Won Hwang, B. Jun Kang, K. Ryoung Park, C. Sun Wonc, D.-K. Park, and J. Kim, "A new iris segmentation method for non-ideal iris images", *Image Vision Comput.*, vol. 28, pp. 254–260, 2010.

[10] R. D. Labati and F. Scotti, "Noisy iris segmentation with boundary regularization and reflections removal", *Image Vision Comput.*, vol. 28, pp. 270–277, 2010.

[11] J. Daugman, "New methods in iris recognition", *IEEE Trans. Sys. Man Cyb., Part B, Cybernetics*, vol. 37, no. 5, pp. 1167–1175, 2007.

[12] J. Daugman, "How iris recognition works?", *IEEE Trans. Circ. Sys. Video Technol.*, vol. 14, no. 1, pp. 21–30, 2004.



**Przemysław Strzelczyk** received his M.Sc. degree in Information Technology in 2005 from the Warsaw University of Technology, and he is currently a PhD candidate in the same field. Since 2005 he is with Reseach and Academic Computer Network (NASK) working for Biometric Laboratories, and within 2007 and 2009 he was research assistant at the Warsaw University of Technology. His main interests include: biometrics, security and methods of artificial intelligence.
e-mail: przemek.strzelczyk@nask.pl
Research and Academic Computer Network (NASK)
Wąwozowa st 18
02-796 Warsaw, Poland

Paper

# Iris Recognition System Based on Zak-Gabor Wavelet Packets

Adam Czajka and Andrzej Pacut

**Abstract—The paper proposes a new iris coding method based on Zak-Gabor wavelet packet transform. The essential component of the iris recognition methodology design is an effective adaptation of the transformation parameters that makes the coding sensitive to the frequencies characterizing ones eye. We thus propose to calculate the between-to-within class ratio of weakly correlated Zak-Gabor transformation coefficients allowing for selection the frequencies the most suitable for iris recognition. The Zak-Gabor-based coding is non-reversible, i.e., it is impossible to reconstruct the original iris image given the iris template. Additionally, the inference about the iris image properties from the Zak-Gabor-based code is limited, providing a possibility to embed the biometric replay attack prevention methodology into the coding. We present the final prototype system design, including the hardware, and evaluate its performance using the database of 720 iris images.**

***Keywords—biometrics, iris recognition, Zak-Gabor transform.***

## 1. Introduction

Biometric researchers are still looking for the *ideal biometrics*, i.e., both a part of the human body and the applied methodology of its symbolic description fused in one system that is characterized by high usability, produces no errors, is robust with respect to variations of attributes of the human body within a large time scale, immune to diseases, resistant to forgery, and produces no social, religious and ethical objections.

The iris is a complex set of interworking muscles, placed anteriorly to the human eye, thus easy to be observed and measured. It is strongly protected by the cornea and eyelids, minimizing the probability of injuries during human life. The structure of the iris tissue is characterized by high stability over the human life span, high degree of structural richness, and almost neglectful dependence on human genotype [1], allowing for recognizing identical twins. In consequence, iris seems to have all of the attributes we'd like to have in biometrics. This paper presents a new iris recognition methodology, that was used to construct fast, highly reliable and non-invasive biometric system.

## 2. Iris Images and Their Preprocessing

### *2.1. Iris Image Capture and Database of Images*

Estimation of the method parameters presented in this work are based on proprietary database of 720 iris images, called here further BioBase. The data was collected for 180 different eyes, with 4 images of each eye. We used 3 images of each eye to calculate the iris templates, and the remaining single image of each eye in the verification stage. Images were collected by IrisCUBE camera (Fig. 1) designed and constructed by the authors to capture the iris from a convenient distance, with the desired speed and a minimal user

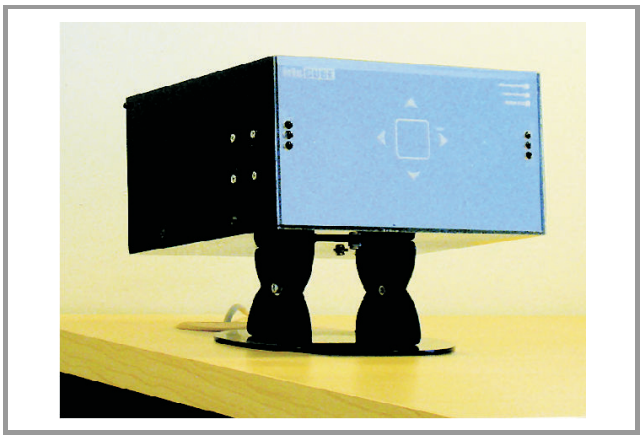


***Fig. 1.*** IrisCUBE camera employed to collect the database used in this work.

cooperation. The camera has an automatic optics to compensate for small depth-of-field that is typical in iris recognition systems. The IrisCUBE camera implements selected aliveness detection methods [2] to deliver actual biometric samples, what still is not a wide practice in commercial iris acquisition systems. The quality of acquired iris images exceeds the highest quality level specified in ISO standard (marked as 'High' in [3]). Since the hardware used allows for one-eye capture, the images taken may be mutually rotated and the rotation of images used in the estimation stage was corrected using the correlation techniques. The remaining fourth image (used in the verification stage) was not altered and the rotation correction mechanism is inscribed in the recognition methodology.

### *2.2. Iris Segmentation*

The raw images contain the iris and its surroundings, thus the iris must be first localized. To detect a boundary between the pupil and the iris, we use a method being a specimen of a commonly applied family of methods sensitive to circular dark shapes, and unresponsive to other dark areas as well as light circles, such as specular reflections. Our implementation is based on a modified Hough transform that employs the image gradient (the so called *directional image*) rather than solely the image gradient value (the so called *edge image*, which neglects the gradient direction).

A boundary between the iris and the sclera is approximated by Daugman's integro-differential operator [1] applied to two opposite horizontally placed angular sectors, 45° each, since the entire circular iris boundary may be partially disturbed by eyelids. The two radii of the resulting arcs are averaged to construct a circle approximating the outer iris boundary.

The iris ring limited by the two circular boundaries may still be disrupted by irregular objects like reflections or eyelashes, hence it is desirable to use occlusion detection that does not assume any particular occlusion shape. We thus assess a non-uniformity along the radial direction within the iris ring and then construct map of irregular occlusions. To do so, we first calculate the sample variances of the iris image intensity for a set of radial sectors along entire iris ring. These variances are then compared to the maximum allowed variance obtained for directions in which the probability of iris occlusion is minimal (i.e., set of directions placed horizontally on both sides of the iris ring). Those directions in which the calculated variance exceeds the threshold value is marked as an occlusion direction, and the appropriate occlusion radius is stored.

Based on the localized occlusions, we select two opposite 90° wide angular iris sectors for coding. The experiments (see also [1]) revealed much higher correlation of the iris image in the radial direction as compared to the angular direction, thus each iris sector is transformed by resampling and smoothing to a $P \times R$ rectangle, where $P = 512$ and $R = 16$. The rows $f_\ell$ of these two rectangles will be further referred to as *iris stripes* (Fig. 2).

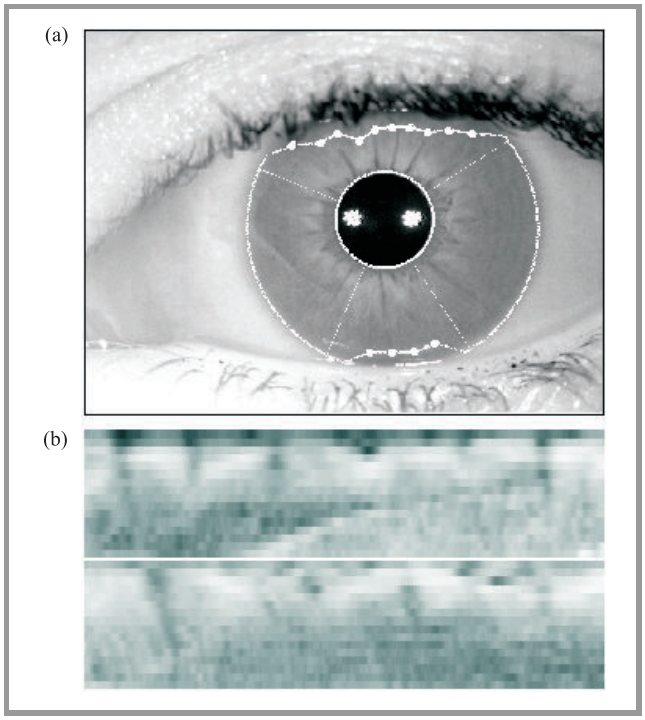



***Fig. 2.*** (a) Raw camera image processed by our system. The eyelids were automatically detected, and the sectors free of occlusions are selected. Star-like shapes on the pupil are reflections of the illumination NIR diodes, and the '+' marks represent the pupil and the iris centers. Full circles correspond to the detected eyelid occlusions. (b) Left and right iris stripes automatically determined for the image shown on (a).

# 3. Iris Features

### *3.1. Choice of Features*

It is often convenient to characterize a discrete-time signal in the frequency domain, thus describing stationary energy distribution. For non-stationary signals, it might be worthwhile to characterize the frequencies locally, and to find the distribution of signal energy in local (possibly overlapping) time segments by application of *time-frequency* or *time-scale* analysis. Similarly, any constant (time-independent) space-homogeneous 1D or 2D pattern (e.g., image) can be characterized in a 1D or 2D frequency domain using "position" statement instead of "time". If a pattern is not space-homogeneous, its spatial frequency contents may be analyzed locally, with the use of *space-frequency* or *space-scale analysis*. Two important candidates for such analysis are *windowed Fourier transform* (WFT, also called "short time Fourier transform", or STFT) and wavelet transform (WT). Both approaches differ. WFT makes use of window functions constant in size, and frequency shifts are achieved by modulating the window. This property is sometimes at a disadvantage, since for each frequency, the number of cycles inscribed in the analysis window differs, resulting in different averaging horizons for different frequencies. In turn, wavelet analysis achieves frequency shifts by *scaling* the position index. Scaling does not change the number of cycles inscribed into the analysis window, thus providing an even analysis for each frequency. These properties of WFT and WT result in different tilings within the 2D position-frequency plane, namely linear for WFT and logarithmic for WT.

The Gabor transform belongs to WT family, and uses a Gaussian window characterized by its width. The window width significantly influences the resulting iris features and must be carefully chosen. We use the space-frequency analysis that employs waveforms indexed by space, scale and frequency simultaneously, what results in a larger set of possible tilling in the space-frequency plane, possibly redundant. This directs our methodology toward a *wavelet packet* analysis. There is a need to select appropriate frequencies and scales simultaneously to make the transformation sensitive to individual features existing in the iris image. In this paper we propose a systematic selection of appropriate scales and frequencies of the iris coding. Although the iris texture makes a 2D pattern, we simplify it to a set of 1D patterns with a certain loss of information and apply the space-frequency analysis locally to the iris circular sectors to describe their local features and to construct a compact iris features set. This approach enables our method to be applied for databases of images of various resolution.

### 3.2. Gabor Expansion

Gaussian-shaped windows are not orthogonal (the inner product of any two of all windows is nonzero), therefore Gabor's expansion coefficients cannot be determined in a simple way. Suggested algorithms include making the window function bi-orthonormal to the Gaussian-shaped elementary function and the matrix-based algorithm [4]. However the Gabor's expansion coefficients determination by application of Zak's transform [5] is considered as the fastest method and it is often referred to as Zak-Gabor's transform. In this section we explain briefly the principles of Gabor expansion coefficient calculation through the Zak's transform for one iris stripe and fixed window width. Denote by $g_s$ a one-dimensional Gaussian elementary function of the width index $s$, sampled at points $0,\ldots,P-1$, namely

$$g_s(p) = e^{-\pi\left((p+\frac{1}{2})/2^s\right)^2}, \quad p = 0,\ldots,P-1, \qquad (1)$$

where $s = 2,3,\ldots,S$, and for the stripe length $P = 512$ we set $S = 8$. If $P$ is (typically) chosen to be even, the $\frac{1}{2}$ term in Eq. (1) makes $g_s$ to be an even function.
Let $M$ be the number of possible translations of $g_s$, and $K$ be the number of frequency shifts, where, following Bastiaans [5], we always take $M = P/K$. A shifted and modulated version $g_{mk;s}$ of the elementary function $g_s$ can be constructed, namely

$$g_{mk;s}(p) = g_s(p - mK)e^{ikp2\pi/K}, \quad p = 0\ldots P-1, \qquad (2)$$

where $m = 0,\ldots,M-1$ and $k = 0,\ldots,K-1$ denote the space and frequency shifts, respectively, and $g_s$ is wrapped around in the $P$-point domain. The finite discrete Gabor transform of the iris stripe $f_\ell$ is defined as a set of complex coefficients $a_{mk;s\ell}$ that satisfy the Gabor signal expansion relationship, namely

$$f_\ell(p) = \sum_{m=0}^{M-1}\sum_{k=0}^{K-1} a_{mk;s\ell}\, g_{mk;s}(p), \quad p = 0\ldots P-1. \qquad (3)$$

Following Bastiaans [5], we further set $K = 2^s$. Note that once the frequency index $k$ is kept constant, $g_{mk;s}$ may be localized in frequency by a modification of $s$. This is done identically as the scaling in a wavelet analysis, hence we call $s$ the *scale* index. The number of Gabor expansion coefficients $a_{mk;s\ell}$ may be interpreted as the signal's number of degrees of freedom. Note that the number $S$ of scales together with the stripe size $P$ determine both $M$ and $K$.

### 3.3. Zak's Transform

The discrete finite Zak transform $\mathscr{Z} f_\ell(\rho,\phi;K,M)$ of a signal $f_\ell$ sampled equidistantly at $P$ points is defined as the one-dimensional discrete Fourier transform of the sequence $f_\ell(\rho + jK)$, $j = 0,\ldots,M-1$, namely [5]

$$\mathscr{Z} f_\ell(\rho,\phi;K,M) = \sum_{j=0}^{M-1} f_\ell(\rho + jK)e^{-ij\phi 2\pi/M}, \qquad (4)$$

where $M = P/K$. Discrete Zak's transform is periodic both in frequency $\phi$ (with the period $2\pi/M$) and location $\rho$ (with the period $K$). We choose $\phi$ and $\rho$ within the fundamental Zak interval [5], namely $\phi = 0,1,\ldots,M-1$ and $\rho = 0,1,\ldots,K-1$. Similarly to the Fourier transformation, one may reconstruct the original function $f_\ell$ from its Zak transform by way of the inverse discrete Zak transform, using the following formula

$$f_\ell(\rho + jK) = \frac{1}{M}\sum_{\phi=0}^{M-1} \mathscr{Z} f_\ell(\rho,\phi;K,M)e^{ij\phi 2\pi/M} \qquad (5)$$

and restricting the domain of the results to the fundamental Zak interval.

### 3.4. Application of Zak's Transform

Application of the discrete Zak transform to both sides of Eq. (3) and rearranging the factors yields

$$\begin{aligned}
&\mathscr{Z} f_\ell(\rho,\phi;K,M) = \\
&= \sum_{j}^{M-1}\left[\sum_{m}^{M-1}\sum_{k}^{K-1} a_{mk;s\ell}\, g_s(\rho + jK - mK)e^{ik\rho 2\pi/K}\right]e^{-ij\phi 2\pi/M} = \\
&= \left[\sum_{m=0}^{M-1}\sum_{k=0}^{K-1} a_{mk;s\ell}e^{-i2\pi(m\phi/M - k\rho/K)}\right]\cdot \\
&\cdot\left[\sum_{j=0}^{M-1} g_s(\rho + jK)e^{-i2\pi j\phi/M}\right] = \\
&= \mathscr{F} a_{s\ell}(\rho,\phi;K,M)\,\mathscr{Z} g_s(\rho,\phi;K,M), \qquad (6)
\end{aligned}$$

where $\mathscr{F} a_{s\ell}[\rho,\phi;K,M]$ denotes the discrete 2D Fourier transform of an array of $a_{s\ell}$ that represents Gabor's expansion coefficients determined for the iris stripe $f_\ell$ and scale $s$, and $\mathscr{Z} g_s[\rho,\phi;K,M]$ is discrete Zak's transform of the elementary function $g_s$.
This shows that Gabor's expansion coefficients can be recovered from the product form Eq. (6). Once $K$ and $M$ are chosen to be powers of 2 (making also the signal length $P$ to be a power of 2), the calculation of both $\mathscr{Z} f[\rho,\phi;K,M]$ and $\mathscr{Z} g[\rho,\phi;K,M]$, and inversion of 2D Fourier series can employ fast Fourier transform thus yielding computation times proportional to those in the FFT.

### 3.5. Definition of Iris Features

Calculation of Gabor's transform for all iris stripes and for all scales results in a set of coefficients $a$ indexed by the quadruple: within-stripe position, frequency index, scale and stripe index $(m,k,s,\ell)$. Inspired by Daugman's work [1], we define the signs of the real and imaginary parts of Zak-Gabor coefficients as the feature set $\mathbb{B}$, namely

$$\mathbb{B} = \left\{\mathrm{sgn}(\Re(a_{mk;s\ell})), \mathrm{sgn}(\Im(a_{mk;s\ell}))\right\}, \qquad (7)$$

where: $m = 0,\ldots,M-1$, $k = 0,\ldots,K-1$, $\ell = 0,\ldots,2R-1$ and $s = 2,3,\ldots,S$.

Since the Fourier transform is symmetrical for real signals, for each position $m$ the coefficients with the frequency index $k > K/2$ can be ignored. Since $M = P/K$, for each $s$ there are $(N-1)P/2$ coefficients to be determined. Taking into account that this analysis is carried out for all iris stripes, and remembering that $R = 16,\ S = 8$ and $P = 512$, the total number of coefficients calculated for the iris image is $R(S-1)P = 57,344$. Both real and imaginary parts are coded separately by one bit, hence $N = |\mathbb{B}| = 114,688$ features may be achieved, where $|\cdot|$ denotes the number of elements in a finite set. The features, positioned identically for each iris, may thus form a binary vector. Thus, matching two features requires only a single XOR operation, and the Hamming distance can be applied to calculate the score.

We stress that $\mathbb{B}$ should not be confused with the so called *iriscode*™ invented by Daugman. The latter one is a result of an iris image filtering, while $\mathbb{B}$ is constructed with Gabor expansion coefficients. We now describe how to select a subset of features $\mathbb{B}$ to be included in a final feature set $\mathbb{B}^*$.

### 3.6. Feature Set Selection

The selection of scales and frequencies of Zak-Gabor coefficients included into the code, and thus selection of the scale $s$ in Eq. (1) and the frequency index $k$, cannot be guessed *a priori*, due to significant and undetermined iris texture variability. Both parameters are interdependent, have a strong influence on the overall method's efficiency and should be considered simultaneously. Moreover, the full feature set $\mathbb{B}$ is significantly oversized, since it consists of features representing all, possibly inadequate, frequencies of the analyzed image. We thus propose a two-stage method that selects optimal Zak-Gabor based features and can be used to estimate optimal feature set given the quality of iris images. Since only certain subset of $\mathbb{B}$ will be included into the final feature set, all elements of $\mathbb{B}$ will be further referred to as *candidate features*.

**Stage one: selection of useful features**. The first selection stage consists of choosing a subset $\mathbb{B}^0$ of candidate features $\mathbb{B}$, called here the *useful features*. To determine $\mathbb{B}^0$, we analyze a variability of candidate features.

For each feature $b$ we calculate the *within-eye sum of squares* $SS^W(b)$, and the *between-eye sum of squares* $SS^B(b)$. Intuitively, a feature is *useful* if at least $SS_n^{(W)} < SS_n^{(B)}$. Typically, the number of bits that meet this requirement is still too high (approximately half of bits $b_n$) and ends up with highly correlated features. We thus introduce a stronger selection mechanisms and categorize features to maximize $SS^B$ and minimize $SS^W$ solving this multicriteria problem by minimizing the distance from the most desired point on $SS^W \times SS^B$ plane. This point was set as $\left(\min_{b\in\mathbb{B}} SS^W(b), \max_{b\in\mathbb{B}} SS^B(b)\right)$, Fig. 3.

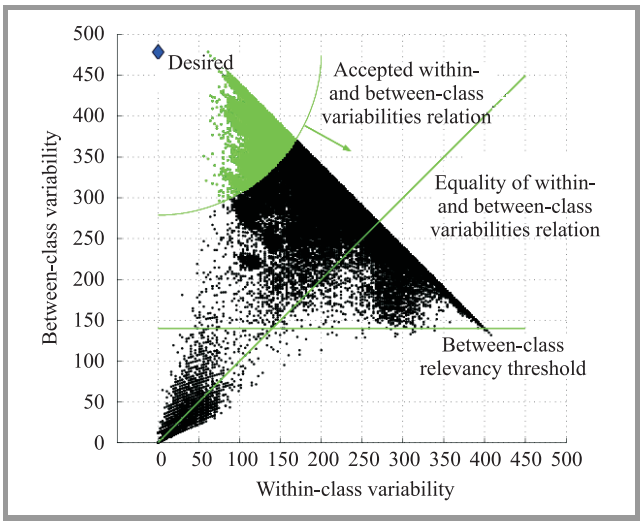



***Fig. 3.*** Within-eye sum of squares versus between-eye sum of squares and the area of *useful features*. Each dot corresponds to real or imaginary part of one Zak-Gabor coefficient. The "desired" point is also marked (we favor coefficients with minimum within-eye and simultaneous maximum between-eye sum of squares).

We use the order introduced by the above procedure in the set of candidate features $\mathbb{B}$ in a procedure removing a high correlation of candidate features to increase an "information density". We include $k$th candidate feature into the set $\mathbb{B}^0$ only if it is *not strongly correlated* with all the features already selected.

We base our useful feature definition on the *decidability coefficient* $d'$ [1] calculated for a given feature subset. We calculate the decidability coefficient for each set of candidate features included into $\mathbb{B}^0$. The decidability varies with the number of candidate features included: it first grows to reach the maximum and then decreases. Experiments show that the decidability $d'$ is highest for the correlation threshold around 0.3, Fig. 4.

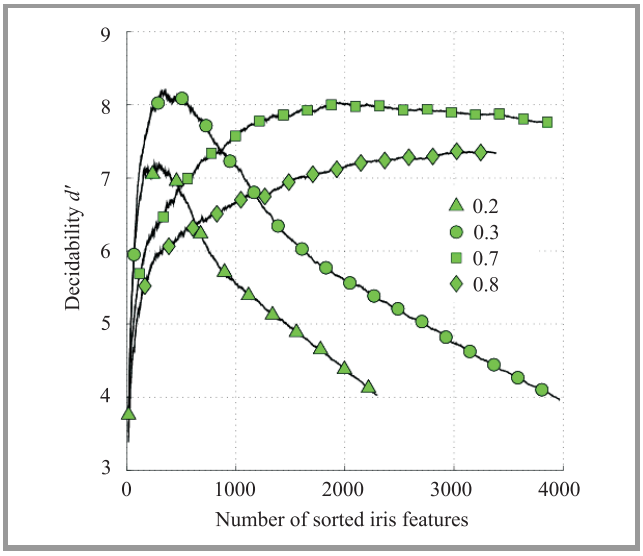



***Fig. 4.*** Decidability coefficient versus number of useful features selected for a few sample correlation coefficients (0.2, 0.3, 0.7 and 0.8) allowed within the *useful features* set.

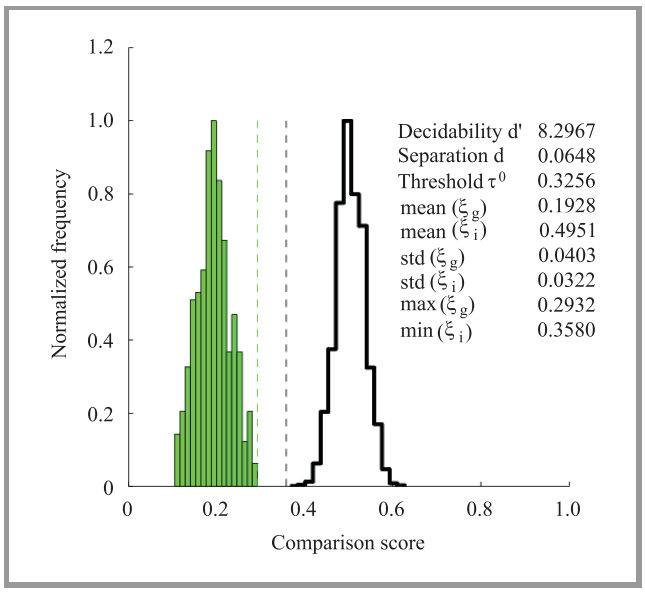



***Fig. 5.*** Score distributions for 360 genuine (left) and 64,440 impostor (right) comparisons, denoted by $\xi_g$ and $\xi_i$, respectively. Only the useful features $\mathbb{B}^0$ are used. No sample errors were registered for a wide range of acceptance threshold $\tau \in (0.2932, 0.358)$, and in particular for optimal threshold $\tau^0 = 0.3256$.

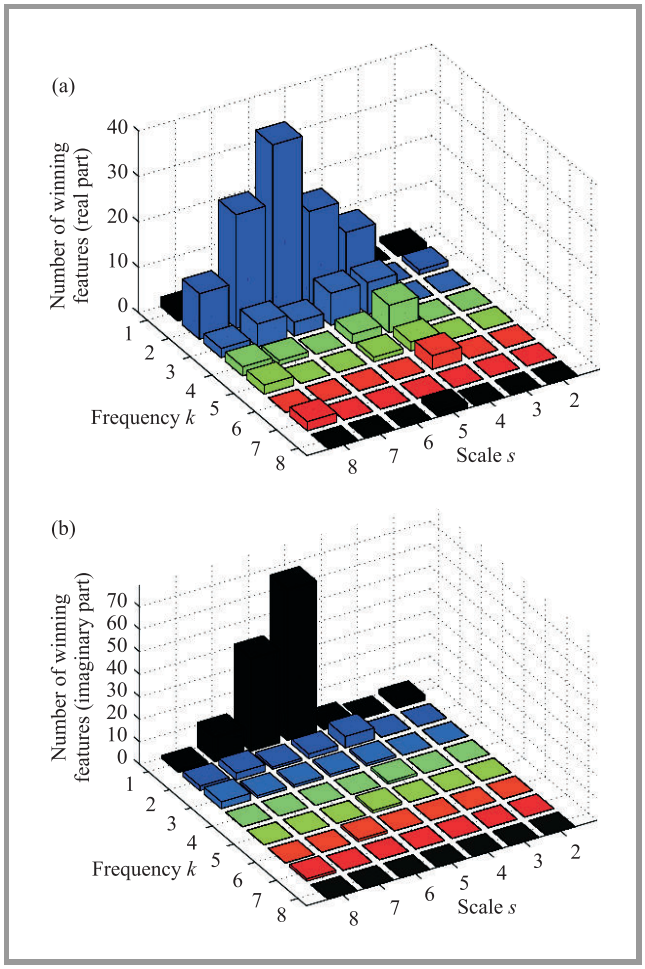



***Fig. 6.*** 2D histogram of how families $\mathbb{B}_{k,s}$ are "populated" by useful features $\mathbb{B}^0$ determined separately for the real (a) and imaginary (b) parts of Zak-Gabor coefficients.

For this solution there is no between-eye – within-eye overlap of sample distributions, i.e., there are no false matches and no false non-match examples in the estimation data set, Fig. 5. The resulting 324 useful features pass to the second feature selection stage. We may add that our procedure included only such features for which $SS^W < SS^B$.

**Stage two: selection of feature families**. Let $\nu(k,s)$ denote the number of useful features in the *candidate features family* $\mathbb{B}_{k,s}$, which represent all candidate features that are labeled by the same scale $k$ and frequency $s$, and differ by space indices $m$ and $\ell$, namely

$$\mathbb{B}_{k,s} = \{\mathrm{sgn}(\Re(a_{mk;s\ell})), \mathrm{sgn}(\Im(a_{mk;s\ell})) : \\ m = 0,\ldots,M-1, \ell = 0,\ldots,2R-1\}. \tag{8}$$

The higher is $\nu(k,s)$, the more important are the frequency indexed by $k$ and the scale indexed by $s$ in iris recognition. To decide for the best frequencies and scales, independently for real or imaginary parts of the Zak-Gabor coefficients, we sort $\mathbb{B}_{k,s}$ by decreasing $\nu(k,s)$ separately for real and imaginary parts of coefficients. Figure 6 depicts the 'population' of scale-frequency families by winning features $\mathbb{B}^0$. Note that the number of winning features is not identical for all families.

We further prioritize the families that are most frequently 'populated' by the useful features. The sorting rule for families of features mirrors the rule used for useful features selection: we sort $\mathbb{B}_{k,s}$ by the decreasing number of useful features $\mathbb{B}^0$ included in a given family, separately for real and imaginary parts of coefficients. We check the decidability $d'$ and separation $d$ (the difference between maximum genuine score and minimum impostor score, note that $d > 0$ denotes perfect separation between distributions of genuine and impostor scores) and chose the families resulting in maximum $d'$ for a given database, Fig. 7.

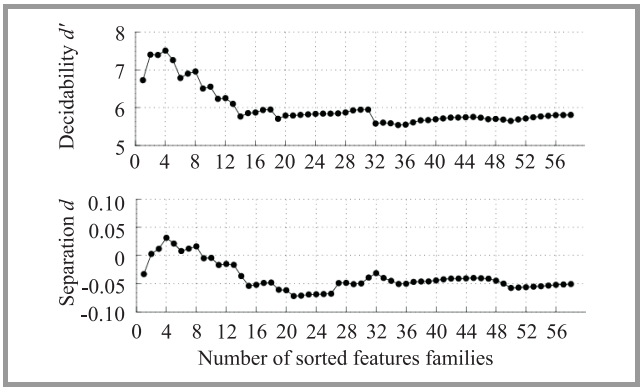



***Fig. 7.*** Decidability $d'$ and separation margin $d$ versus the number of sorted feature families $\mathbb{B}_{k,s}$ included in the feature set.

This rule allows finding the frequency-scale pairs of real and imaginary parts of Zak-Gabor coefficients, which, if chosen as iris features, result in the best separation of distributions of genuine and impostor comparison scores. The feature families set resulting in maximum $d'$ constitutes the *final feature set* $\mathbb{B}^*$, which in our case contains only

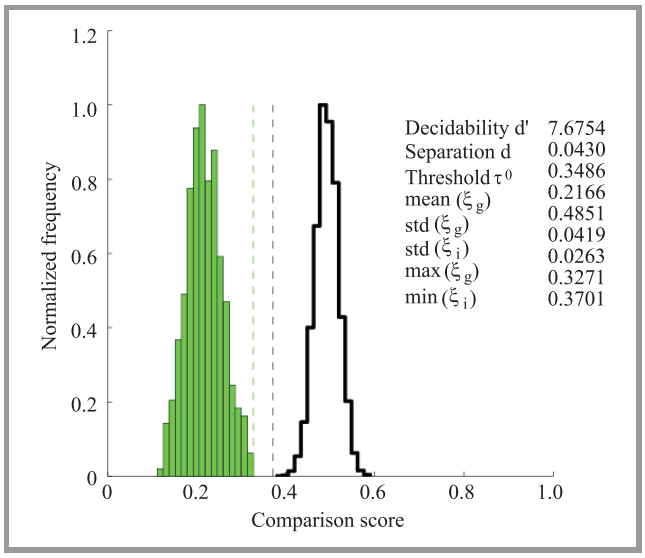



***Fig. 8.*** Same as in Fig. 5, except the selected feature families are used to build the iris feature set. No sample errors are registered for a wide range of acceptance threshold $\tau \in (0.3271, 0.3701)$, and in particular for optimal threshold $\tau^0 = 0.3486$.

four families, what ends up with 1024 bit code. For this final feature set, we still achieved no sample verification errors, Fig. 8.

### *3.7. Personalize Feature Subsets*

Once the optimal feature families, namely the best scale-frequency pairs indexed by $s$ and $k$, are selected, the iris features set is calculated for those chosen $s$ and $k$ and all $m = 1,\ldots,M-1$, and $\ell = 0,\ldots,2R-1$. Each Zak-Gabor coefficient can 'measure' the correlation between the modulated Gaussian elementary function $g_{mk;s}$ and the corresponding iris stripe. The question arises how 'robust' are the consecutive Zak-Gabor coefficients against noise, and iris tissue elastic constrictions and dilations.

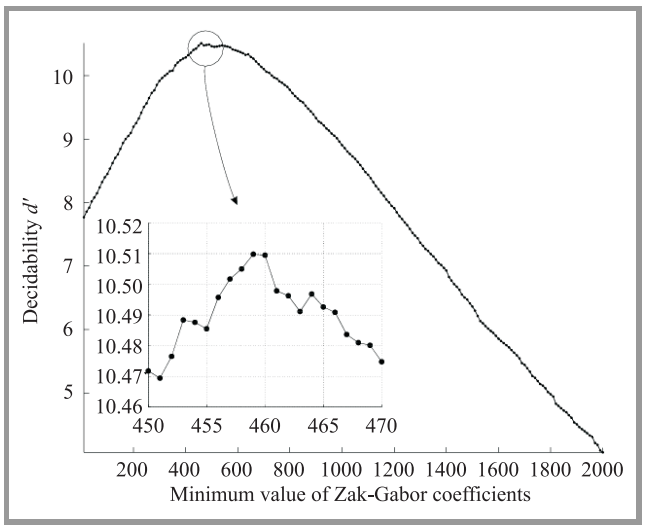



***Fig. 9.*** Decidability versus minimum value of Zak-Gabor coefficients, required to be included in the feature set. For BioBase data, best decidability $d' = 10.5095$ is achieved if using coefficients $a_{mk;s\ell} \geq a^{\text{thr}} = 459$

Due to a significant variability of the iris tissue, some $g_{mk;s}$ may not conform with the iris body, resulting in small coefficients. Such a situation is dangerous, since once the coefficients are close to zero, their signs may depend mostly on a camera noise, and consequently may weaken the final code. This motivates personalization of the iris feature sets that employ only those Zak-Gabor coefficients that exceed experimentally determined threshold $a^{\text{thr}}$, which is the minimum value of Zak-Gabor coefficiens $a_{mk;s\ell}$ required to become a *relevant coefficient* being a base of an iris feature.

To answer the question how big the value of $a^{\text{thr}}$ should be, we maximize the decidability $d'$ using $a^{\text{thr}}$ as a parameter given a database of iris images. By increasing $a^{\text{thr}}$ a greater number of coefficients are neglected. According to observations (Fig. 9) the system reliability first increases, then deteriorates, and the maximum $d' = 10.51$ for $a^{\text{thr}} = 459$ can be found, which is higher than $d' = 7.67$ achieved for a full set of 1024 feature bits (cf. Fig. 8). Although the separation margin $d$ is not increased significantly for the determined $a^{\text{thr}}$, the distribution tails are larger, and the score averages for the comparison of the same and different eyes are spaced more widely compared to non-personalized technique, Fig. 10.

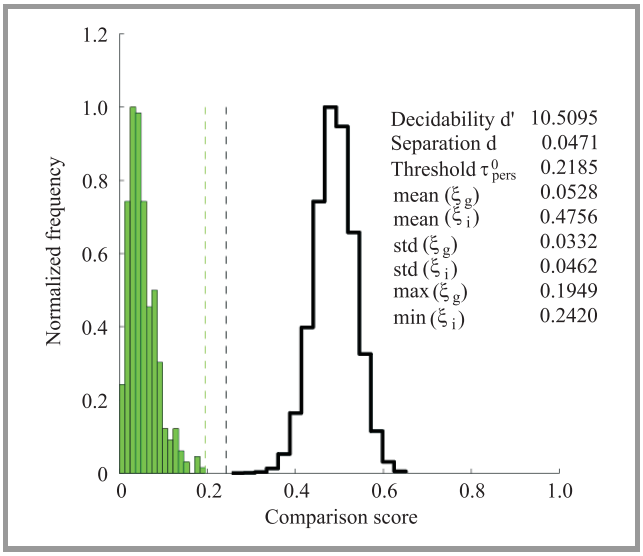



***Fig. 10.*** Same as in Fig. 5, except the personalized features are used for each person, in the way to guarantee the best possible decidability $d'$ for BioBase data, and the optimal threshold $\tau^0_{\text{pers}} = 0.2185$.

To distinguish between relevant and irrelevant features within the final feature set $\mathbb{B}^*$, we introduce a set of masking bits, thus enlarging twice the required data for the iris template (we need to add 1024 bits to the existing set of 1024 bits representing iris features).

# 4. Iris Recognition System

### *4.1. Iris Template Creation*

Image quality influences the reliability of the feature set. Consequently, a quality check is usually performed during enrollment which is slightly longer than the verification stage. We propose a two-stage procedure that leads to template internal consistency. This procedure has been applied in the prototype system. The first stage encompasses raw image quality check (calculation of the focus factor, eyelids/eyelashes coverage, identification of existence of specular reflections). After a successful check, eyeball rotation is corrected using correlation methods for all three images used for template creation. Since the aim is to enroll samples which are close in terms of the comparison score, the second stage investigates the consistency of acquired images as measured within the feature space. To check this, all possible comparisons are made between template feature sets. To pass the consistency check, all resulting scores should be lower than the acceptance threshold. We used thresholds established at the estimation stage (Subsecs. 3.6 and 3.7) as those values guarantee no sample errors, yet it is a choice of system administrator who may tune the template quality settings adequately to his needs. Consequently, as the iris template we select this feature set, for which the distance to the remaining feature sets is minimal (best candidate approach).

### *4.2. Eyeball Rotation Correction and Iris Verification*

Small eyeball rotations in consecutive images may lead to considerable deterioration of within-eye comparison scores, Fig. 11. Since during verification the iris image

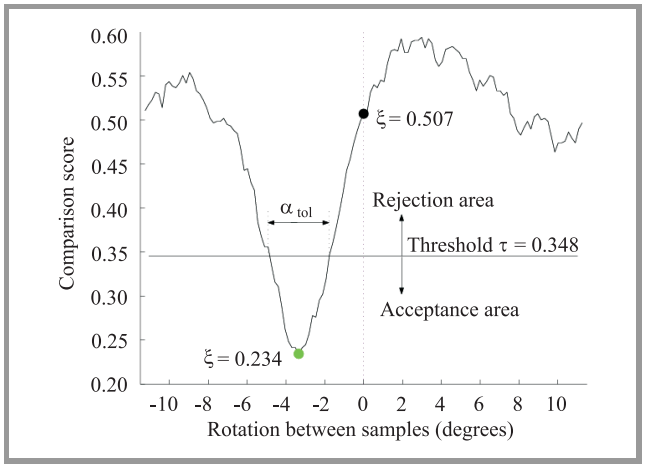



***Fig. 11.*** Comparison score $\xi$ vsersus mutual rotation angle $\alpha$ for two images of the same iris (results shown for one sector). Note that without rotation compensation, a non-match would be observed. The eye rotation tolerance $\alpha_{\text{tol}}$ (given the acceptance threshold) of a single code, illustrating its robustness to eyeball rotations is also marked; for BioBase data and the acceptance threshold $\tau^0 = 0.3486$ the average tolerance $\overline{\alpha}_{\text{tol}} \approx \pi/60$.

corresponding to the template is unavailable, the rotation cannot be corrected by maximizing the correlation between the images and another methodology must be applied.

We use an iterative minimization of the comparison score between Zak-Gabor-based features determined for a set of small artificial shifts of the iris stripes being verified. This method is applied to both iris sectors independently, and the resulting codes corresponding to both sectors are compared separately. Obtained scores are averaged into the final score. However, correcting each incoming image is not reasonable, since a number of them do not require additional action, due to initial code robustness to eyeball slope. Thus, we apply a staged verification procedure that compensates for eye rotation only if necessary, i.e., if the comparison score does not drop below the acceptance threshold. Such an approach takes into account engineering aspects, since this minimizes the verification time. Approximately 55% of iris images in BioBase captured for verification do not require rotation correction for the threshold $\tau^0 = 0.3486$, and in the remaining genuine transactions we needed only 5 iterations (i.e., calculation of iris template for artificially shifted iris stripes) to find the best match.

# 5. System Evaluation

### *5.1. Recognition Methodology Performance*

For the purpose of evaluation, $N_g = K = 180$ genuine and $N_i = K(K-1)/2 = 32,220$ impostor comparisons were made, where $K = 180$ is the number of verification im-

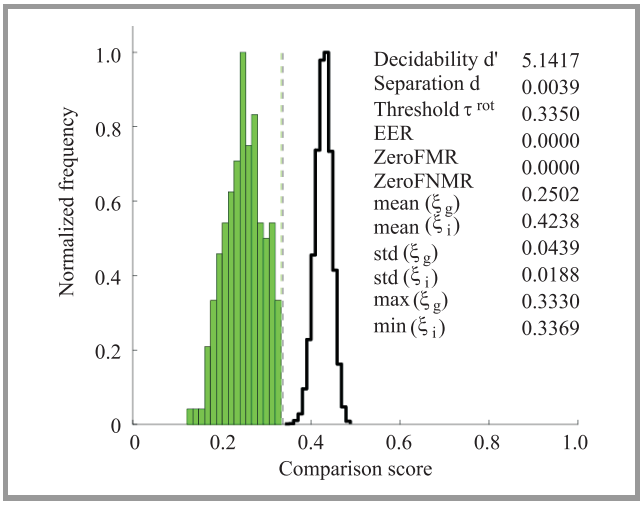



***Fig. 12.*** Sample distributions of scores for non-personalized recognition method employing iterative minimization for eyeball slope correction. No sample errors were encountered for BioBase data and selected threshold $\tau^{\text{rot}} = 0.3350$.

ages, each representing different eye. Prior to the eyeball slope correction procedure, one should select the acceptance threshold $\tau^{\text{rot}}$. Note that the thresholds $\tau^{\text{rot}} = 0.3486$

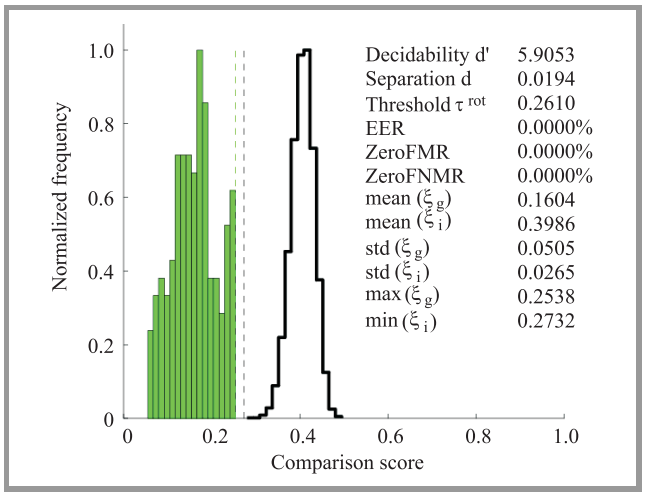



***Fig. 13.*** Same as in Fig. 12, except the personalized recognition method is employed. No sample errors were encountered for BioBase data and selected threshold $\tau^{\text{rot}}_{\text{pers}} = 0.2610$.

used for Zak-Gabor-based coding for initially rotated samples (and $\tau^{\text{rot}}_{\text{pers}} = 0.2185$ when the personalized coding is used) are no longer valid, since the score calculation in this approach differs, i.e. it is iteratively minimized. Thus, to select the operating thresholds, full inspection is performed for all possible eyeball rotations and we set $\tau^{\text{rot}} = 0.3350$ and $\tau^{\text{rot}}_{\text{pers}} = 0.2610$ for non-personalized and personalized coding, respectively. No sample errors were encountered for BioBase data and selected thresholds, Figs. 12 and 13.

### *5.2. Operating Times*

The methodology (and its variants) presented in the paper was implemented as the Software Development Kit [6] and was integrated with the IrisCUBE camera forming a standalone recognition system prototype. Current prototype is based on a PC workstation equipped with a 2.0 GHz Pentium processing unit, 1 GB RAM, controlled by Windows XP operating system, yet the system requirements guarantying reasonable processing times are much lower.

Table 1
Iris image acquisition and processing times achieved by the prototype system employing IrisCUBE camera, averaged for all BioBase acquisition sessions

| Task | Average time [s] |
|---|---|
| Head positioning by skilled volunteer | 2.5 |
| Acquisition of frames | 1.0 |
| Best frame selection | 1.5 |
| Iris boundary localization and occlusions detection | 0.8 |
| Representation of iris image as a sequence of stripes | 0.5 |
| Zak-Gabor coefficients calculation and transformation into a features vector | 0.05 |
| Matching (with iterative minimization) | 0.25 |
| Total | 6.6 |

Table 1 summarizes acquisition and processing times measured for all IrisCUBE acquisitions for BioBase. Although short training was offered to volunteers prior to data collection, the acquisition times are prone to relatively high uncertainty, due to the huge variability and unpredictability of human skills while positioning the subject's head. Processing times depend only on software implementation, hence the results may be predicted with a higher certainty in comparison to the volunteer behavior. The entire verification time, including the volunteer's mandatory cooperation and image acquisition, typically does not exceed 7 s, which is recognized by volunteers as an acceptable result.

## 6. Summary

The iris recognition project detailed in this paper encompasses the entire recognition system. We proposed a systematic approach of selection of the Zak-Gabor based coding parameters employing variance analysis of the iris features. The procedure allows selecting the frequency and scale of the image transformation appropriate in terms of the system reliability to the given iris image quality and resolution. This feature selection procedure can be applied also to other iris coding methods based on wavelet analysis. Presented methodology was used in a number of applications, for instance in remote access scenario and biometric smart card development. It was also evaluated in iris recognition system prototype with eye aliveness detection.

## Acknowledgements

This paper has been financed by the Ministry of Science and Higher Education grant OR00 0026 07 "*A platform of secure biometrics implementations in personal verification and identification*".

## References

[1] J. Daugman, "Biometric personal identification system based on iris analysis", *United States Patent* US 5,291,560, assignee: IriScan Inc., NJ, USA, March 1, 1994.

[2] A. Pacut and A. Czajka, "Aliveness detection for iris biometrics", in *Proc. 40th IEEE Int. Carnahan Conf. Secur. Technol. ICCST 2006*, Lexington, USA, 2006.

[3] *Information technology – Biometric data interchange formats – Part 6: Iris image data*, ISO/IEC International Standard 19794-6:2005(E)

[4] T. T. Chinen and T. R. Reed, "A performance analysis of fast gabor transform methods", *Graph. Mod. Image Proces.*, vol. 59, no. 3, pp. 117–127, 1997.

[5] M. J. Bastiaans, "Gabor's expansion and the zak transform for continuous-time and discrete-time signals", in *Signal and Image Representation in Combined Spaces*, J. Zeevi and R. Coifman, Eds. Academic Press, 1995, pp. 1–43.

[6] A. Czajka and A. Pacut, "BiomIrisSDK – software development kit for iris recognition", *NASK Review*, pp. 34–39, 2009.



**Adam Czajka** Adam Czajka received his M.Sc. in computer control systems in 2000 and Ph.D. in control and robotics in 2005 from Warsaw University of Technology, Poland. Since 2003 he is with Warsaw University of Technology, and since 2002 with Research and Academic Computer Network NASK working for Biometric Laboratories. He is a member of the NASK Research Council (2006–), NASK representative in Technical Committee No. 182 on Information Security in IT Systems (2007–) and in Technical Committee No. 309 on Biometrics (2010–) of Polish Committee for Standardization PKN. He is also a member of the IEEE (2002–) and recently served as the Secretary of the IEEE Poland Section (2006–2009).
e-mail: Adam.Czajka@nask.pl
Biometric Laboratories
Research and Academic Computer Network (NASK)
Wawozowa st 18
02-796 Warsaw, Poland



**Andrzej Pacut** holds Ph.D. in electronics and D.Sc. in control and robotics. Since 1969 he is with Warsaw University of Technology, being presently a professor and the head of Biometrics and Machine Learning Group in the Institute of Control and Computation Engineering. Since 2001 he is also the head of Biometric Laboratory of Research and Academic Computer Network NASK. He was Visiting Prof. in the Lefschetz Center for Dynamic Systems at Brown University, Providence, Rhode Island 1980–1981, and Visiting Prof. in the Department of Electrical and Computer Engineering of Oregon State University, Corvallis, Oregon, 1986–1991. He is a senior member of the IEEE and served as the Vice-Chair and then Chair of Poland Section of the IEEE in 2002–2009. He is the head of Technical Committee 309 on Biometrics of Polish Committee for Standardization PKN. He is interested in learning systems, biometrics, identification, bio-inspired modeling and control, and related areas.
e-mail: A.Pacut@ia.pw.edu.pl
Institute of Control and Computation Engineering
Warsaw University of Technology
Nowowiejska st 15/19
00-665 Warsaw, Poland

Biometric Laboratories
Research and Academic Computer Network NASK
Wąwozowa st 18
02-796 Warsaw, Poland

Paper

# Image Preprocessing for Illumination Invariant Face Verification

Mariusz Leszczyński

**Abstract—Performance of the face verification system depend on many conditions. One of the most problematic is varying illumination condition. In this paper 14 normalization algorithms based on histogram normalization, illumination properties and the human perception theory were compared using 3 verification methods. The results obtained from the experiments showed that the illumination preprocessing methods significantly improves the verification rate and it's a very important step in face verification system.**

***Keywords—DLDA, face verification, histogram normalization, homomorphic filtering, illumination normalization, LDA, PCA, preprocessing techniques, quotient image, retinex.***

## 1. Introduction

Face is one of the most commonly used by people to recognize each other. Over the course of its evolution, the human brain has developed highly specialized areas dedicated to the analysis of the facial images [1]. In the past decades, face recognition has been an active research area and many types of algorithms and techniques has been proposed to equal this ability of human brain. It is however questioned whether the face itself is a sufficient basis for recognizing, a person from large population with great accuracy. Indeed, the human brain also relies on many contextual information and operate on limited population.

The most problematic perturbation affecting the performance of face recognition systems are strong variations in pose and illumination. Variation between images of different faces in general is smaller than taken from the same face in a variety of environments [2]. In face verification system authenticates a person's claimed identity and decide that claimed identity is correct or not. In this case we have limited user group and in the most cases we can forced or demand frontal pose orientations. Unfortunately we still have problems with illumination condition. Face recognition tests [3]–[6] revealed that the lighting variant is one of the bottlenecks in face recognition/verification. If lighting conditions are different from the gallery identity decision is wrong in many cases.

There are two approaches to this problem. Model-based [7], [8] and preprocessing-based. Model-based attempt to model the light variation. Unfortunately, this requires large amount of training data and sometimes fall when we have complicated lighting configuration.

The second approach using preprocessing methods to remove lighting influence effect without any additional knowledge. In this paper, we compare 14 normalization algorithms using 3 verification methods.

## 2. Histogram Normalization

Illumination preprocessing on 2D images can be divided into two groups: histogram transformation and photometric normalization.

### *2.1. Histogram Equalization (HQ)*

Histogram normalization is one of the most commonly used methods. In image processing, the idea of equalizing a histogram is to stretch and/or redistribute the original histogram using the entire range of discrete levels of the image, in a way that an enhancement of image contrast is achieved. The most common used histogram normalization technique is histogram equalization where one attempts to change the image histogram into a histogram that is constant for all brightness values. This would correspond to a brightness distribution where all values are equally probable. For image $I(x,y)$ with discrete $k$ gray values histogram is defined by:

$$p(i) = \frac{n_i}{N}, \tag{1}$$

where: $i \in 0,1,\ldots,k-1$ grey level and $N$ is total number of pixels in the image.

Transformation to a new intensity value is defined by:

$$i_{out} = \sum_{i=0}^{k-1} \frac{n_i}{N} = \sum_{i=0}^{k-1} p(i). \tag{2}$$

Output values are from domain of $[0,1]$. To obtain pixel values in to original domain, it must be rescaled by the $k-1$ value.

Figure 1 shows two face images with different light condition and preprocessed images with corresponding histograms.

### *2.2. Histogram Truncation and Stretching (HT)*

Histogram stretching consists in distributing the pixel appearance frequencies over the entire width of the histogram.

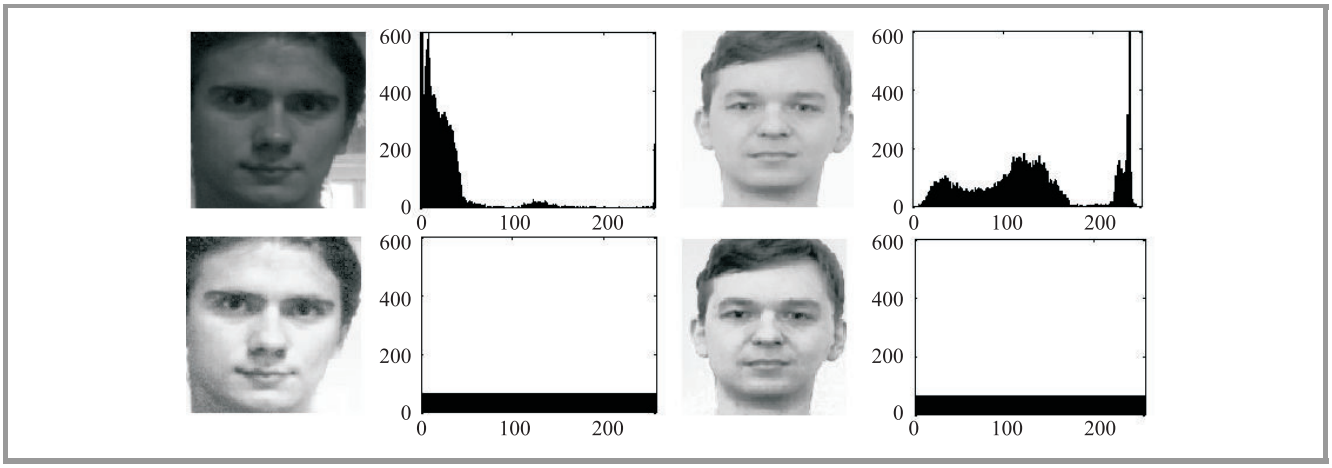



*Fig. 1.* Two sample images with histogram before (upper) and after (lower).

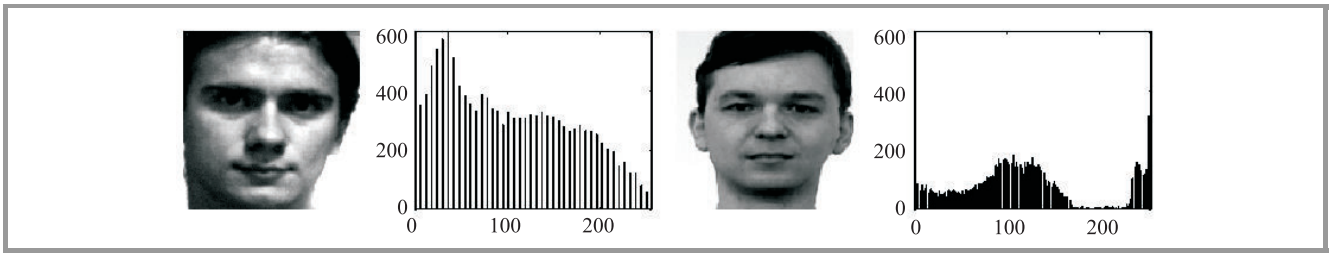



*Fig. 2.* Two sample images with histogram after histogram truncation and stretching.

Thus, it is an operation that consists in modifying the histogram in such a way as to distribute the intensities on the scale of values available as well as possible. This amounts to extending the histogram so that the value of the lowest intensity is zero and that of the highest is the maximum value. In this way, if the values of the histogram are very close to each other, the stretching will make it possible to provide a better distribution in order to make light pixels even lighter and dark pixels closer to black. Additional ten percentage of the lower and upper ends of an image histogram are truncated. This solves the problem when few very bright or dark pixels have the overall effect of darkening or brightening the rest of the image after rescaling (see Fig. 2).

### *2.3. Histogram Modeling*

Histogram preprocessing is not only limited for HQ and stretching. We can model it with different density function where brightness distribution will be mapped to a specific probability distribution. According to [9], [10] general mapping function for the distribution function $f(x)$ may be calculated from:

$$\frac{N-R+0.5}{N} = \int_{x=-\infty}^{t} f(x)dx, \tag{3}$$

where: $R$ is rank of the pixels ordered from smallest intensity to the largest intensity value with assigned rank from 1 to $N$.

The right side of Eq. (3) represents target cumulative distribution function (CDF). The searching $t$ parameter will be computed by from the inverse CDF of the left side Eq. (3).

### *2.4. Normal Distribution (ND)*

The first consider distribution is normal distribution, which is the most commonly observed probability distribution. It was first described by De Moivre in 1733. Laplace used the normal curve in 1783 to describe the distribution of errors. Subsequently, Gauss used the normal curve to analyze astronomical data in 1809. The normal curve is often called the Gaussian distribution and its defined by the following equation:

$$f(x) = \frac{1}{\sigma\sqrt{2\pi}} \exp\left(\frac{-(x-\mu)^2}{2\sigma^2}\right), \tag{4}$$

where: $\mu$ is the mean and the second, $\sigma$ is the standard deviation.

In our experiments (Fig. 3) we use the standard normal distribution, where $\mu = 0$ and $\sigma^2 = 1$.

### *2.5. Lognormal Distribution (LN)*

The lognormal distribution is an asymmetric distribution. Many physical, chemical, biological, toxicological, and statistical processes tend to create random variables that follow

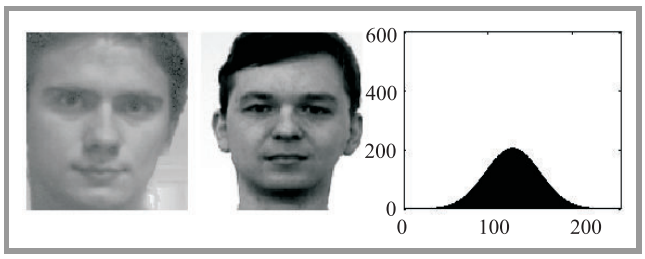



***Fig. 3.*** Two sample images with histogram after mapping the histogram to a normal distribution.

lognormal distributions. For example, lognormal distributions can model certain instances, such as the change in price distribution of a stock or survival rates of cancer patients or failure rates in product tests.
Density function for this distribution is defined by:

$$f(x) = \frac{1}{x\sigma\sqrt{2\pi}} \exp\left(\frac{-(\ln x - \mu)^2}{2\sigma^2}\right), \qquad (5)$$

for our experiments (Fig. 4) mean $\mu = 0$ and standard deviation $\sigma = 0.25$.

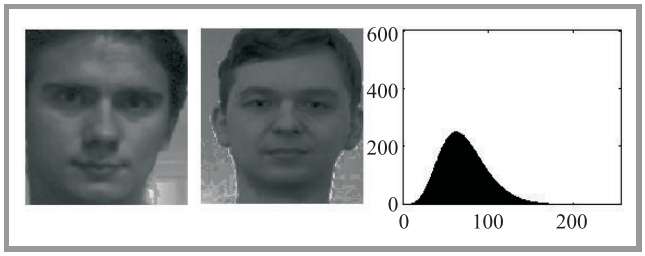



***Fig. 4.*** Two sample images with histogram after mapping the histogram to a lognormal distribution.

### *2.6. Extreme Value Distribution (EV)*

The third distribution is called extreme value distribution and appropriate for modeling many rare events, and has the following probability density function:

$$f(x) = \sigma^{-1} \exp\left(\frac{x-\mu}{\sigma}\right) \exp\left(-\exp\left(\frac{x-\mu}{\sigma}\right)\right), \qquad (6)$$

where: $\mu$ is the location parameter, and $\sigma$ is the distribution scale (set to 0 and 1 in the experiments, see Fig. 5).

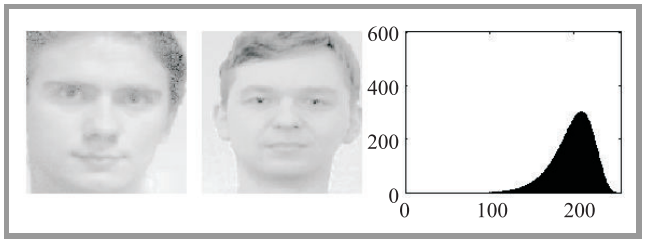



***Fig. 5.*** Two sample images with histogram after mapping the histogram to a extreme value distribution.

### *2.7. Exponential Distribution (EN)*

The exponential distribution is a commonly used distribution in reliability engineering. Density function for this distribution is defined by:

$$f(x) - \frac{1}{\beta} \exp\left(\frac{-x}{\beta}\right), \qquad (7)$$

where $\beta$ is the scale parameter. In our experiments (Fig. 6) we use the standard exponential distribution, where $\beta = 1$.

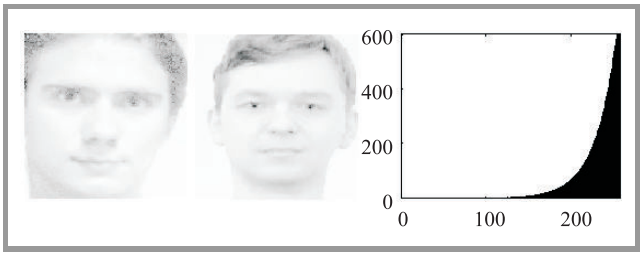



***Fig. 6.*** Two sample images with histogram after mapping the histogram to a exponential distribution.

## 3. Photometric Normalization

The second approach for illumination normalization is based on human perception theory and illumination properties.

### *3.1. Single Scale Retinex (SSR)*

In 1971 Land and McCann introduce the idea that image $I(x,y)$ is the product of two components, illumination $L(x,y)$ and reflectance $R(x,y)$ [11].

$$I(x,y) = L(x,y)\,R(x,y)\,. \qquad (8)$$

Illumination contains geometric properties of the scene (i.e., the surface normals and the light source position) and reflectance contains information about the object. Based on the assumption that the illumination varies slowly across different locations of the image and the local reflectance may change rapidly across different location, the processed illumination should be drastically reduced due to the high-pass filtering, while the reflectance after this filtering should still be very close to the original reflectance. The reflectance can be also finding by dividing the image by the low pass version if the original image, which is representing illumination components.
Land proposed a technique called retinex, which is a combination of the words retina and cortex. Its try to explain model of the human visual system. The most interesting point for illumination normalization is the assumption, that perception depends on the relative or surrounding illumination. It means that reflectance $R(x,y)$ equals the quotient of image $I(x,y)$ and the illumination $L(x,y)$ calculated by the neighborhood of $I(x,y)$. It improves the visibility of

dark object while maintaining the visual different of the light area.
Single scale retinex algorithm proposed by Jobson and Woodell [12] defines a Gaussian kernel to estimate the neighborhood illumination. Additional the logarithmic transformation is employed to compress the dynamic range. Reflectance image is takes from the form:

$$R_{SSR}(x,y) = \log I(x,y) - \log\left[F(x,y) * I(x,y)\right], \quad (9)$$

where: $*$ denotes the convolution operation and $F(x,y)$ is the surround Gaussian function.
Figure 7 shows two sample face images received from single scale retinex.



***Fig. 7.*** Two sample images received from SSR.

### *3.2. Mutli Scale Retinex (MSR)*

Rahman [13] improved previous method by estimating illumination as a combination of several weighting ($\omega_n$) Gaussian filters with different scales ($N$). Reflectance image is defined by:

$$R_{MSR}(x,y) = \sum_{n=1}^{N} \omega_n \left\{ \log I(x,y) - \log\left[F(x,y) * I(x,y)\right] \right\}. \quad (10)$$

Two sample face images received from multi scale retinex are shown in Fig. 8.



***Fig. 8.*** Two sample images received from MSR.

### *3.3. Adaptive Single Scale Retinex (ASR)*

ASR was presented by Park in [14]. The proposed method estimates illumination by iteratively convolving the input image with a $3 \times 3$ smoothing mask weighted by a coefficient via combining two measures of the illumination discontinuity at each pixel, see Fig. 9.



***Fig. 9.*** Two sample images received from ASR.

### *3.4. Homomorphic Filtering (HOMO)*

Homomorphic filtering [15] using the same properties as previous methods, that reflectance is connected with high frequency. In this case high-pass filtering is performed in frequency domain using Fourier transform. The processed image can be found by following equation:

$$I' = e^{\mathrm{Re}\left(IFT(FT(\log I) * H)\right)}, \quad (11)$$

where: $H$ is a high-pass Butterwoth's filter, $FT$ – Fourier transform, $IFT$ – inverse Fourier transform.
In Fig. 10 are shown two sample images received from homomorphic filtering.



***Fig. 10.*** Two sample images received from HOMO.

### *3.5. Single Scale Self Quotient Image (SSQ)*

The self quotient image was developed by Wang [16] in 2004 and is based on Land's human vision model. From Eq. (8) it can be derived that the reflectance is given by:

$$I(x,ky)\frac{1}{L(x,y)} = R(x,y). \quad (12)$$

Because illumination can be consider as the low frequency component then, it can be estimated as:

$$L(x,y) \approx F(x,y) * I(x,y), \quad (13)$$

with $F(x,y)$ is a low pass filter.
From Eqs. (12) and (13) the self quotient image $Q(x,y)$ is defined as:

$$Q(x,y) = \frac{I(x,y)}{F(x,y) * I(x,y)} \approx R(x,y). \quad (14)$$

Two sample face images received from single scale self quotient are shown in Fig. 11.



*Fig. 11.* Two sample images received from SSQ.

### *3.6. Multi Scale Self Quotient Image (MSQ)*

Properties of the previous $Q(x,y)$ are dependent on the kernel size of filter $F(x,y)$. If it will be to small than $Q \approx 1$ and all reflectance information will be lost. On the other hand if kernel size will be too large then will appear halo effects near edges. To avoid this problems Wang propose multi scale approach where:

$$Q(x,y) = \sum_{k=1}^{n} m_k T\{Q_k(x,y)\}, \tag{15}$$

where: $m_k$ are weighting factors, $T$ is nonlinear function and $Q_k$ are quotient images corresponding to $k$ scale.

$$Q_k(x,y) = \frac{I(x,y)}{\left(\frac{1}{N} W_k G_k\right) * I(x,y)}, \quad k = 1,\ldots,n, \tag{16}$$

where: $N$ is normalization factor, $W_k G_k$ are weighted Gaussian kernels.
Figure 12 shows two sample face images received from multi scale self quotient.



*Fig. 12.* Two sample images received from MSQ.

### *3.7. DCT-Based Normalization (DCT)*

This technique [17] is based on fact that illumination can be consider as the low frequency component. First image is transform into frequency domain using discrete cosine transform (DCT) and then some number of DCT coefficients are sets to zero. This removes some of the low-frequency information contained in the images and reduce illumination influence (see Fig. 13). Target image is obtained after applying inverse discrete cosine transform (IDCT).



*Fig. 13.* Two sample images received from DCT-based normalization.

### *3.8. Wavelet-Based Normalization (WAV)*

Next method combine two approaches based on histogram normalization and illumination low frequency properties. In the first step discrete wavelet transform is used to decompose the facial image into approximation, horizontal, vertical and diagonal components. The approximation components represents low level image components. Next equalizes the histogram of the approximation coefficients matrix. As a final step it performs an inverse wavelet transform to recover the normalized image.
In Fig. 14 are shown two sample face images received after wavelet-based normalization.



*Fig. 14.* Two sample images received after wavelet-based normalization.

## 4. Feature Discrimination

Biometric pattern verification is conceptually different from traditional class membership verification. This is involving with following terms:

1. We deal always with a subset of the whole collection of classes.
2. The number of classes used in training time of recognition system is small and usually different from classes which are recognized in exploiting time.

Since natural human centered pattern classes cannot be used in person verification biometric systems, another categorization has to be sought. It appears that differences of human features for the biometric measurements of the same person (within-class differences) and for different persons (between-class features) create a consistent categorization including two specific classes. The specificity of this two classes follows from the fact that means of these two classes are both equal to zero. Moreover, for the within-class feature variation ($\mathrm{var}_w$) could be sometimes greater than between-class feature variation ($\mathrm{var}_b$), i.e., usually the squared within-class errors are of the same magnitude as

squared between-class errors. Therefore, it is natural to look for such a linear transformation $W : R^N \rightarrow R^n$ of original measurements $x \in R^N$ (e.g., vectorized pixel matrix of face image or its 2D frequency representation) into a target feature vector $z = W^t x$ for which intra-class differences are decreased while inter-class differences are increased. This is the problem of the classical linear discriminant analysis (LDA) [18]. However from the previous works described in [19] it is already known that in case of face verification the dual linear discrimination analysis (DLDA) leads to better results than the optimization of Fisher ratio (LDA). Difference between DLDA and Fisher LDA if a way we founding optimal $W$:

$$\begin{aligned} &\text{Fisher } LDA \quad W = \arg\max_w \frac{\text{var}_b(Z)}{\text{var}_w(Z)}, \\ &\text{dual } LDA \qquad W = \arg\min_w \frac{\text{var}_w(Z)}{\text{var}_b(Z)}, \end{aligned} \tag{17}$$

where: $Z = [z_1, \ldots, z_L]$ and $L$ number of images. In our next experiments we are using LDA, DLDA and oldest method principal components analysis [20]. PCA-based face recognition method was proposed in [21] and became very popular. Using PCA method we find a subset of principal directions (principal components) in a set of the training faces. Then like in LDA we project faces into the space of these principal components and get the feature vectors.

## 5. Experimental Results

The experiments are carried out on normalized images taken from the following databases (Fig. 15):

- Altkom (80 persons 1680 images),
- Banca (52 persons 474 images),
- Valid (106 persons 1575 images),
- WUT database (143 persons 769 images).

Which gives 391 persons with 4525 images. Picture from this databases were taken in different light conditions and except the Altkom database in some time interval. According to [22] images are normalized to the size $46 \times 56$ based on fixed eye center position.

To quantify verification performance we are using receiver operating characteristic (ROC). This characteristic shows the tradeoff between two types of verification's errors false rejection error against false acceptance error. To more clarity presentation in Table 1 we show only the equal error rate (EER), which is the value were false rejection and false acceptance errors are equal. Based on the results we can conclude that except homomorphic filtering all compared methods gives verification improvements, especially using DLDA as the discriminative algorithm. The best results was conducted using multi scale quotient images where we get 38% less errors.



*Fig. 15.* Face databases – from upper Altkom, Banca, Valid, WUT.

Table 1
Performance comparison of different normalization methods

| | EER | | |
|---|---|---|---|
| | OCA | LDA | DLDA |
| ORG | 0.2138 | 0.2571 | 0.2173 |
| HQ | 0.1737 | 0.1868 | 0.166 |
| HT | 0.1935 | 0.1976 | 0.1656 |
| ND | 0.1681 | 0.1746 | 0.164 |
| LN | 0.1772 | 0.2058 | 0.1857 |
| EV | 0.1585 | 0.1857 | 0.1521 |
| EN | 0.1535 | 0.1845 | 0.165 |
| SSR | 0.1941 | 0.2178 | 0.1896 |
| MSR | 0.1987 | 0.2073 | 0.1883 |
| ASR | 0.1423 | 0.1898 | 0.1458 |
| HOMO | 0.3464 | 0.3376 | 0.26 |
| SSQ | 0.1566 | 0.1957 | 0.15 |
| MSQ | 0.1494 | 0.1865 | **0.1346** |
| DCT | 0.226 | 0.2287 | 0.1854 |
| WAV | 0.1813 | 0.17 | 0.1435 |

# 6. Conclusion

In this paper we analyze 14 illumination invariant algorithms. The performance of the presented methods were compared on database contains 4525 images of 391 persons taken in different light conditions. The results obtained from the experiments showed that the illumination preprocessing methods significantly improves the verification rate. The best results were achieved using human perception related MSQ algorithm, with 38% less verification errors compared to the same DLDA discriminant method with using not preprocessed images. Very promising seems to be combination of both analyzing approaches (histogram and photometric normalization) as a field to future work.

# References


[1] N. Kanwisher, J. McDermott, and M. Chun, "The fusiform face areas: a module in human extrastriate cortex specialized for face perception", *J. Neurosci.*, vol. 17, no. 11, pp. 4302–4311, 1997.

[2] Y. Adini, Y. Moses, and S. Ullman, "Face recognition: the problem of compensating for illumination changes", *IEEE Trans. Patt. Anal. Mach. Intell.*, vol. 19, pp. 721–732, 1997.

[3] P. J. Philips, H. Moon, S. A. Rizvi and P. J. Rauss, "The FERET evaluation methodology for face-recognition algorithms", *Trans. PAMI*, vol. 22, pp. 1090–1104, 2000.

[4] P. J. Philips, P. Grother, R. Micheals, D. M. Blackburn, E. Tabassi and M. Bone, "Face recognition vendor test 2002: evaluation report", Tech. Rep. NISTIR 6965, National Institute of Standards and Technology, 2003.

[5] P. J. Philips, P. J. Flynn, T. Scruggs, K. W. Bowyer and W. Worek, "Preliminary Face Recognition Grand Challenge Results", in *Proc. 7th Int. Conf. Autom. Face Gesture Recogn.*, Southampton, UK, 2006, pp. 15–24.

[6] P. J. Philips, W. T. Scruggs, A. J. O'Toole, P. J. Flynn, K. W. Bowyer, C. L. Schott and M. Sharpe, "FRVT 2006 and ICE 2006 large-scale results", NISTIR 7408, National Institute of Standards and Technology, 2007.

[7] A. S. Georghiades, P. N. Belhumeur and D. J. Kriegman, "From few to many: illumination cone models for face recognition under variable lighting and pose", *IEEE Trans. Patt. Anal. Mach. Intell.*, pp. 643–660, 2001.

[8] M. Vasilescu, D. Terzopoulos, "Multilinear analysis of image ensembles: tensor-faces", in *Proc. Eur. Conf. Comp. Vision ECCV 2002*, Copenhagen, Denmark, 2002, pp. 447–460.

[9] M. Abramowitz and I. A. Stegun, *Handbook of Mathematical Functions*. New York: Dover, 1964.

[10] M. Evans, N. Hastings, and B. Peacock, *Statistical Distributions*. 2nd ed., New York: Wiley, 1993.

[11] E. Land and J. McCann, "Lightiness and retinex theory", *J. Opt. Soc. America*, vol. 61, pp. 1–11, 1971.

[12] D. J. Jobson, Z. Rahman, and G. A. Woodell, "Properties and performance of a center/surround retinex", *IEEE Trans. Image Process*, vol. 6, no. 3, pp. 451–462, 1997.

[13] Z. Rahman, G. Woodell, D. Jobson, "A Comparison of the Multiscale Retinex with other Image Enhancement Techniques", in *Proc. 50th IS&T Anniv. Conf.*, Boston, USA, 1997.

[14] Y. K. Park, S. L. Park, and J. K. Kim, "Retinex method based on adaptive smoothing for illumination invariant face recognition," *Sig. Proces.*, vol. 88, no. 8, pp. 1929–1945, 2008.

[15] R. Gonzalez and R. Woods, *Digital Image Processing*. 2nd ed. Boston: Addison-Wesley Longman, 1992.

[16] H. Wang, S. Z. Li and Y. Wang, "Face recognition under varying lighting condition using self quotient image", in *Proc. 6th IEEE Int. Conf. Autom. Face Gesture Recog.*, Seoul, Korea, 2004, pp. 819–824.

[17] Štruc V., Pavešić, N. "Performance evaluation of photometric normalization techniques for illumination invariant face recognition", in *Advances in Face Image Analysis: Techniques and Technologies*, Y. J. Zhang, Ed. Hershey: IGI Global, 2010.

[18] K. Fukunaga, *Introduction to Statistical Pattern Recognition*. San Diego: Academic Press, 1992.

[19] W. Skarbek, K. Kucharski, and M. Bober, "Dual LDA for face recognition", *Fundamenta Informaticae*, vol. 61, 303–334, 2004.

[20] K. Pearson, "On lines and planes of closest fit to systems of points in space", *Philosophical Mag.*, vol. 2, no. 6, pp. 559–572, 1901.

[21] M. A. Turk and A. P. Pentland, "Face recognition using eigenfaces", in *Proc. IEEE Comp. Soc. Conf. Comp. Vision Patt. Recogn. CVPR'91*, Maui, USA, 1991, pp. 586–591.

[22] ISO/IEC 15938-3:2002 (2002).





**Mariusz Leszczyński** was born in Kozienice, Poland, on April 1979. He received B.Sc. in radiocommunication and multimedia engineering and M.Sc. in telecomunications degrees from the Warsaw University of Technology (WUT) in 2002 and 2004, respectively. Currently he is a Ph.D. student. His research interest include image segmentation, image processing and face recognition and verification.

e-mail: mariuszleszczynski@gmail.com
Institute of Radioelectronics
Warsaw University of Technology
Nowowiejska st 15/19, 00-665 Warsaw, Poland

Paper

# Face Tracking and Recognition with the Use of Particle-Filtered Local Features

Łukasz A. Stasiak and Andrzej Pacut

**Abstract—A consistent particle filtering-based framework for the purpose of parallel face tracking and recognition from video sequences is proposed. A novel approach to defining randomized, particle filtering-driven local face features for the purpose of recognition is proposed. The potential of cumulating classification decisions based on the proposed feature set definition is evaluated. By applying cumulation mechanisms to the classification results determined from single frames and with the use of particle-filtered features, good recognition rates are obtained at the minimal computational cost. The proposed framework can operate in real-time on a typical modern PC. Additionally, the application of cumulation mechanisms makes the framework resistant to brief visual distortions, such as occlusions, head rotations or face expressions. A high performance is also obtained on low resolution images (video frames). Since the framework is based on the particle filtering principle, it is easily tunable to various application requirements (security level, hardware constraints).**

***Keywords—biometrics, face recognition, particle filtering, video analysis.***

## 1. Introduction

### *1.1. Problem Statement*

We consider an identification scenario with the use of a video input signal. It is assumed that individuals entering the controlled zone cannot be effectively tracked over the entire stay-in-the-zone period, e.g., due to the large number of people walking along the main routes or due to the complicated topography of the zone. However, multiple identity recognitions with the use of local cameras installed in various locations around the controlled zone are possible. Cameras are installed in a way that enables frontal capture of subjects' faces, i.e., at the average height of a human and in specific places where frontal face images can be captured. Such *places* could be near paintings in galleries, supermarket shelves, shop windows, advertising posters, mirrors, elevator exits, escalators, at the ends of narrow corridors, etc. Based on multiple identifications, a rough track of an individual's (sequence of visited places) can be retrieved or an alarm can be raised when the selected individual enters a prohibited area in the controlled zone. It is assumed that the controlled zone is relatively small, so that the number of individuals to identify simultaneously is limited. Additionally, we assume that individuals who enter the controlled zone were previously enrolled to the system or are enrolled on entry. Consequently, a closed-set identification scenario is considered.

The presented usage scenario may be primarily regarded as *tracking by identifying* and is similar to the usage scenario of the Face Cataloger from IBM [1], [2]. Both solutions are used to answer the question *who is where?* within the controlled zone. Information gathered from tracking by identifying can be useful for warehouse, museum or gallery management, since it permits assessment of the attractiveness of the presented items. The application can also be effectively used to control higher security regions within the controlled zone, particularly, when two groups of subjects are considered. such as, e.g., employees and visitors. Unlike in the IBM's Face Cataloger scenario, we assume utilization of video-specific information not only for the purpose of tracking but also for the purpose of identity recognition. Additionally, the Face Cataloger utilizes a badge identification system for the purpose of subject identification at the entrance. We do not utilize any external systems and we assume low computational requirements for the proposed framework.

### *1.2. Particle Filtering*

Particle filtering is used as a basis for the proposed framework and therefore it is shortly presented here. By definition, particle filters are sequential Monte Carlo methods based upon point mass representations (*particles*) of probability densities. Such representations can be applied to any state space model and generalize the traditional Kalman filtering methods [3]. The key idea of the Monte Carlo methods is to approximate a difficult analytical problem by a much simpler problem represented by a statistical sample [4]. The stochastic nature of the Monte Carlo simulation in computer environment is achieved by the use of pseudo-random number generators. The Monte Carlo simulation is considered to be one of the most influential and landmark algorithms of the 20th century [5].

An implementation of the particle filtering principle – particularly well known in the computer vision research area – is the Condensation algorithm of Isard and Blake [6]. The Condensation is also utilized within our proposed framework. For the purpose of tracking, it requires at least two models to be defined, namely the object model (usually including object's dynamics model) and the observation model. Selection of the models is essential for performance of the whole solution.

Probably the most practically useful property of the particle filters is that they do not require any functional assumptions (linearity, Gaussianity, unimodality) about the densities. Initial state density $p(\vec{x}_0)$ must however be given, i.e., some initialization (e.g., initial face detection) must be done. Common drawback of the particle filtering techniques is a *degeneracy problem*, which consists in concentration of most of the weight on a relatively small subset of particles [3], [7], [8]. Full discussion of the particle filtering and degeneracy problem can be found in the cited literature.

Within the proposed framework, the particle filtering principle is utilized for the purpose of face tracking with the use of local face features (called primary face features), defined as small face patches. Color distribution within these features is analyzed for the purpose of tracking. For the purpose of recognition, the frequency analysis of the features is run. As a result, two types of secondary face features are obtained from the primary face features, namely color distribution- and frequency analysis-based secondary features. Definitions of the secondary face features as well as comments on selecting the primary features are given in the next section.

# 2. Local Face Features

## *2.1. Primary Face Features*

Typically, particle filtering-based trackers assume that particle model is very similar (or identical) to the object model. In the context of the proposed framework, it would mean that particles are face candidate locations, and thus the observation area associated with each particle is of a size of the face candidate [9], [10], [11]. The final object state vector would then be calculated as mean value of all the particles. However, processing such *big* particles is usually computationally expensive. Therefore we utilize *small* particles (of a size between $10 \times 10$ and $40 \times 40$ pixels), which refer to local face patches and are understood as primary face features. Such solution results in computational time savings per each particle at the cost of the more complicated procedure of estimation of the whole face area. Namely, face area cannot be straightforwardly determined as the mean value of all particles. Instead, the distribution of all particles in image space must be analyzed to obtain rough face area estimation. For the case of many faces in the scene, this must involve automatic clustering of the particle set. Finer face area estimations are retrieved with the use of *dust filtering* procedure, which we proposed previously [12]. The dust filtering consists in classifying single pixels as skin or non-skin pixels only within the initial, roughly estimated face areas. The enhanced results of the single-pixel-classifications are then used to determine face areas more precisely. Details of the procedure of rough face area estimation from the particle distribution and of the dust filtering procedure can be found in [12].

Primary face features, i.e., particles are resampled accordingly to the Condensation schema. Each $n$th primary feature has assigned a weight $\pi_n$, which is used for the purpose of resampling. Additionally, random diffusion and deterministic drift are applied to steer the particle motion in the image space. For the purpose of determining the drift, a tracking history of the normalized face area locations ($\hat{R}(t)$, where $t < 0$) is stored. Predicting a new location of the face area $\hat{R}(0)$ is based on a simple model, namely

$$\hat{R}(0) = \hat{R}(-1) + (\hat{R}(-1) - \hat{R}(-2)) + \varepsilon(0), \tag{1}$$

where $\varepsilon(t)$ is an i.i.d. zero-mean noise.

## *2.2. Secondary Face Features for Tracking*

For the purpose of face tracking, primary face features (particles) must be resampled. This is done with the use of color distribution features retrieved from the primary face features and compared to a universal skin color model. Skin color is a low level feature, which appears to be highly discriminative and computationally fast. It is easy to understand and robust to geometrical changes. As many research studies have shown, the skin tones of different ethnical groups differ mainly in their intensity values [13], [14], [15], [16], [17], being clustered in chrominance values. This makes it possible to use a universal skin color model to represent all skin types. Main disadvantage of color features is that cameras are not able to distinguish changes of the actual surface colors from changes caused by varying illumination. Consequently, illumination is the most influential factor, which changes color values recorded by a camera. Lighting compensation techniques have been proposed to reduce this problem [13], [18], [16].

In the proposed framework, color is the main cue used for the purpose of tracking by particle filtering and for the purpose of quick face normalization by means of the previously proposed dust filtering method [12]. Due to utilization of color features, real time processing can be achieved with the use of a typical modern PC. We represent color distributions of the local face features as $64 \times 64$ bin hue-saturation (HS) histograms of HSV colorspace. The V-channel (value/intensity) is ignored. We compare the HS histograms of local patches to the reference skin color model with the use of the Bhattacharyya distance $d_n = d_{Bhatt}[p_n, q]$, where $p_n$ is the HS histogram determined from the $n$th particle and $q$ is the reference color histogram. The Bhattacharyya distance is defined with the use of the Bhattacharyya coefficient $\rho[p,q]$, which is a similarity measure between two color distributions $p(u)$ and $q(u)$, namely

$$\rho[p,q] = \int \sqrt{p(u)d(u)}du. \tag{2}$$

In the context of discrete densities represented by histograms $p = \{p^{(u)}\}_{u=1\ldots64\times64}$ and $q = \{q^{(u)}\}_{u=1\ldots64\times64}$, the Bhattacharyya coefficient is defined as

$$\rho[p,q] = \sum_{u=1}^{64\times64} \sqrt{p^{(u)}d^{(u)}}. \tag{3}$$

For two identical normalized histograms we obtain $\rho = 1$, indicating the perfect match. The Bhattacharyya distance $d_{Bhatt}[p,q]$ is then defined as

$$d_{Bhatt}[p,q] = 1 - \rho[p,q]. \tag{4}$$

Particles (primary features) are then re-weighted accordingly to the Condensation schema: the new weight $\pi_n$ of $n$th particle is calculated as

$$\pi_n = \exp(-\lambda\, d_n^2). \tag{5}$$

We use value of $\lambda = 20$ as suggested in [9], [19].

### *2.3. Secondary Face Features for Recognition*

Particles which are found to be located within the dust-filtered face area, are then used for retrieval of secondary features for the purpose of identity recognition. We define these secondary features as the discrete cosine transform (DCT) coefficients of the respective primary features (particles). Whereas the secondary features for the purpose of tracking where retrieved from H- and S-channels of HSV colorspace, the secondary features for the purpose of recognition are calculated with the use of V-channel. We selected the DCT coding mainly due to its ease of application, known successful applications to face recognition [20], [21], and the potential of introducing identity recognition mechanisms into the existing compression schemes, which already utilize the DCT commonly.

Having precisely estimated the face area, relative location of each primary feature within the face area can be retrieved and thus the corresponding feature in the template can be found and compared against a given feature. This means that – for the purpose of recognition – the features are valid only in combination with their relative location within face area. The combination of frequency and location properties is similar to other existing face recognition approaches, where localization data is used in combination with some transformed local features, e.g., elastic bunch graph matching (EBGM) [22] or active shape models (ASM) [23]. However, in the previously known methods, features to be detected are precisely defined and the feature detection is the most computationally expensive part of these algorithms. In our proposed approach, we introduce primary features into the particle filtering framework. Consequently, the costly detection is skipped, *randomized* feature locations are utilized and feature sets used for the purpose of recognition differ from video frame to video frame. Such feature sets can be easily processed in real time, still providing good image exploration. Furthermore, such definition makes it possible to employ low resolution face images in which accurate detection of facial landmarks is hardly achievable [24].

The main drawback of the proposed feature set definition is that the set of stored template features should be extensive. Since any feature locations (as a result of particle filtering) can be achieved within the actually processed image, features for all possible locations within the templates should be pre-calculated. They might also be calculated on-demand, but this would lead to the high extension of the processing time. Furthermore, the face areas being processed should be well aligned with the template images, so that the actually corresponding features in the image and in the template can be compared (precise alignment or feature detection is an issue for all face recognition methods). For the purpose of testing the proposed face feature set definition approach, we utilized fast face normalization with the use of the dust filtering. Additionally, sizes of faces in test sequences were compliant with sizes of template face images. As presented in next sections, this provided good recognition rates at a low computational cost. Application of more precise normalization procedures is expected to further improve the recognition quality (at a higher computational cost).

## 3. Recognition from Video

A video sequence provides more information in comparison to a still face image. This information is distributed over video frames, which have some relation to the *real time* that passes during the video recording. Such distributed information can be cumulated in order to provide a stronger decision than a single-frame (or still image) based decision. The sequential hypothesis testing paradigm may be applied for the purpose of identity recognition over a sequence. The initial *weak* classifications become *stronger* when new video frames become available. An input to the recognition module are dust-filtered face areas from the tracking module. The tracking module provides consistency of the track, i.e., it assures that consecutive faces passed to the recognition module are correctly labeled as belonging to a given individual (though the individual's identity remains *unknown* to the tracking module).

As described above, secondary features from each frame are retrieved and compared to the respective secondary features of the stored templates. For the purpose of comparison, the DCT secondary features are zig-zag reorganized [25] to form feature vectors. The feature vectors can then be compared directly with the use of various distance metrics. We evaluated $L_1$, $L_2$ and $L_\infty$ distances and $L_2$ resulted in the best performance. Therefore, it is used for the purpose of performance evaluation further in this article. The distance between DCT feature vectors is expressed as

$$d_{L_2}(\vec{x},\vec{y}) \quad = \quad \sqrt{\sum_{i=1}^{n}(x_i - y_i)^2}, \tag{6}$$

where $\vec{x} = [x_1,\ldots,x_n]^T$ and $\vec{y} = [y_1,\ldots,y_n]^T$ are DCT-transformed feature vectors to be compared, and $n+1$ is the length of the feature vector (including 0 indexed coefficient). The DCT coefficient indexed 0 represents the average of the image patch and therefore is ignored for the purpose of recognition.

### *3.1. Cumulation of Classification Results*

A result of comparing the particle-related local features is a set of distance measures between those features (retrieved

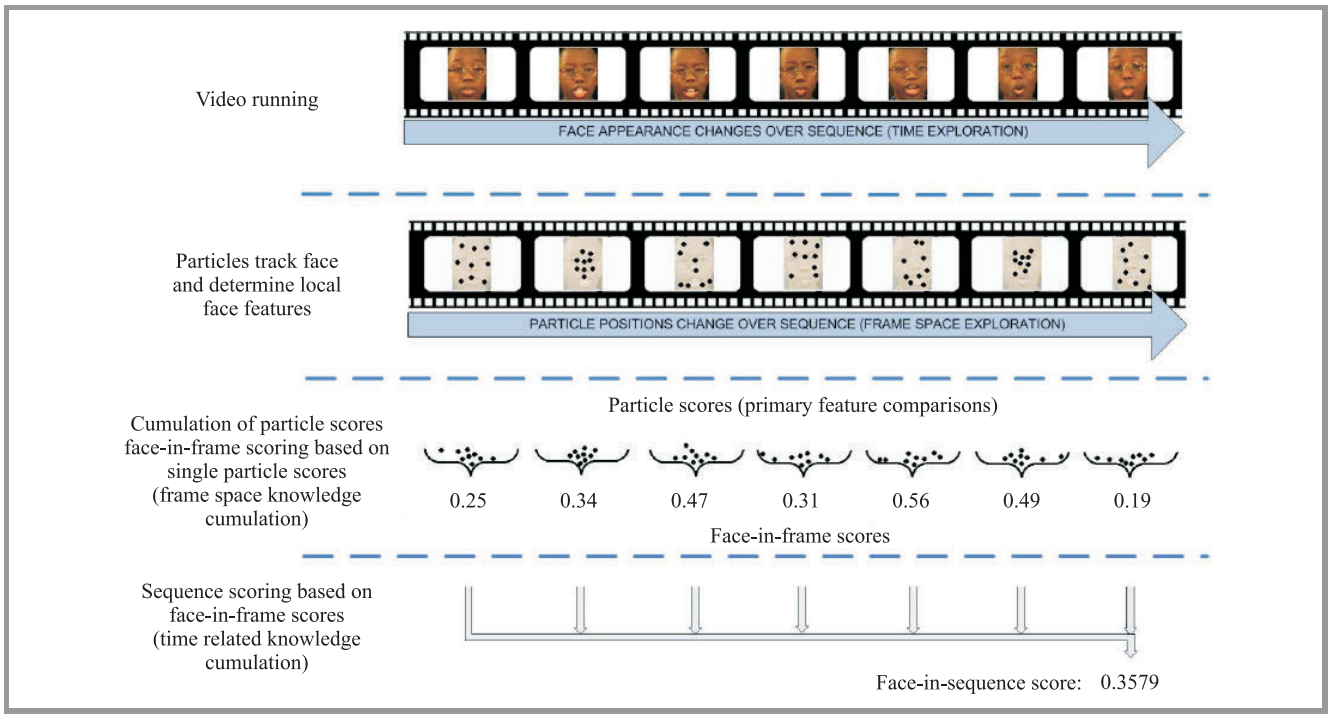



***Fig. 1.*** Face image space and time exploration achieved by applying particle filtering to a talking head video sequence. Randomized positions of particles in each frame provide image space exploration and determine a set of face features used for the purpose of recognition. Single particle-related distances from each frame are cumulated to give a face-in-frame distance. Face-in-frame distances are cumulated to obtain a face-in-sequence distance. Sample video frames used for the presented processing were taken from [26].

from a given face image) and corresponding features of a template. These measures can be cumulated in order to provide a distance between the whole face area and the template, later referred to as the *face-in-frame distance*. The face-in-frame distance is simply calculated as an average of particle-related distance measures taken for only those particles which fall within the normalized face area, namely

$$D(F,T) = \text{avg}\left(\text{all } d_{L_2}(\vec{x},\vec{x_T}) : x \in F\right), \tag{7}$$

where $\vec{x_T}$ is a DCT feature vector in template $T$ corresponding to DCT feature vector $\vec{x}$ retrieved from a given face image $F$, and $x$ is a particle (primary feature) from which $\vec{x}$ is determined.

Face-in-frame distances are then cumulated for the purpose of comparing the whole sequences of faces against given templates, resulting in *face-in-sequence distances*. Consequently, it can be concluded that *distance cumulation* is applied at two different levels:

- Space level: cumulation of distance values (scores) of particle-related local features (distributed over a face image) in order to obtain a single face-in-frame distance.
- Time level: cumulation of face-in-frame distance values (distributed over a video sequence) in order to obtain a face-in-sequence distance.

Comparisons of local features can also be understood as weak local classifications, which are then cumulated to provide stronger frame-related classifications (face-in-frame distances). Cumulation of frame-related distances provides yet stronger classification of video sequences (face-in-sequence distances). Face image space exploration is a result of the integrated framework by which the probabilistic nature of the particle filtering-based tracking is passed to the recognition task. Consequently, more face image space can be explored and tested for the purpose of recognition without significantly increasing the processing burden. Two levels of distance cumulation within the proposed framework are depicted in Fig. 1. The face-in-sequence distances are then utilized for the purpose of recognition, which is done by building a ranking of identities. Since we consider a closed set scenario, all stored templates are compared (scanned) against the actually processed face image and the ranking is determined.

Integration of the tracking and recognition within one framework brings additional advantage, namely identity cue can be used for the purpose of tracking corrections. In case of processing multiple faces in the scene, some face-to-track assignment conflicts occur. A feedback from the recognition module can additionally support conflict resolution mechanisms and thus improve tracking accuracy. We described this idea in [27].

### *3.2. Multi-Image Template*

Quality of a template has a great influence on the overall recognition performance. The template quality can be improved by using more than one image for template creation.

The simple template extension technique, which directly uses several images to create a multi-image template, can be used: the query face image $F$ is compared to each image of the multi-image template $T^*$ and the best match between the query image and template image is selected as the actual face to multi-image template distance $D^*$, namely, for $K$-image multi-image template

$$D^*(F,T^*) = \min_{1\le k\le K} D(F,T_k^*), \tag{8}$$

where $T_k^*$ is $k$th image of the multi-image template and $D$ is the face-in-frame distance calculated between the given face and a single template image. Performance improvement achieved by extending the template representation and with the use of the simple comparison procedure is analyzed in the following sections.

# 4. Definition of the Framework and Its Parameters

### *4.1. Testing Environment*

For the purpose of testing the performance of the proposed framework, we used 55 video sequences of the Open Video Project (OVP) [26]. The downloaded sequences are *talking heads* videos of different individuals. The length of sequences varies from 851 to 8265 frames. Talking head videos of 35 different individuals have been extracted, resulting in 100 to 1187 talking head frames per individual (340 talking head frames per individual on average). Talking head sequences contain frontal and almost-to-frontal shots (less than 30 degrees profile) without any extra constraints on the individual's head motion or facial expression. Changes in captured head size due to camera zoom or head motions are present. A few sample frames from the test video sequences are presented in Fig. 2.



***Fig. 2.*** Variations in test video sequences from Open Video Project [26]. Test videos are frontal and almost-to-frontal talking head sequences without any additional constraints on the individual's head motion and facial expression. Variations in head size, video quality and background type are noticeable.

It is to notice that – since we consider a sequential recognition from video sequences – the beneath reported cumulative match characteristics (CMC) are *cumulated* not only over identities but also over sequence time. For example, the 80% 1-rank identification rate means that the actual subject identity was returned in the first position in the ranking for 80% of the time (video frames) in all test sequences. Various aspects of the proposed framework have been evaluated and the results are presented beneath.

### *4.2. Optimal Feature Vectors*

Selection of DCT coefficients for the purpose of building feature vectors for identity recognition can influence the overall recognition performance. The DCT coefficients selection is related to the question of how much identity-specific information is carried by various signal frequencies. Ekenel and Stiefelhagen [28] showed that selection of the number of coefficients influences performance and that extending this number over a certain limit does not significantly improve performance. Sanderson et al. [29] showed that increasing the dimensionality from 15 to 21 provides only a small recognition improvement, while it significantly increases the computational requirements.

In order to find an optimal set of DCT coefficients we run several tests for the closed-set scenario on the full test database, but with the use of different feature vector definitions. The testing was done with the use of the whole proposed framework. Although all framework parameters will only be introduced hereinafter, we think it is reasonable to present the results regarding selection of the DCT coefficients here. For the purpose of this testing fixed particle positions were used, so that the results of different test runs (for different feature vector definitions) are comparable. The identification rates obtained with the use of differently defined feature vectors are gathered in Table 1 and depicted in Fig. 3.

Table 1
Influence of the number of DCT coefficients on the identification rates obtained for full database testing in the closed-set scenario

| Indexes of selected DCT coefficients | Identification rate [%] | | |
|---|---|---|---|
| | 1-rank | 5-rank | 10-rank |
| 1-5 | 66.64 | 79.49 | 85.66 |
| 1-10 | 65.15 | 79.15 | 86.21 |
| 1-15 | 64.16 | 78.92 | 85.70 |
| 1-20 | 62.16 | 77.50 | 84.74 |
| 1-25 | 60.48 | 76.54 | 83.89 |
| 6-15 | 41.24 | 60.27 | 74.74 |
| 16-25 | 22.70 | 44.58 | 59.92 |

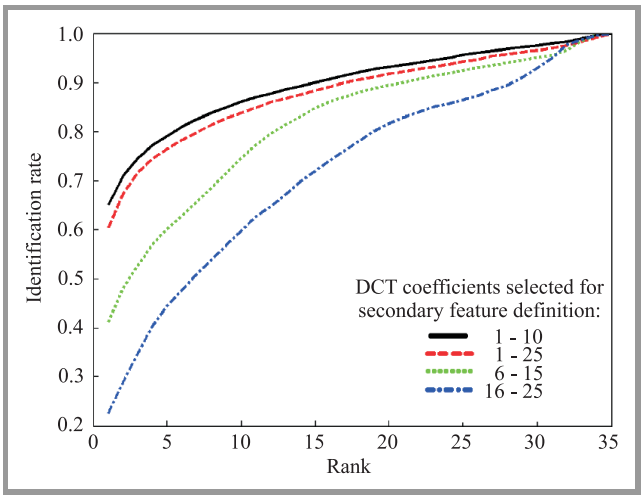



***Fig. 3.*** Influence of the feature vector definition on CMCs obtained from full database testing in the closed set scenario.

The obtained results show that lower frequencies (low coefficient indexes) contain most of the identity-related information. Extending the feature vector by higher frequencies does not improve the performance significantly and excluding lower frequencies drastically reduces performance quality. It can also be observed that differences between the cases of 1-5, 1-10, 1-15, 1-20 and 1-25 are minor. However, in the literature [30], [28], [29], it is rare that as few as five coefficients are suggested. We finally selected DCT coefficients 1-10 to be used as the local feature representation in our framework. Such a definition results in good recognition performance and keeps the representation compact and is used for the evaluation presented in the following sections. It is to comment that before calculating the DCT coefficients, contrast of the whole face area (in a given frame and in a template) is enhanced by the histogram equalization technique, which improves system performance [28]. No other illumination compensation techniques are applied.

### *4.3. Face-in-Sequence Scoring and Classification*

Having obtained face-in-frame distances against a set of templates, a cumulated distance for the video sequence, i.e., face-in-sequence distance, can be obtained. Cumulation of the distances can be done in the following ways:

- Fixed lag cumulation. Face-in-sequence distance is based on the distances of $n$ previous frames (cumulation lag = $n$). Results are available at any $i$th frame (time) of the sequence, such as $i > n$.
- Fixed point (growing lag) cumulation. Face-in-sequence distance is based on all previous frames. Results are available at any frame of the sequence.
- Adaptive lag cumulation. Face-in-sequence distance is based on the varying number of previous frames. Results may be available at any frame of the sequence (but with different *strength*) or when a given minimal number of frames is available.

In all cases *previous frames* must be understood as face areas retrieved from previous frames and with reference to a given track.

In order to achieve good classification when large lag values are used (a high number of previous frames is considered), it must be ensured that tracks are consistent, i.e., the subject-track pairs are not swapped during tracking. Otherwise, classification of a sequence (track) containing face images of various subjects will be dominated by the prevailing subject. In consequence, application of the fixed point scoring is not appropriate for high security scenarios and should rather be applied to other non-security scenarios, e.g., for the purpose of video summarization [31]. For security applications, utilization of the fixed or adaptive lag is more appropriate. The lag value does not only influence classification strength, but it also defines response delay (e.g., updating identity classification result or raising an alarm), when the subject identity within a track changes (which, first of all, may be a result of tracking error). Response delay can also be understood as a resistance to brief misclassifications: the higher the lag value, the more the duration of the misclassification (e.g., caused by occlusion) will not affect classification result. The trade-off between a quick response to identity change and the resistance to misclassifications is actually the problem of tuning a biometric system to achieve optimal false acceptance and false rejection rates (FAR and FRR). An optimal solution does not seem to exist in general and should be found with respect to application specific requirements, such as security level, environmental conditions, input video quality, usability requirements, hardware requirements (e.g., memory requirements for storing previous frame distances). It may be concluded that optimally the lag value should change within some predefined range $[\text{lag}_{\min}, \text{lag}_{\max}]$. The value of $\text{lag}_{\min}$ should be derived from the required minimal classification strength and misclassification-resistance, whereas $\text{lag}_{\max}$ should be derived from maximal acceptable response delay.

Level of the face-in-frame distances is dependent on the input frame conditions, such as e.g. head rotation or frame quality. It means, that though the ranking of identities can be preserved between the frames, the absolute level of distance values can vary strongly and influence the cumulated distance. Therefore, for building the cumulated rankings, a distance value normalization is required. For this purpose we utilize min-max normalization of face-in-frame distances, namely

$$D'_{FT} = \frac{D_{FT} - D_{F_{\min}}}{D_{F_{\max}} - D_{F_{\min}}}, \tag{9}$$

where: $D_{FT} = D(F,T)$ is the calculated (unnormalized) distance between frame $F$ and template $T$, $D'_{FT}$ is the normalized frame-to-template distance, $D_{F_{\min}}$ and $D_{F_{\max}}$ are respectively minimal and maximal distances between the given frame $F$ and any template from the template set. As a result of normalization the values of $D'_{FT}$ within the range of $[0,1]$ are obtained.

### 4.4. Occurrence of Classification Errors

During evaluation we have observed that erroneous or *weak* classifications are usually a result of distortions in video sequence. In such cases most of the templates seem to be *almost equally* distant to the given frame. In other words, the given frame is not particularly similar to any given template. On the other hand, if the recognition is strong, there are usually only a few *good* matches being clearly separated from others. This effect is illustrated in Fig. 4.

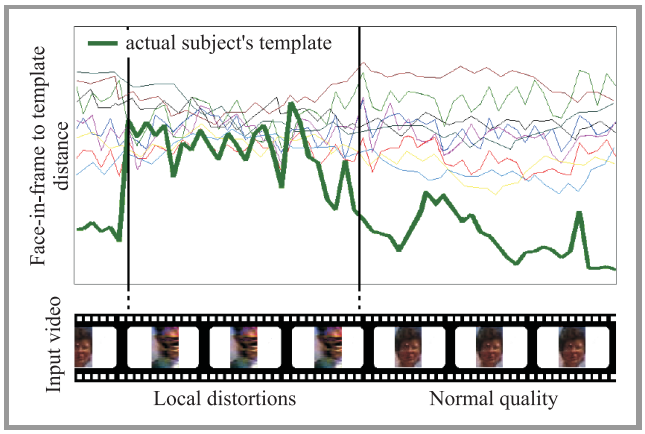



***Fig. 4.*** Local distortions in video cause face-in-frame distance values to *gather* around an average value. In good quality frames, best-matches are clearly separated from other matches. Here distortions are caused by digital storage medium errors.

Based on this observation it may be concluded that strong classifications are possible when the subject's face in the video can be seen well. Since *bad* matches are similarly bad for all the templates, distance value cumulation should enable the brief erroneous classifications to be overcome when a longer period of time is considered.

## 5. Distance Value Cumulation Mechanisms

### 5.1. Utilizing Video Sequentiality

By video sequentiality we mean the high dependency of a video frame on previous frames. It may be informally said that almost every video frame is very similar to the preceding frame. In the identity recognition context the sequentiality may be utilized to overcome brief misclassifications, since an identity recognized in a given frame is very likely to have also appeared in previous frames. This property can be utilized by applying a cumulation mechanism, i.e., ranking the identities of each video frame with respect to previous frames. A simple approach is to use the sum of face-in-frame distances on a lag of $k$ previous frames as a cumulated distance. Namely, the cumulated distance $D_{cum}$ for the $j$th frame $F_j$ of the sequence against a given template $T$ on a lag of $k$ previous frames is defined as:

$$D_{cum}(F_j,T) = \frac{1}{k}\sum_{i=0}^{k-1} D(F_{j-i},T). \tag{10}$$

Implementing the cumulation mechanism results in a higher recognition rate (as calculated per every video frame). Similarly, extending template representation from one image to a three image multi-image template increases recognition rates. The observed performance improvement is summarized in Table 2 and depicted in Fig. 5. It is observed that increasing the template quality improves the performance more than introducing cumulation mechanism only.

Table 2
Identity recognition improvement obtained as a result of introducing cumulation mechanism (with lag $L = 10$) and extending template representation: 1-rank identification rates presented

| Solution | 1-rank ident. rate [%] |
|---|---|
| One-image templates, no cumulation | 34 |
| One-image templates with cumulation | 41 |
| Three-image templates, no cumulation | 50 |
| Three-image templates with cumulation | 61 |

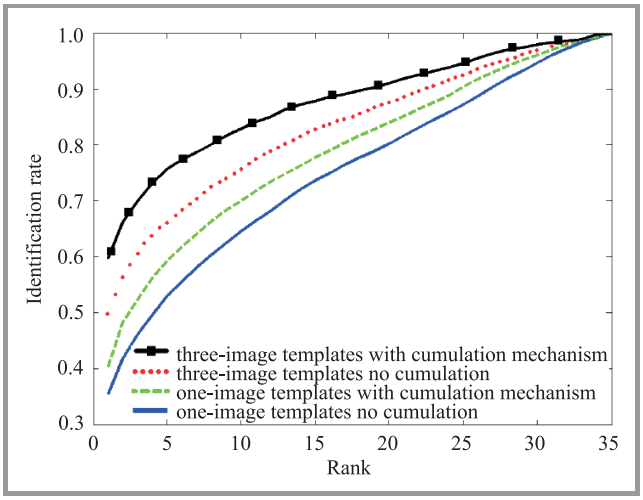



***Fig. 5.*** Identity recognition improvement obtained as a result of introducing cumulation mechanism (with lag $L = 10$) and extending template representation: CMCs depicted.

### 5.2. Cumulation Schemes

Fusion of face-in-frame distances can involve a simple sum rule or can be combined with extra distance value transformation. Let us denote $D$ as the original distance determined by scoring a given face-in-frame and $D'$ as the distance after transformation. We propose and evaluate the following transformation methods:

- linear mapping $D' = D$, (11)
- square root transformation $D' = \sqrt{D}$, (12)
- quadratic transformation $D' = D^2$. (13)

Linear mapping is used by a basic sum rule approach: it corresponds to a simple summation of all frame distances over the sequence. Square root and quadratic transformations are meant to emphasize Eq. (12) or de-emphasize Eq. (13) differences between *similarly good* matches. As described above, all face-in-frame distances are min-max normalized before applying transformations of Eqs. (11), (12), (13).

CMCs for different fusion approaches with the cumulation lag of $L = 10$ and three-image template representations are depicted in Fig. 6. The best performer, namely the square root fusion approach, achieved **1-rank** identification rate of **61%**, **5-rank** rate of **77%** and **10-rank** rate of **85%**. The simple sum rule (linear mapping) performed almost equally well.

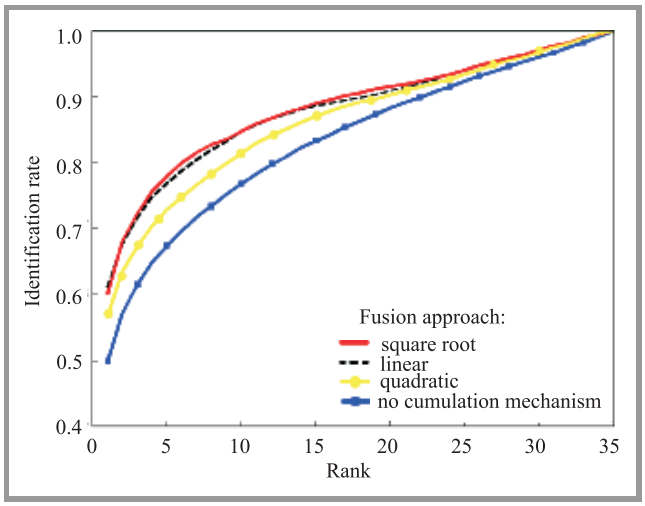



***Fig. 6.*** CMCs obtained with the use of different fusion schemes, cumulation lag of $L = 10$ frames and three-image template representations. Application of cumulation mechanisms significantly improves performance in comparison to the non-cumulation approach.

It is again observed that application of any cumulation mechanism improves the recognition performance significantly in comparison to single frame based identification. It is also concluded that the simplest summation rule may be optimal solution, since differences in performance quality between simple rule and square root fusion are minor.

### *5.3. Definition of the Cumulation Lag and the Influence of Input Video Frequency*

During evaluation it was discovered that defining the cumulation lag by the number of frames was confusing – it is rather the time period which should be defined as a cumulation lag. Time-based definition of the cumulation lag remains independent of video frequency, unlike definition by number of frames. Time-based definition describes the admissible range of appearance changes better than does the number of frames. We observed that given a time defined cumulation lag, changes in recognition performance caused by different input video frequencies are minor and can be disregarded. In other words, the most influential factor on identification rates is the period of distortions in video in relation to the cumulation lag (i.e., how long can distortions maximally last), and this property is easier to describe by defining the cumulation lag in time units.

### *5.4. Optimal Lag Value Selection*

Selection of the lag value $L$ determines the recognition performance of the system and its response-delay to the *identity change* of the observed individual. Choice of an optimal lag value is application-specific but a general policy can be defined:

- For high security the lag value should be low to provide a quick response.
- For high user-friendliness the lag value should be higher to minimize the number of false rejections.

To evaluate influence of the lag $L$ on the overall performance, we evaluated the proposed framework with the use of various cumulation lag values. Lag $L = 1$ is equivalent to the case with no cumulation mechanism. For this configuration the 1-rank identification rate of 50% was obtained, which confirms the *weak* nature of the used frame classifier. For a cumulation lag equal to 8 s (200 frames) **1-rank** identification rate of **81%**, **5-rank** identification rate of **90%** and **10-rank** identification rate of **93%** were obtained. CMCs for various cumulation lag values with the use of the square root fusion approach Eq. (12) are depicted in Fig. 7.

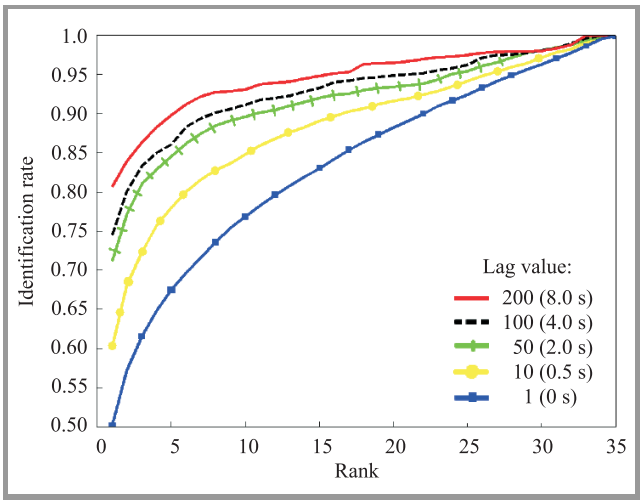



***Fig. 7.*** CMCs for the square root fusion approach and various cumulation lag values $L$. Increasing the $L$ value improves recognition performance. Graphs for 25 fps sequences.

The obtained results confirm expectation that extending the cumulation lag improves the recognition performance of the framework. It is also observed that cumulating only a few previous frame distances improves performance significantly. The more frame distances are already cumulated, the less influence on performance is observed by adding new distances. It is however to remember, that no track

swapping errors were considered during this test: i.e., an indexed track have never *skipped* to another individual (assumption of tracks consistency). In the target application, as mentioned previously, not only recognition rate, but also response delay must be considered when selecting the optimal lag value.

### 5.5. Fixed Point Approach

Increasing the lag improves recognition performance. As a result, it may be expected that the fixed point (growing lag) cumulation approach would provide higher identification rates. The drawbacks of this approach, as described previously, include the risk of high response delay when a tracking error occurs. For the purpose of testing the progress of identification rate in fixed point approach we extracted a subset of 100-frame long sub-sequences from the testing database. Rank 1 and rank 5 recognition rate changes observed over frames of the extracted video sequences are depicted in Fig. 8.

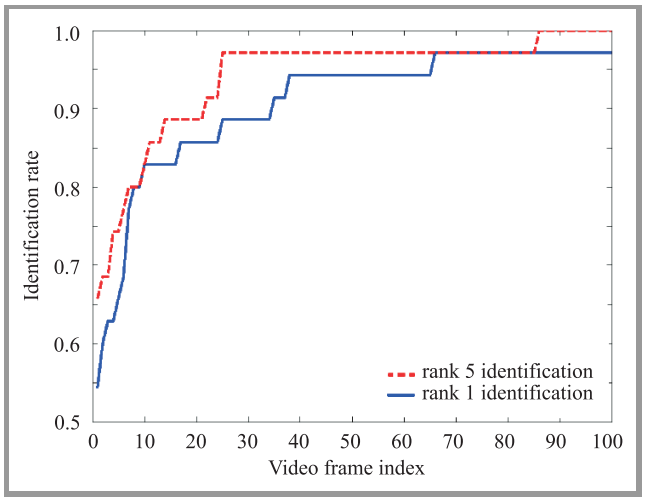



***Fig. 8.*** Rank 1 and rank 5 identification rates over 100-frame sequences with the fixed point cumulation approach (all previous frame distances considered).

From the results it can be concluded that extending the cumulation lag does not need to be indefinite. In early processing steps, as the cumulation lag value is low, the recognition performance increases rapidly with new frames. Later on however, as many previous frames are already considered, the cumulated score becomes saturated. This leads again to the idea of utilizing an adaptive lag as the most practical approach. It should be remembered that this test was run with sub-sequences of the original testing database, therefore the reported results vary from those presented previously for the whole database.

### 5.6. Processing Times and Further Enhancements

The proposed framework was tested on an Intel Core 2 Duo E6750 computer, 2.67 GHz with 2.00 GB RAM. Processing times were tested in a whole framework combined with a detection module, which is not described in this paper and which initializes the tracking process. The detection, tracking and normalization functionalities were implemented in the Visual C++ environment and with the use of OpenCV 1.0 library [32]. The recognition module was implemented in Matlab 6.5. No special code optimization was applied.

Processing times were calculated for the following configuration parameters and environmental conditions: input video frame of size $320 \times 240$ pixels, particles of size $8 \times 8$ pixels, normalized face areas of size $64 \times 64$ pixels, face detection by Haar-like face detector [33], generic skin color model represented by $64 \times 64$ hue-saturation histogram, face area normalization by means of dust filtering with pixel-step equal to 4, 10-subject closed set identification scenario, three-image template representations, one person in the scene.

In the basic configuration, 335 ms per frame were needed to run all the tasks of detection, tracking and recognition with the use of 50 particles. Face detection was the most time consuming task – tracking and recognition itself needed 85 ms per frame on average. Frame preprocessing, which involved data retrieval from the video buffer and transformation from RGB to HSV colorspace, required 3 ms of the processing time.

Some process optimizations, which should further reduce computational requirements, are possible. First of all, face-in-frame to template comparisons are currently realized by linear scanning of the whole set of templates. Therefore, the recognition processing time is proportional to the number of subjects in the database. Effective indexing and sorting techniques are subject to further research with the aim of ensuring that a quick search during identification can be carried out.

Additionally, average cost of face detection can be reduced by minimizing the frequency of running the detection process. The detection can be, for example, triggered by an external event, such as door open etc. Furthermore, tracking and recognition do not need to operate on every single frame, but can wait until detection is finished. This would result in the longest processing time per frame of 250 ms occurring during the detection phase. After detection, the whole frame processing would require **85 ms** (tracking and recognition only) – this means that a speed of **11 fps** can be achieved and 5 fps is regarded as sufficient for handling *normal* head motions [34]. For the purpose of detection in a testing environment the frontal face detector and profile face detector were utilized. Profile detection involved horizontal mirroring of the whole frame. Reducing detection to frontal faces only can save 190 ms of processing time.

Consequently, simultaneous frontal face detection, tracking and recognition in every frame results in a processing speed of more than **6 fps**. A lower accuracy in detection may be accepted for many applications, in particular due to utilizing video as an input signal: in many practical cases it may be reasonable to invoke detection less accurately (i.e., detecting frontal faces only), but more

frequently. Detection by tracking approach [12], [35] can also be utilized to further reduce processing times. Achieving the optimal architecture of the modules (sub-processes) is, however, a non-trivial and application-specific issue.

The advantages of using the distributed hardware architecture should be considered for the proposed framework. Due to construction of sub-processes, detection can be easily realized by other processing units than tracking and recognition. Separating tracking and recognition between different processing units is also possible. In the distributed environment each process would run independently and retrieve required data from the supporting process (e.g., tracking from detection or recognition from tracking). A further degree of parallelization could be achieved by computing the DCT for various particles on separate units.

# 6. Conclusions

We proposed and evaluated a consistent particle filtering-based framework for face tracking and recognition from video. Presented results proved that sequentiality of the video signal can be effectively used for the purpose of increasing identification rates. This is achieved by applying distance cumulation mechanisms. Even utilization of weak classifiers, which result in the 1-rank identification rate of 50% when no cumulation mechanism is applied, can lead to 1-rank identification rate of 81% when a cumulation lag of 8 s (200 frames) is used. The strength of the classification increases as more frame distances are collected for the purpose of classification. The classification result is available at any video frame, so it can be obtained even at early steps of the sequence processing, though with lower accuracy. The number of previous video frames used for the purpose of classification, i.e., the cumulation lag, can be adapted to the needs of a particular application.

The proposed particle filtering-based determination of local face features enables good exploration of a face space over a video sequence and results in high recognition performance, while keeping computational requirements at a modest level. Consequently, real-time processing can be achieved on an ordinary modern PC. The trade-off between quality and computational requirements can easily be optimized for the purpose of specific applications by tuning the number of particles. Additional tuning is possible by adapting the cumulation lag to given environmental conditions or application requirements.

A particle filtering-based definition of a face feature set, in combination with cumulation mechanisms, is resistant to small rotation- and expression-caused appearance changes. It also provides good recognition performance from low resolution input videos and performs well in combination with fast dust filtering-based face normalization (in low resolution videos, precise classical normalization, such as that based on eye positions, is often not possible at all).

The proposed system opens new fields for future research. One of the most promising directions is integration of our solution with speaker recognition technology. Both approaches can operate on data retrieved from talking head video sequences, provided that voice is recorded. The integration should ensure mutual support between face-based and voice-based recognition, in particular in cases when one of two signal sources becomes unclear or temporarily unavailable. Hardware focused research should also be conducted in order to examine the advantages of the parallelizing sub-processes of the proposed system, and thus decrease the overall processing time. Additionally, research on the usage of particle filtering-determined (randomized) feature sets for the purpose of recognition in other scenarios, including recognition from still images, should provide interesting conclusions for applications with limitations on processing time. In-depth examination of distance fusion schemes – both at the frame level and sequence level – could bring further performance improvements. The possibility of using other classifiers (instead of Euclidean distance-based one) to determine face-in-frame distances should be examined. Other secondary feature representations (apart from DCT), e.g., derived from training-based methods, such as the PCA, should also be evaluated within the proposed framework.

# Acknowledgements

The work of the second Author has been financed by the Ministry of Science and Higher Education grant OR00 0026 07 "*A platform of secure biometrics implementations in personal verification and identification*".

# References

[1] A. Hampapur, L. Brown, J. Connell, A. Ekin, N. Haas, M. Lu, H. Merkl, S. Pankanti, A. Senior, C.-F. Shu, and Y. L. Tian, "Smart video surveillance – exploring the concept of multiscale spatiotemporal tracking", *IEEE Sig.. Proces. Mag.*, vol. 22, pp. 38–51, 2005.

[2] "IBM Smart Surveillance System", 2009 [Online]. Available: http://www.research.ibm.com/peoplevision

[3] S. Arulampalam, S. Maskell, N. Gordon, and T. Clapp, "A tutorial on particle filters for on-line non-linear/non-gaussian bayesian tracking", *IEEE Trans. Sig. Process.*, vol. 50, no. 2, pp. 174–188, 2002.

[4] C. Andrieu, N. de Freitas, A. Doucet, and M. I. Jordan, "An introduction to MCMC for machine learning", *Machine Learn.*, vol. 50, no. 1-2, pp. 5–43, 2003.

[5] J. Dongarra and F. Sullivan, "The top 10 algorithms", *Comput. Sci. Eng.*, vol. 2, pp. 2–97, 2000.

[6] M. Isard and A. Blake, "Condensation-conditional density propagation for visual tracking", *Int. J. Comp. Vis.*, vol. 2, pp. 5–28, 1998.

[7] P. Perez, J. Vermaak, and A. Blake, "Data fusion for visual tracking with particles", *Proc. IEEE*, vol. 92, no. 3, pp. 495–513, 2004.

[8] A. Doucet, N. de Freitas, and N. Gordon, "An introduction to sequential Monte Carlo methods" in *Sequential Monte Carlo Methods in Practice*, New York: Springer, 2001, pp. 3–14.

[9] P. Perez, C. Hue, J. Vermaak, and M. Gangnet, "Color-based probabilistic tracking", in *Proc. 7th Eur. Conf. Comp. Vis. ECCV 2002*, Copenhagen, Denmark, 2002.

[10] D. Comaniciu, V. Ramesh, and P. Meer, "Real-time tracking of non-rigid objects using mean shift", in *Proc. IEEE Int. Conf. Comp. Vis. Patt. Recogn. CVPR 2000*, Hilton Head Island, USA, 2000, pp. 142–149.

[11] K. Nummiaro, E. Koller-Meier, and L. Van Gool, "An adaptive color-based particle filter", *Image and Vis. Comput.*, vol. 21, no. 1, pp. 99–110, 2003.

[12] Ł. Stasiak and A. Pacut, "Particle filtering in multilevel color context for face detection and tracking in real-time video sequences", in *Proc. 42nd Ann. IEEE Int. Carnahan Conf. Secur. Technol. ICCST 2008*, Prague, Czech Republic, 2008.

[13] R.-L. Hsu, M. Abdel-Mottaleb, and A. K. Jain, "Face detection in color images", *IEEE Trans. Patt. Anal. Mach. Intell.*, vol. 24, no. 5, pp. 696–706, 2002.

[14] S. Spors and R. Rabenstein, "A real-time face tracker for color video", in *Proc. IEEE Int. Conf. Acoust. Speech. Sig. Proces. ICASSP'01*, Salt Lake City, USA, 2001.

[15] V. Vezhnevets, V. Sazonov, and A. Andreeva, "A survey on pixel-based skin color detection techniques" in *Proc. Graphicon-2003*, Moscow, Russia, 2003.

[16] J. Yang, W. Lu, and A. Weibel, "Skin-color modeling and adaptation". Tech. Rep. CMU-CS-97-146, School of Computer Science, Carnegie Mellon University, May 1997.

[17] J. Birgitta Martinkauppi and M. Pietikainen, "Facial skin color modeling" in *Handbook of Face Recognition*, S. Z. Li and A. K. Jain, Eds. New York: Springer, 2005, pp. 113–135.

[18] J. Yang and A. Waibel, "Tracking human faces in real-time", Techn. Rep., CMU-CS-95-210, School of Computer Science, Carnegie Mellon University, Pittsburgh, November 1995.

[19] K. Okuma, A. Taleghani, N. de Freitas, J. J. Little, and D. G. Lowe, "A boosted particle filter: multitarget detection and tracking", in *Proc. 8th Eur. Conf. ECCV 2004*, Prague, Czech Republic, 2004.

[20] Z. M. Hafed and M. D. Levine, "Face recognition using the discrete cosine transform", *Int. J. Comp. Vis.*, vol. 43, no. 3, pp. 167–188, 2001.

[21] Y. Chen and Y. Zhao, "Face recognition using dct and hierarchical RBF model", in *Proc. 7th Int. Conf. IDEAL 2006*, Burgos, Spain, 2006, *Lect. Not. Comp. Sci.*, vol. 4224, pp. 355–362.

[22] L. Wiskott, J.-M. Fellous, N. Krueger, and C. von der Malsburg, "Face Recognition by Elastic Bunch Graph Matching", in *Intelligent Biometric Techniques in Fingerprint and Face Recognition*, L. C. Jain, U. Halici, I. Hayashi, S. B. Lee, and S. Tsutsi, Eds. Boca Raton: CRC Press, 1999, pp. 355–396.

[23] T. F. Cootes, C. J. Taylor, D. H. Cooper, and J. Graham, "Active shape models – their training and application", *Comp. Vis. Image Understand.*, vol. 61, no. 1, pp. 38–59, 1995.

[24] W. Zhao, R. Chellappa, P. J. Phillips, and A. Rosenfeld, "Face recognition: a literature survey", *ACM Comput. Sur.* , vol. 35, no. 4, pp. 399–458, 2003.

[25] C. Fogg, "Questions that should be frequently asked about MPEG, April 1996 [Online]. Available: http://bmrc.berkeley.edu/research/mpeg/faq/mpeg2-v38/faq_v38.html

[26] "The Open Video Project" [Online]. Available: http://www.open-video.org/

[27] Ł. A. Stasiak and R. Vicente-Garcia, "Identity recognition-based correction mechanism for face tracking", in *Proc. 6th Conf. AMDO*, Andratx, Spain, 2010.

[28] H. K. Ekenel and R. Stiefelhagen, "Local appearance based face recognition using discrete cosine transform", in *Proc. 13th Eur. Conf. EUSIPCO 2005*, Antalya, Turkey, 2005. *Sig. Proces.*, vol. 83, no. 5, pp. 931–940, 2003.

[29] C. Sanderson and K. K. Paliwal, "Features for robust face-based identity verification",

[30] J. Stallkamp, H. K. Ekenel, and R. Stiefelhagen, "Video-based face recognition on real-world data", in *Proc. IEEE 11th Int. Conf. ICCV 2007*, Rio de Janeiro, Brazil, 2007, pp. 1–8.

[31] Y. Li, T. Zhang, and D. Tretter, "An overview of video abstraction techniques", Tech. Rep. HP-2001-191, HP Laboratory, July 2001.

[32] "Open computer vision library" [Online]. Available: http://sourceforge.net/projects/opencvlibrary/

[33] P. Viola and M. Jones, "Rapid object detection using a boosted cascade of simple features", in *Proc. IEEE Int. Conf. CVPR 2001*, Kauai, USA, 2001, pp. 511–518.

[34] S. Z. Li and A. K. Jain, Eds., *Handbook of Face Recognition*. New York: Springer, 2005.

[35] Ł. Stasiak and A. Pacut, "Particle filters for multi-face detection and tracking with automatic clustering", in *Proc. IEEE Int. Worksh. Imag. Sys. Techniq.*, Cracow, Poland, 2007.



**Łukasz A. Stasiak** was born in 1980 in Warsaw, Poland. He received the M.Sc. (with honors) in computer science from Warsaw University of Technology in 2004. From 2003 to 2006 and in 2008 he was with Biometric Laboratories, Research and Academic Computer Network NASK, Warsaw, Poland as a researcher in the field of biometric systems, with particular interest in hand- and face-based methods. He was involved in creation of first scientific biometric database in Poland. He also participated in development of systems for iris and hand signature recognition. In 2007 and 2008 he worked in a software company in Warsaw in area of Geographic Information Systems (GIS). From 2009 he is with the department for Security Technology at the Fraunhofer-Institute for Production Systems and Design Technology IPK in Berlin, Germany.

e-mail: lukasz.stasiak@ipk.fraunhofer.de
Fraunhofer IPK
Pascalstrasse 8-9
10587 Berlin, Germany

**Andrzej Pacut** – for biography, see this issue, p. 18.

*Invited paper*

# Rule Based Speech Signal Segmentation

Mindaugas Greibus and Laimutis Telksnys

**Abstract—This paper presents the automated speech signal segmentation problem. Segmentation algorithms based on energetic threshold showed good results only in noise-free environments. With higher noise level automatic threshold calculation becomes complicated task. Rule based postprocessing of segments can give more stable results. Off-line, on-line and extrema types of rules are reviewed. An extrema-type segmentation algorithm is proposed. This algorithm is enhanced by a rule base to extract higher energy level segments from noise. This algorithm can work well with energy like features. The experiments were made to compare threshold and rule-based segmentation in different noise types. Also was tested if multi-feature segmentation can improve segmentation results. The extrema rule-based segmentation showed smaller error ratio in different noise types and levels. Proposed algorithm does not require high calculation resources. Such algorithm can be processed by devices with limited computing power.**

***Keywords—rule base, speech analysis, speech endpoint detection, speech segmentation.***

## 1. Introduction

Speech segmentation is a process of labeling signal areas with symbolic information in some application. Speech segmentation is important to various automated speech processing algorithms: speech recognition, speech corpus collection, speaker verification etc. In many papers speech segmented using wavelet [1], fuzzy methods [2], artificial neural networks [3] and hidden Markov models [4]. Such segmentation algorithms give high accuracy, but also require large amount of calculation resources. In some cases this is not possible, such as mobile devices, when calculation power is weak and/or network speed is limited. In such situations, it is needed to have an algorithm to extract segments as accurate as possible and to send only them through network for external processing. Common approach to speech signal zone identification is using a threshold value.

Threshold based segmentation works in this way: feature samples which exceed chosen threshold $TH$ are marked as useful signal areas, see Fig. 1. In this case, if threshold is too low $TH_{low}$ the various noisy segments will be marked as signal, if too high $TH_{high}$ – important information at the beginning and the end may be lost. If it is known, that in the signal there is only one segment, it is possible to calculate threshold by evaluating noise samples in the beginning and the end of speech signal [5]. This algorithm for continuous speech will not work if there is no enough noisy signal at the ends of the signal. To have more accurate results, Lu proposed to use multi-feature segmentation supported by rule base to discriminate speech from music [6].

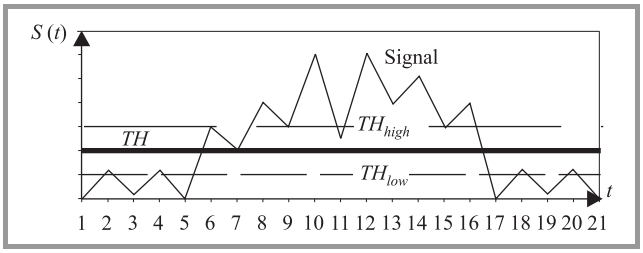



***Fig. 1.*** Threshold-based segmentation.

To improve segmentation it is possible to use background knowledge of vocal tract and peculiarities of the language. Common segmentation errors: short peaks in the signal are noise segments, two segments with short space in between can be a sibilant consonant. This knowledge can be defined as a rule base, and be used in postprocessing of initial segmentation results. Rule based postprocessing of segments can give more stable results [7]. Off-line and on-line rules are working with different types of signals. Signals retrieved from corpus is possible to postprocess with off-line rule base. Such rules will not perform good in on-line mode, when the signal being processed from microphone.

### *1.1. Off-Line Rules*

The postprocessing using a rule base can fix errors like segment interruption at the ends and short segments of noise. Waheed [7] proposed to use two rules ($l_i$ – the $i$th segment length; $d_{i\,i+1}$ – distance between $i$ and $i+1$ segment):

- if $l_i < minLength$ and $d_{i\,i+1} > minSpace$, then the segment $i$ is discarded, similarly if $l_{i+1} < minLength$ and $d_{i\,i+1} > minSpace$, then the segment $i+1$ is discarded,
- if ($l_i$ or $l_{i+1}$) $> minLength$ and $d_{i\,i+1} > minSpace$ and $(l_i + l_{i+1}) < maxLength$, then the two segments are merged, and anything between the two segments that was previously left, is made part of the speech.

The rule model proposed by Waheed works in off-line mode, when an earlier recorded signal is processed. This model for postprocessing requires to have two segments, that causes processing latency, $l_i + d_{i\,i+1} + l_{i+1}$, which can be not acceptable for automated speech recognition. On-line rules must cause smaller latency.

### *1.2. On-Line Rules*

In on-line mode, when signal processing is part of a recording stream, a segment can be recognized faster if the recognition is done in parallel, as soon as this segment starts. For that purpose it is needed to have rules for current frame [8], see Fig. 2. Such rule base has validation latency. *minLength*, until recognition module can start to process the segment in parallel and *minSpace* – till the end of segment is found.

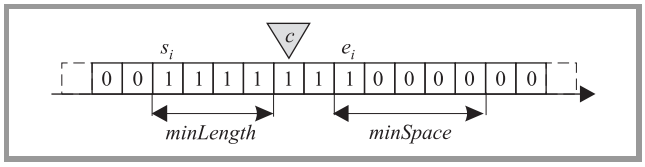



***Fig. 2.*** Postprocessing rules for each frame: 0 – noise frame, 1 – speech frame, $c$ – current position, $s_i$ – segment start, $e_i$ – segment end.

The rule base engine consists of working memory, assertions and a set of IF-THEN rules. The rule base inference is defined as a forward chaining system. The interpreter walks through the rules and applies them in order to take certain action. The rule is selected using a "first applicable" conflict resolution strategy, where rules have a specified order. Thus firing the rule with highest priority that matches current frame facts.
Rules specify how to act on the assertion set:

- **R1** $signalFrame \wedge noiseState$ : mark frame as segment start,
- **R2** $signalFrame \wedge startState \wedge validLength$ : accept start marker,
- **R3** $signalFrame \wedge endState$ : join to previous segment,
- **R4** $noiseFrame \wedge startState \wedge \neg validSpace$ : reject segment,
- **R5** $noiseFrame \wedge segmentState$ : mark frame as segment end,
- **R6** $noiseFrame \wedge endState \wedge validSpace \wedge \neg validLength$ : reject segment,
- **R7** $noiseFrame \wedge segmentState$ : accept marked segment end.

The noise states can last as long as noise frames are processed. Same thing is applied for signal state with segment frames. From the start state the machine can go to a noise state if segment start has rejected (R6) or to a segment state if it approved (R2). From the end state it can go to noise state when segment end is approved (R7) or to segment state if segments are joined (R3).
On-line and off-line rule-based approaches are dependent on result that gives a segmentation algorithm. Threshold algorithms can be used for such frame classification, but this approach has weaknesses that were mentioned before. Automated syllables-like strong segments extraction was described by [9]. Similar extrema-based segmentation can be used to find strong elements in the signal. This paper presents such extrema-type rule-based algorithm.
This article is organized in four sections: Section 2 describes proposed segmentation algorithm, in Section 3 experiment results are presented and Section 4 is a conclusion.

## 2. Extrema-Based Segmentation with On-Line Rule-Based Processing

Proposed segmentation is based on detecting local minima and maxima of signal feature. Segments are constructed using extrema and processed with the help of a rule base (see Fig. 3).

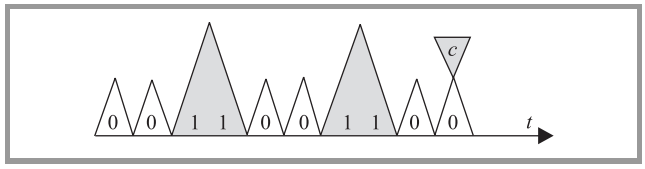



***Fig. 3.*** Extrema-based segmentation: 0 – noise segment, 1 – speech segment, $c$ – current position.

### *2.1. Extrema-Based Segmentation*

Extrema-based segmentation is working with energy-like signal features. First of all a feature value is calculated for given signal sample. Local minimum and maximum are calculated. Atomic segments are initialized. A single atomic segment contains two minima and a single maximum in between. In the next step atomic segments are processed with a rule-based module and complex segments are constructed. A complex segment has its own features, as: $A$ – area(power), $S$ – number of sub-segments, $L$ – length (see Fig. 4). Complex segments are classified into classes

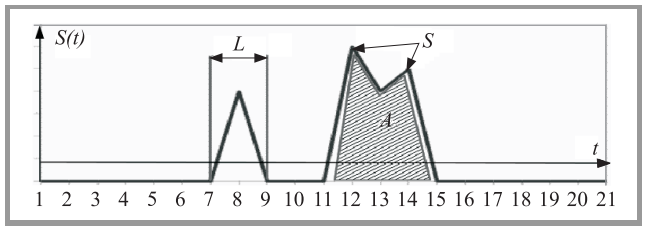



***Fig. 4.*** Segment features.

using such defined segment features. Most powerful segments represent parts of vowels. Less powerful segments represent transitions between vowels (consonants) and others segments are background noises.

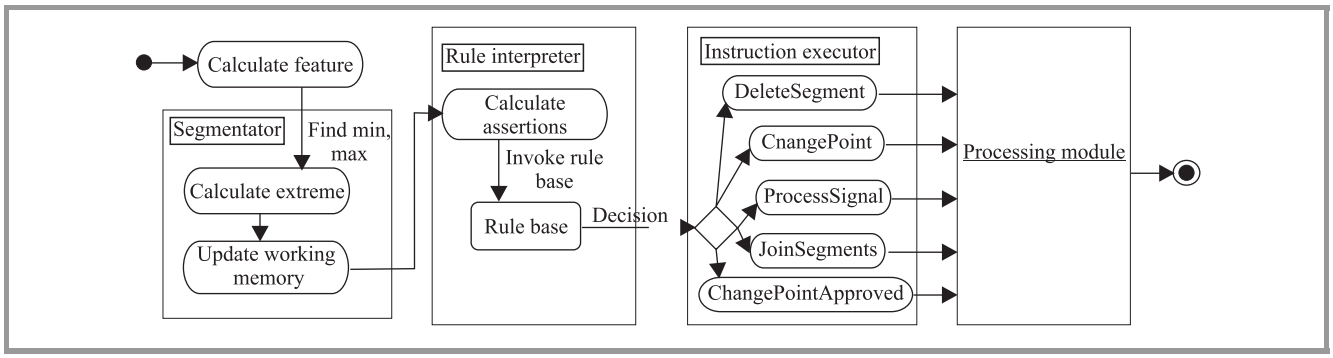



***Fig. 5.*** Extrema-based segmentation with rule-based postprocessing.

### *2.2. On-Line Rule-Based for Extrema Segmentation*

The segmentation algorithm state diagram is depicted in Fig. 5. The rule base receives calculated values from the segmentation module and it calculates assertions. Decision are made using these assertions in the rule interpreter module. Instruction executor send an event, dependent on the decision, to the processing module: automated speech recognition, automated speech corpus collector etc.
A simplified signal feature model shows how rules can be used, see Fig. 6. The processed signal has a lot of atomic segments. Extracted atomic segments has to be rejected or grouped into complex segments that point the areas where speech signal exists.

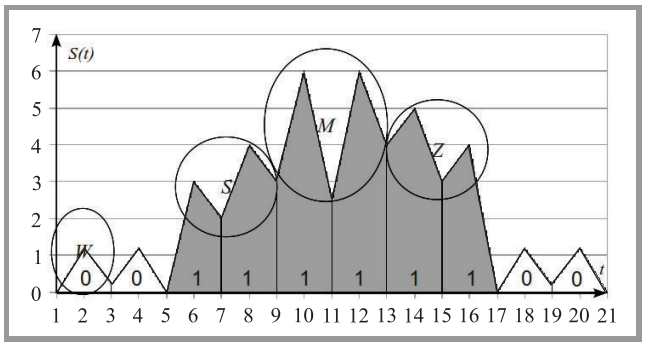



***Fig. 6.*** Feature simplified segment model.

It is possible to define rules that can be used for segment union and rejection. Examples of possible rules: segments labeled by $S(signalIncreasing)$, $Z(signalDecreasing)$ and $M(variation)$ can be joined and $W(weakSegment)$ can removed.
Rules specify how to act on the assertion set:

- R1 $isMinimum$: a change point detected,
- R2 $isMaximum \wedge signalIncreasing$: join previous and current segments,
- R3 $isMaximum \wedge signalDecreasing$: join previous and current segments,
- R4 $isMaximum \wedge previousWeakSegment$: reject previous segments,
- R5 $isMaximum$: accept previous segment.

Rules interpreter is responsible for invoking certain actions and context changes, see Fig. 5. Such rules cause segment detection latency up to time point when a next maximum is found.
Proposed algorithm should perform better than a threshold algorithm in different environment types. The extrema-based algorithm should adapt automatically to different noises. It should work well with different features such as spectral flux, signal entropy, loudness, envelope, LPC residual. Experiments show [8] that rules-based segmentation results can be improved by using not a single feature, but several features in parallel.

## 3. Experiment

For the experiments there were compared 3 types of segmentation algorithms: threshold, dynamic (adaptive) threshold and the extrema rule-based algorithm. Threshold-based segmentation calculates the global statistics (histogram) of a feature for a complete recording and then determines a fixed decision threshold [10]. This algorithm showed good performance for single word detection in a signal. Dynamic threshold-based algorithm was working on same principle, only threshold was adjusted every 10 frames. Rule based segment extractor was working on proposed algorithm. Segmentation results of all 3 algorithms were postprocessed by the rule base [8] to make the results more accurate.
In experiment the noisy speech corpus Noizeus was used. It was developed to facilitate comparison of speech enhancement algorithms among research groups [11]. The noisy database contains 30 sentences (from three male and three female speakers), corrupted by 8 different real-world noises and with different SNRs. The noise signals were taken from the AURORA database: suburban train noise, babble, car, exhibition hall, restaurant, street, airport and train-station noise.
From the Noizeus speech corpus following speech samples were taken: “The birch canoe slid on the smooth planks.”, “He knew the skill of the great young actress.”, “Her purse was full of useless trash.”, “Read verse out loud for pleasure”, “Men strive but seldom get rich.”, “The sky that morning was clear and bright blue.”, “The set of china hit

the floor with a crash.". The Noizeus corpus distributes audio signal in 8 kHz 16 bit mono audio format. The noise signals were added to the speech signals at SNRs of 0 dB, 5 dB, 10 dB, and 15 dB. 198 speech samples were taken in total.

Every segmentation algorithm was applied to signals decomposed into 30 ms frames and with 66% overlap. Hamming window was used to minimize the speech signal discontinuities at the beginning and end of each analysis frame. First order infinite impulse response filter was used for pre-emphasis [5]. Automatically extracted segments were compared with marked by the expert segments.

### *3.1. Result Evaluation*

Experiment results are evaluated with modified voice activity detector minimum performance standard TIA/EIA-136-250 [12]. In this standard there are 3 types of frames:

- Onset – few frames of speech at the beginning of speech segment.
- Steady – speech frames between onsets and offsets.
- Offset – few frames of speech at the end of speech segment.

Performance metrics are used:

- Probability of clipping speech onsets.
- Probability of detecting steady-state speech.
- Probability of clipping speech offsets.
- Normalized difference voice activity factor from truth.

These 4 error values are combined into one criterion. Ideal and testing segmentation results are processed in the same time frame by frame. When an ideal segment is started, the testing segment with the nearest boundaries are compared, see Fig. 7. Also delta voice activity factor is estimated in parallel.

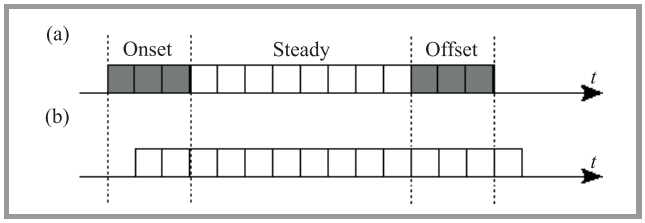



***Fig. 7.*** Segment evaluation by TIA/EIA-136-250: (a) ideal signal, (b) testing signal.

### *3.2. Experiment Results*

There were executed two types of experiments. The first one is to prove that rule based segmentation performs better than threshold segmentation. In the second it is tested if proposed algorithm shows better results in multi-feature segmentation.

### *3.3. Single Feature Segmentation*

In the first experiment series the spectral flux feature was used as main segmentation feature. Spectral flux shows rate of spectral change in a signal. It was chosen as it showed good performance in speech segmentation [8]. The experimental results are shown in Fig. 8.

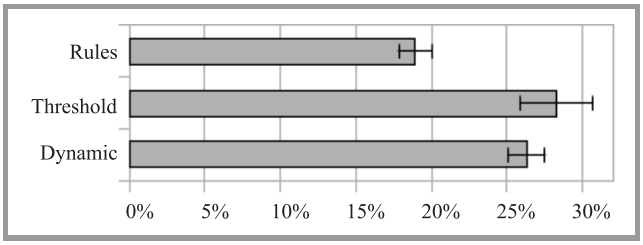



***Fig. 8.*** Segmentation errors in results of threshold, dynamic threshold and rule-based segmentation.

Proposed rule algorithm overall performed better by 7.36% in comparison with dynamic threshold and 9.33% in comparison with static threshold. As expected dynamic threshold should perform slightly better than static one.

By comparing results in different type of noise types, extrema type rule base segmentation performed with smallest error ratio in all tested noise environments except the restaurant noise (see Fig. 9). In the restaurant environment

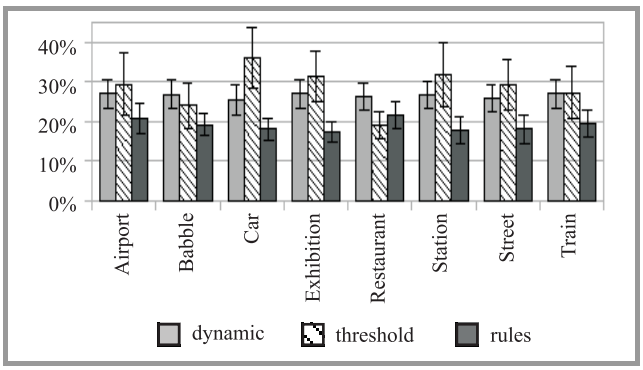



***Fig. 9.*** Segmentation error ratio comparison by noise type.

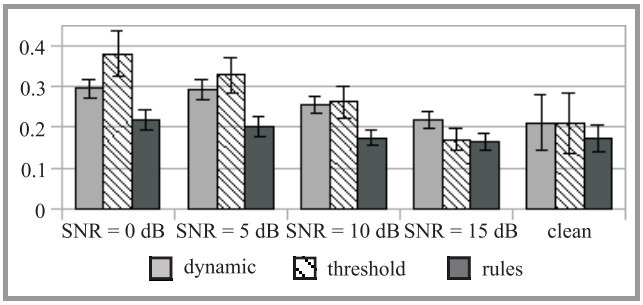



***Fig. 10.*** Error ratio in rule, threshold and dynamic threshold-based segmentation when SNR rises.

Table 1
Error ratios indexed by SNR and noise types

| Segmentation | SNR [dB] | Airport | Babble | Car | Exhibition | Restaurant | Station | Street | Train |
|---|---|---|---|---|---|---|---|---|---|
| Dynamic | 0 | 29.93% | 30.75% | 29.46% | 29.78% | 29.31% | 28.77% | 29.31% | 28.79% |
| | 5 | 28.94% | 29.57% | 30.13% | 28.65% | 28.63% | 28.46% | 28.32% | 30.69% |
| | 10 | 26.73% | 24.51% | 22.59% | 26.59% | 26.29% | 24.66% | 24.94% | 26.98% |
| | 15 | 21.95% | 22.32% | 19.53% | 23.18% | 20.39% | 24.74% | 20.18% | 21.13% |
| Threshold | 0 | 47.57% | 32.18% | 55.54% | 40.36% | 17.38% | 47.33% | 28.65% | 34.64% |
| | 5 | 25.63% | 26.10% | 44.75% | 37.83% | 22.60% | 42.29% | 30.32% | 32.41% |
| | 10 | 30.42% | 25.12% | 25.88% | 29.80% | 16.10% | 24.33% | 32.38% | 25.53% |
| | 15 | 13.97% | 12.59% | 18.19% | 17.00% | 19.55% | 13.37% | 25.50% | 16.03% |
| Rules | 0 | 23.36% | 23.64% | 17.50% | 20.10% | 26.80% | 20.19% | 19.03% | 23.02% |
| | 5 | 23.61% | 16.66% | 20.94% | 17.87% | 24.92% | 17.89% | 19.23% | 20.79% |
| | 10 | 18.56% | 19.25% | 18.59% | 14.78% | 16.94% | 14.97% | 18.50% | 18.05% |
| | 15 | 17.16% | 17.12% | 15.52% | 16.05% | 17.24% | 17.80% | 15.58% | 15.53% |

music is played in background. The rule base made mistakes as this noise has many fluctuations. Proposed rules failed to merge and reject segments that normally were expected so. Future investigation of rule base enhancement may fix this problem. In the second place street and car noises leaded to most mistakes. Dynamic threshold method had stable results in all tested environments.

Figure 10 presents error ratio of segmentation results for different noise level. Extrema type rule-based segmentation is more stable in higher noise levels, than threshold-based segmentation. However dynamic threshold-based segmentation performed similar in SNR = 15 dB, this was caused by friendly conditions for noise level estimation. Rule-based showed better performance in SNR = 5 dB and SNR = 0 dB, as fluctuations created smaller segments and allows more accurate segment detection.

In Table 1 error ratios are listed by noise levels and types for each segmentation algorithm. Threshold segmentation showed the worst result of all segmentation algorithms in car noise with SNR = 0 dB, equals to 55.54% and the best result in babble with SNR = 15 dB, equals to 12.59%. The rule-based segmentation achieved the best result in exhibition noise, SNR = 10 dB equals to 14.78% and the worst result in restaurant noise, SNR = 0 dB equals to 26.80%. Threshold segmentation is less stable in different noise types, but can show good performance in non-nosy environments.

The processing speed of each segmentation algorithm was measured to find out how much CPU time extrema segmentation process requires. In average a signal of length 2.3 s was processed in: 50 ms threshold, 55 ms dynamic and 69 ms extrema. Threshold-based and rule-based segmentation time differ by 19 ms. This shows that increased power demand is not significant. Such an algorithm can be used in devices that have limited computation power.

### *3.4. Multi-Feature Segmentation*

The aim experiment of the multi-feature segmentation experiment was to find out if parallel feature calculation can perform better that single feature.

For the second experiment 6 features were chosen: spectral flux, loudness, LPC residual, signal entropy, energy, envelope. These features showed good result in speech segmentation [8]. 41 possible combinations were created with 1 feature (6 combinations), 2 features (15 combinations) and 3 features (20 combinations). Every feature of each combination was processed in 3 steps:

- calculated feature values,
- segmented by selected algorithm,
- segmentation results were merged.

Merged final segmentation result was used to compare with expert segmentation.

Multi-feature segmentation results are presented in Fig. 11. It can be noted that the best result was showed by spectral flux and loudness features. Other feature combinations with spectral flux are in the top of the best results. spectral flux feature alone is in the 6th place. This experiment showed, that multiple feature usage may improve extrema type rule-based segmentation results. Although, it is needed deeper to investigate which feature group works better in noisy environments.

Feature combination selection for threshold-based segmentation was studied previously [8]. The order of feature combinations presented in Fig. 11 is similar to the experimental results with threshold segmentation. Although, in our experiment continuous speech signals were tested instead of short word commands, like in [8], also more types of noises were tested. It was expected that segmentation error ratio will increase for continuous speech segmentation.

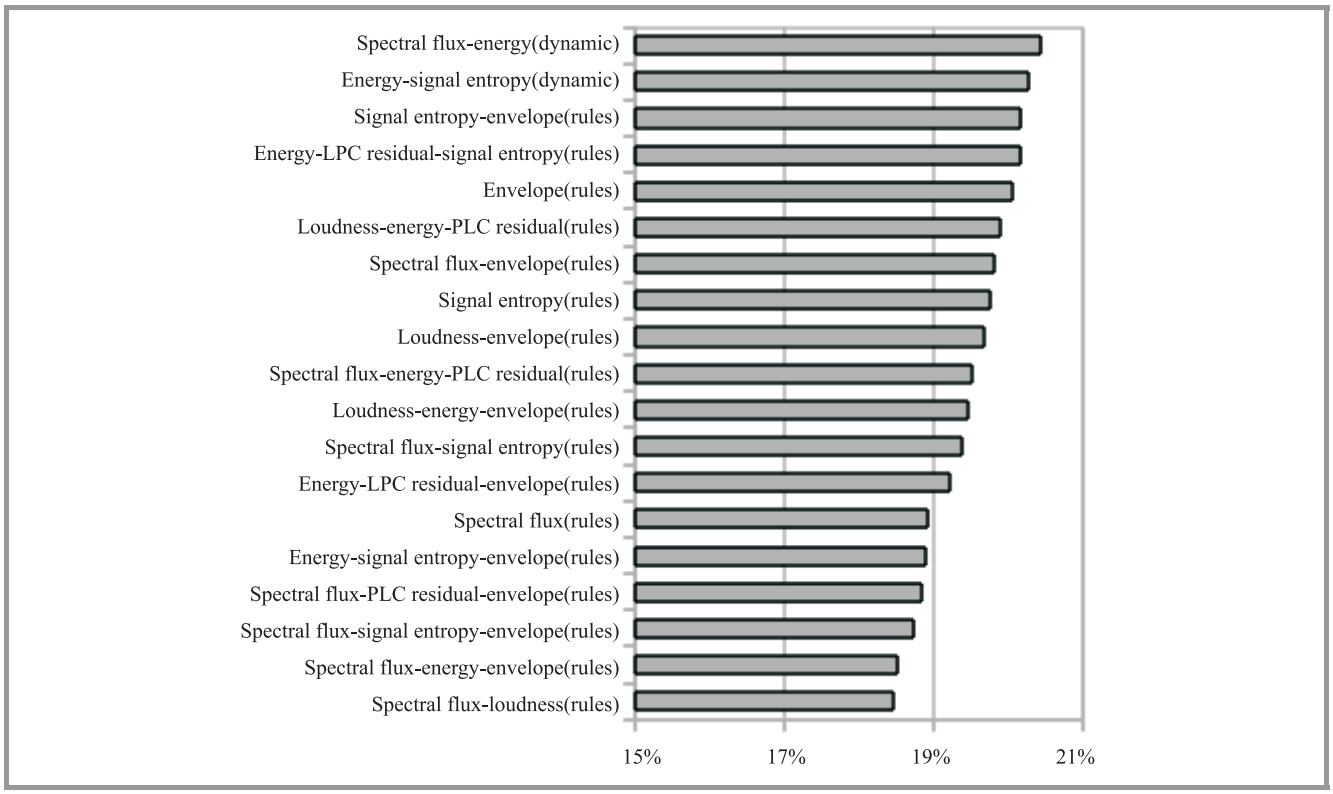



***Fig. 11.*** Multi-feature segmentation results with rule, threshold and dynamic threshold algorithms.

The segmentation with features: spectral flux and envelope was higher error ratio in this experiment by 7.97% as in [8] experiment.

## 4. Conclusions

In this paper an extrema type rule-based segmentation algorithm was proposed:

- The main novelty is the postprocessing of extrema-based segments by using several rules of segment merging and deleting.
- The experiments showed that extrema rule-based segmentation performed better than threshold or dynamic threshold-based algorithms by around 7%.
- The rule-based approach showed better result in high noise levels environments.
- Multi-feature segmentation experiment showed that the proposed algorithm it may also improve results when using 2 or 3 features in parallel.

The rule-based algorithm showed weak results in some noise type environments, this will be improved by tuning rules in future works.

## References

[1] Y. Hioka and N. Hamada, "Voice activity detection with array signal processing in the wavelet domain", *IEICE Trans. Fund. Elec. Commun. Comput. Sci.*, vol. 86, no. 11, pp. 2802–2811, 2003.

[2] F. Beritelli and S. Casale, "Robust voiced/unvoiced speech classification using fuzzy rules", in *IEEE Worksh. Speech Cod. Telecommun. Proc.*, Pocono Manor, USA, 1997, pp. 5–6.

[3] Y. Qi and B. R. Hunt, "Voiced-unvoiced-silence classifications of speech using hybridfeatures and a network classifier", *IEEE Trans. Speech Audio Proces.*, vol. 1, no. 2, pp. 250–255, 1993.

[4] S. Basu, "A linked-HMM model for robust voicing and speech detection", in *IEEE Int. Conf. ICASSP'03*, Hong Kong, China, 2003, vol. 1, pp. 816–819.

[5] J. Lipeika, A. Lipeikiene, and L. Telksnys, "Development of isolated word speech recognition system", *Informatica*, vol. 13, no. 1, pp. 37–46, 2002.

[6] L. Lu, H. Jiang and H. J. Zhang, "A robust audio classification and segmentation method", in *Proc. 9th ACM Int. Conf. Multimedia*, Ottawa, Canada, 2001, p. 211.

[7] K. Waheed, K. Weaver, and F. Salam, "A robust algorithm for detecting speech segments using an entropic contrast" in *Proc. IEEE Midwest Symp. Circ. Sys. MWCAS 2002*, Tulsa, USA, 2002, vol. 5.

[8] M. Greibus and L. Telksnys, "Speech segmentation features selection", *Inf. Technol.*, vol. 15, no. 4, pp. 33–45, 2009.

[9] P. Mermelstein, "Automatic segmentation of speech into syllabic units", *J. Acoust. Soc. Am*, vol. 58, no. 4, pp. 880–883, 1975.

[10] S. Van Gerven and F. Xie, "A comparative study of speech detection methods", in *Proc. 5th Eur. Conf. EUROSPEECH'97*, Rhodes, Greece, 1997, vol. 3, Rhodes, Greece pp. 1095–1098.

[11] Y. Hu and P. C. Loizou, "Subjective comparison and evaluation of speech enhancement algorithms", *Speech Commun.*, vol. 49, no. 7–8, pp. 588–601, 2007.

[12] *TIA/EIA-136-250 Standard*, "TDMA third generation wireless – minimum performance standards for acelp voice activity detection. Tech. Rep., TIA, 2001.



**Mindaugas Greibus** was born in Kaunas, Lithuania, in 1979. In 2005 he received the M.Sc. from the Vytautas Magnus University, Kaunas, Lithuania. He currently is pursuing a Ph.D. at the Institute of Mathematics and Informatics.

e-mail: mindaugas.greibus@exigenservices.com
Institute of Mathematics and Informatics
Recognition Processes Department
Gostauto 12
LT-01108 Vilnius, Lithuania



**Laimutis Telksnys** is Professor, Doctor Habilitatis in informatics, Doctor Honoris Causa of the Kaunas University of Technology, member of Lithuanian Academy of Sciences, head of Recognition Processes Department at the Institute of Mathematics and Informatics, Vilnius, Lithuania. He is the author of an original theory of detecting changes in random processes, investigator and developer of a computerized system for statistical analysis and recognition of random signals. His current research interests are in analysis and recognition of random processes, cardiovascular signals and speech processing and computer networking.
e-mail:
Institute of Mathematics and Informatics
Recognition Processes Department
Gostauto 12
LT-01108 Vilnius, Lithuania

*Invited paper*

# Speech Segmentation Algorithm Based on an Analysis of the Normalized Power Spectral Density

Dzmitry Pekar and Siarhei Tsikhanenka

**Abstract—This article demonstrates a new approach to speaker independent phoneme detection. The core of the algorithm is to measure the distance between normalized power spectral densities in adjacent, short-time segments and verify it based on velocity of changes of values of short-time signal energy analysis. The results of experiment analysis indicate that proposed algorithm allows revealing a phoneme structure of pronounced speech with high probability. The advantages of this algorithm are absence of any prior information on a signal or model of phonemes and speakers that allows the algorithm to be speaker independent and have a low computation complexity.**

***Keywords— phoneme segmentation, power spectral density, short-term signal energy, speaker independent, voice systems.***

## 1. Introduction

Modern research in the field of speech technologies aimed at creating robust speech systems that can be used as a basis of voice interfaces for managing the information environment.

This article describes an algorithm for converting acoustic information in speech segments into phonemes. The algorithm is based on an analysis of speech spectral characteristics. Human speech is a sequence of phonemes that have their own unique spectral parameters: power spectral density (PSD) [1], the centers of gravity of the sub-bands of the spectrum [2], the slope spectrum [3], the fundamental frequency [4], formants [5], etc. An analysis of the normalized PSD of the temporal segment of the speech signal is used in the presented algorithm.

## 2. General Structure of the Algorithm

The general structural scheme of the proposed algorithm is presented in Fig. 1.

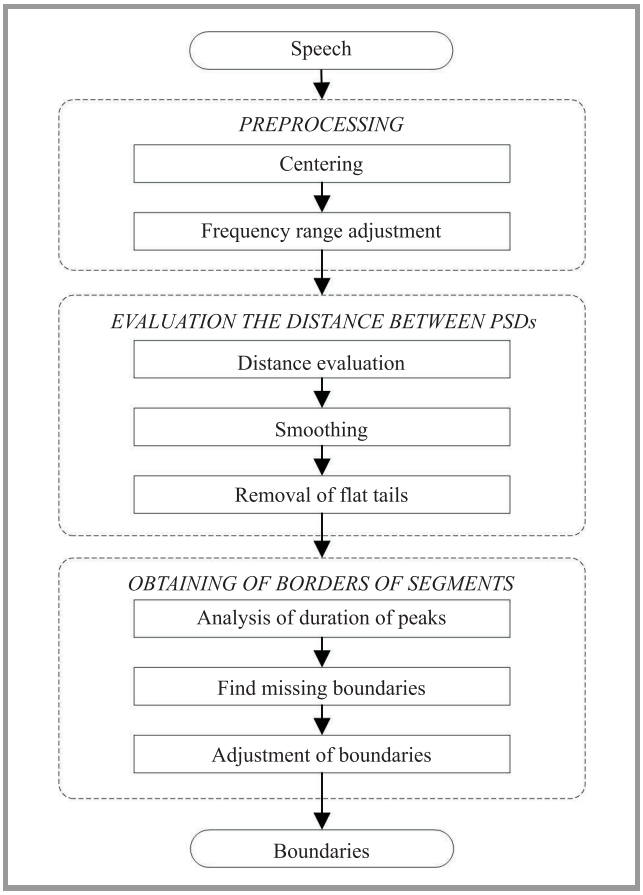



***Fig. 1.*** General structural scheme of the algorithm.

## 3. Preprocessing

Preliminary signal processing is carried out to improve the characteristics of the signal, such as reducing inserted distortions and adjustment of the frequency range. The first step is centering the signal – removing the constant component introduced by the hardware of the system according to the expression:

$$\overline{s}_i = s_i - s_{mean}\,, \tag{1}$$

where

$$s_{mean} = \frac{\sum_{i=1}^{L} s_i}{L}\,, \tag{2}$$

where: $\overline{s}_i$ — samples of the corrected signal, $s_i$ – samples of the original signal, $L$ – number of samples.

The second step is pre-emphasizing which is performed to adjust frequency range and to strengthen the information contained in higher frequency components [6], [7]:

$$H(z) = 1 - 0.95z^{-1}\,. \tag{3}$$

# 4. Distance Evaluation between PSD Adjacent Short-term Segments

After completing speech signal preprocessing, the next step is carried out for the extraction of the numeric feature, which is the normalized power spectral density. Human speech has emotion, intonation, stresses, and therefore the signal takes significant fluctuations in energy, even within the same phoneme, which may lead to incorrect segmentation. Energy normalization is used to reduce the influence of the mentioned factors: the energy of each time segment should be equal to unity, which allows to take into account only the energy distribution in frequency, rather than its absolute value. Calculating the distance between PSDs is performed according to the algorithm, which is shown in Fig. 2.

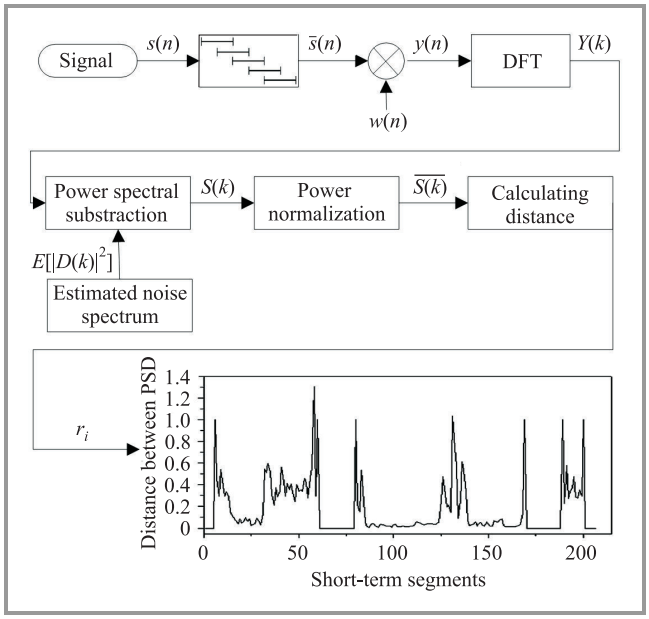



***Fig. 2.*** Structural scheme of algorithm for computing the distance between PSDs.

In the first step the segmentation of the speech signal into portions with 50% overlap is carried out, which improves the temporal localization of spectral change. To reduce the influence of boundary effects arising from the segmentation of the signal, a weighted Chebyshev window is used, which is determined by the following expression:

$$w(n) = \sum_{k=0}^{n} \frac{(-1)^k (n-k-2)! z_0^{-2k}}{(N-n-k-1)!k!(n-k)!}, \tag{4a}$$

$$z_0 = \mathrm{ch}\left(\frac{1}{\mathrm{N}-1}\mathrm{arch}\frac{1}{\mathrm{h}}\right), \tag{4b}$$

where $N$ – number of samples in short-term segment.

In Eqs. (4a) and (4b) the coefficients $a(n)$ are defined for $n$ from 0 to $L$, where $L = N/2-1$ for an even $N$. The values are determined from the condition of even symmetry of the weight function. These coefficients are normalized by equating their sum to unity.

At the next step the time segment undergoes a discrete Fourier transform to obtain the signal spectrum:

$$Y(k) = \sum_{n=0}^{N-1} y(n)\exp(-j2\pi kn/N), \tag{5}$$

where: $y(n) = w(n)s(n)$ – weighted samples of speech signal, $k = 1,2,\ldots,N-1$ – spectral components number.

Taking into account the additive nature of noise, the signal in time domain can be represented as:

$$y(n) = s(n) + d(n), \tag{6}$$

where: $s(n)$ – source signal, $d(n)$ – noise.

Then, the power spectrum is described by the following expression [8]:

$$|Y(k)|^2 = |S(k)|^2 + |D(k)|^2 + S^*(k)D(k) + D^*(k)S(k), \tag{7}$$

where: $S(k)$ – spectrum of the source signal, $Y(k)$ – spectrum of the real measured signal, $D(k)$ – noise spectrum, which is calculated during pauses in the absence of speech signals, $S^*(k)$, $D^*(k)$ – imaginary part of the spectrum of not noisy signal and noise signal, respectively.

Terms on the right side of expression (7) can be estimated as $E[|D(k)|^2]$, $E[S^*(k)D(k)]$, $E[D^*(k)S(k)]$, where the complex assessment is eliminated assuming that the noise $d(n)$ is a zero mean and is not correlated with the signal $s(n)$.

Then the estimation is calculated according to the following expression:

$$|S(k)|^2 = |Y(k)|^2 - E[|D(k)|^2]. \tag{8}$$

In the next step the signal energy is normalized according to the expression:

$$\overline{S}(k) = \frac{S(k)}{\left(\sum_z |S(z)|^p\right)^{1/p}}, \tag{9}$$

where: $S(k)$ – original $k$th spectral component, $\overline{S_k^i}$ – normalized spectral component, $p$ – parameter, which is $p = 2$.

Then the distance between $i$th and $j$th PSDs is defined by the expression:

$$r_i = \sum_k \left|\overline{S}(k)_i - \overline{S}(k)_j\right|^E, \tag{10}$$

where the exponent $E$ is a parameter that determines the sensitivity of the algorithm to the spectral changes.

The final step aims to give a threshold to the distance between PSDs and to remove the small values from the characteristics:

$$\overline{r} = \begin{cases} 0, r < E[r] \\ r, r \geq E[r] \end{cases}, \tag{11}$$

where $E[r]$ – average distance between two PSDs, calculated across the signal.

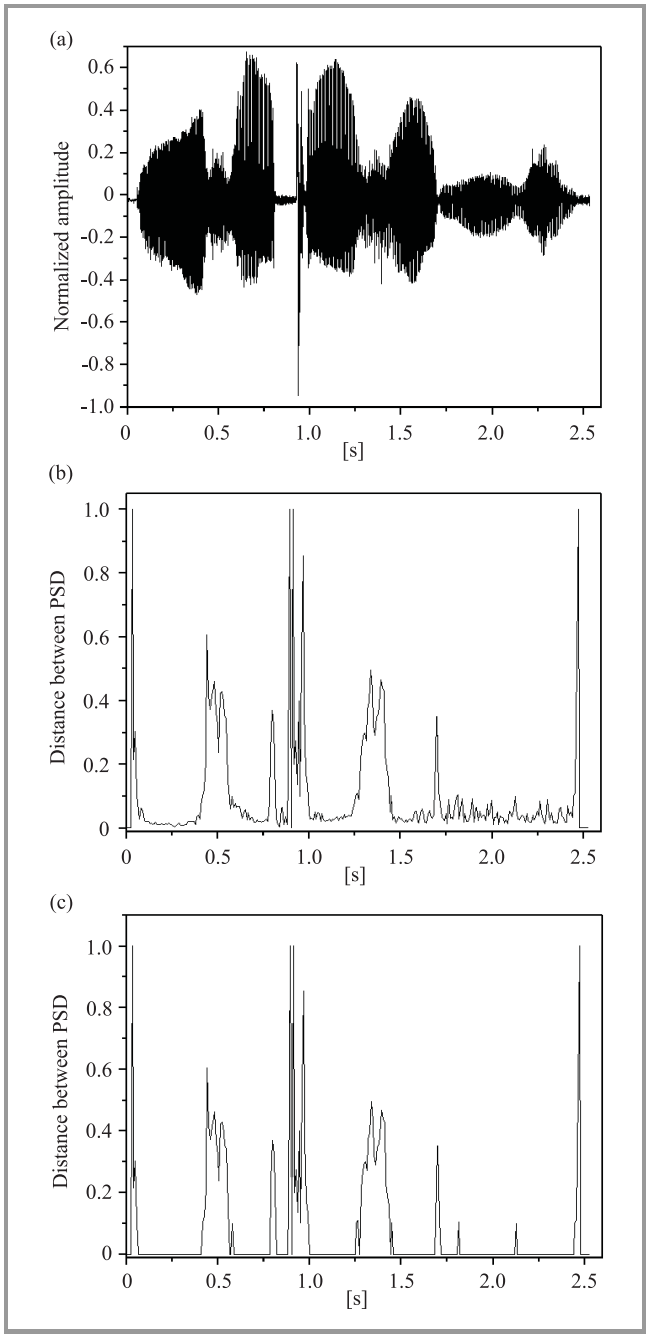



***Fig. 3.*** The stages of calculating the distance between twp PSDs: (a) original signal, (b) distance between PSD, (c) distance between PSD after threshold.

Figure 3 shows the stages of calculating the distance between PSDs. There are three types of fragments in Fig. 3c: fragments of the long zero which correspond to quasi-stationary segments, broad emissions which are associated with noise segments and narrow peaks which are related to the transitions between the quasi-stationary segments or random spectral changes. Since the vocal tract has inertia, then the sequence of transition segments appears as closely spaced peaks which correspond to the restructuring of the vocal organs. To isolate them in a typical continued fragment, smoothing is performed by the method of least squares, which is similar to passing the signal through a low pass filter, which is determined by:

$$\widetilde{r}_k = \frac{1}{2L_f + 1} \sum_{n=L_f}^{L_f} r_{k-n} b_n \,, \tag{12}$$

where: $L_f$ – length of the filter, in this particular case $L_f = 9$, $b_n$ – filter coefficients. As the filter coefficients it is advisable to use the values: $b_n = \{(-21, 14, 39, 54, 59, 54, 39, 14, -21)/231\}$. This allows achieving the required smoothing while maintaining the necessary detail.

The application of smoothing shifts the edges of the peaks. To neutralize the mentioned effect the algorithm for analysis of the growth rate peak is used in accordance to which flats edges are cut which rate of change in each point Eq. (13) is less than average rate over all peak Eq. (14) multiplied by weighting factor *thr* (see Fig. 4):

$$r'_k = \frac{-\widetilde{r}_{k+2} + 4\widetilde{r}_{k+1} - 3\widetilde{r}_k}{2} \,, \tag{13}$$

$$E[r'] = \frac{1}{W-2} \sum_{i=2}^{W-1} r'_i \,, \tag{14}$$

where W – peak length.

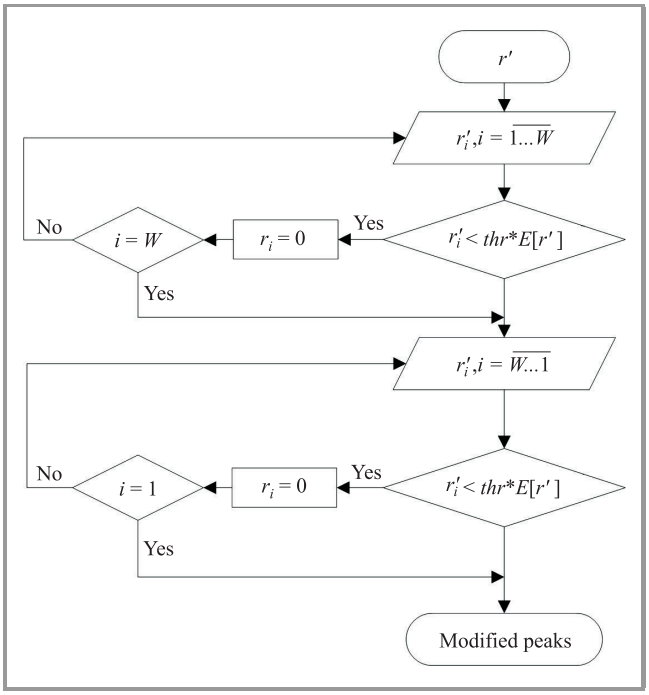



***Fig. 4.*** Structural scheme of the analysis procedure of the edges of the peaks.

Next, the analysis of the edges of the peaks is performed, according to procedure in Fig. 4.

Experimental selected value of the weighting factor *thr* was in the range of 12%–17%. Figure 5 shows result of smoothing emissions.

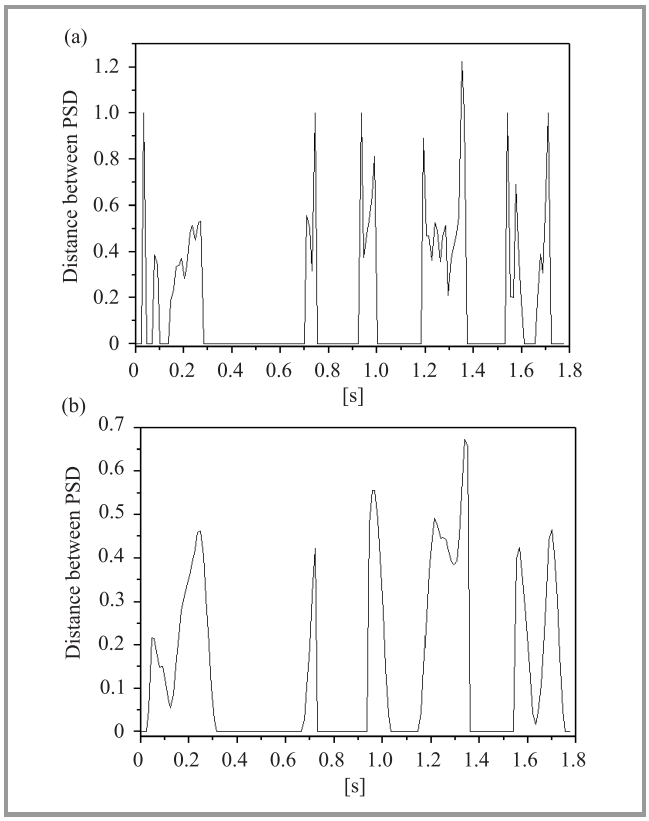



***Fig. 5.*** The results of smoothing: (a) before smoothing, (b) after smoothing.

## 5. Delimitation Segments

The purpose of this stage is to define the representatives of smoothed PSD, representing individual phonemes and the transition boundaries between phonemes. In first step, the duration of the peaks is analyzed, according to the algorithm presented in Fig. 6.

In the above algorithm: $L_{peak}$ – the duration of peak, $L_{\min}$ – the minimum allowable length of the segment. A phoneme length takes values from 50 ms to 250 ms [9], therefore, 50 ms can be taken as the minimum acceptable duration of the segment. When the duration exceeds the peak threshold, it exhibits two boundaries on the peak width, limiting the found segment, otherwise, put up one, to the maximum peak.

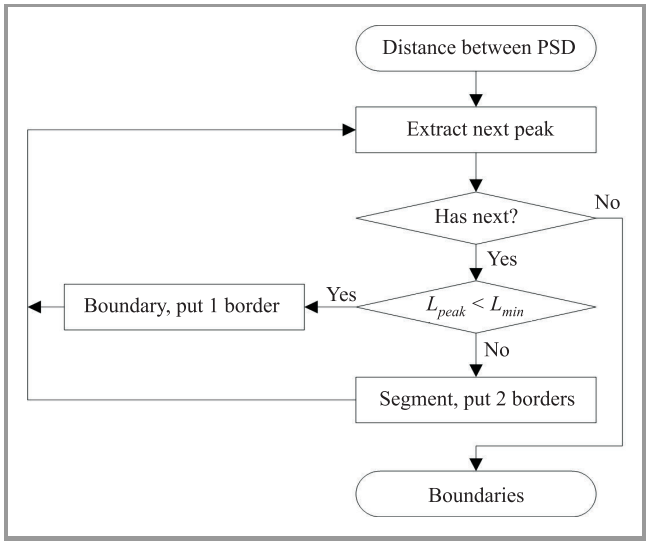



***Fig. 6.*** Structural scheme of algorithm analysis of the duration of the peaks.

After a preliminary determination of boundaries a procedure that checks for the presence of missed boundaries is performed. Based on testing of different criteria for deciding on the loss of the border, the most efficient test was the difference between the lengths of the two largest segments. If the longest segment has a duration of twice or more greater than the second length of the segment, it is necessary to check first on the presence of a merger of phonemes. Since the analysis in the frequency domain has not given the significant differences in the characteristics, then further analysis is conducted in the time domain, where the analyzed characteristic is the rate of change of short-time signal energy, which is determined according to the algorithm shown in Fig. 7.

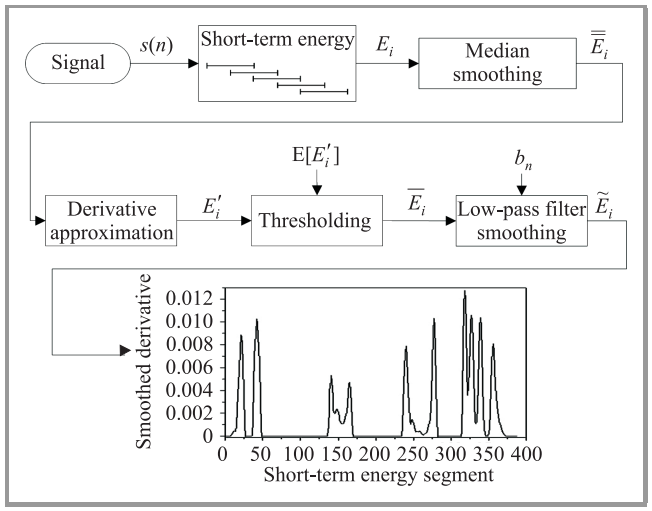



***Fig. 7.*** Structural scheme of the algorithm for evaluation values of rates of change in short-term signal energy.

Calculation of short-time energy is carried out according to the expression:

$$E_i = \sum_{j=0}^{N/2} s_{i,j}^2 , \tag{15}$$

where: $E_i$ – energy of the $i$th time segment, $s_{i,j}$ – $j$th count in the $i$th time segment.

Median filter removes random emissions, makes the curve of the values $E_i$ smoothed and allows avoiding big leaps in rate of change $E'_i$ of short-time energy $E_i$ which is calculated similar to the expression (13). Threshold processing $\overline{E}_i$ is conducted according to expression (11) with the threshold $E[E']$ defined by the expression (14). To isolate specific fragments $\widetilde{E}_i$ curve of rate of change of signal energy, all values are passed through a low pass filter Eq. (12). After these steps, the output is a similar to that which arises in the analysis of the distance between PSDs, which allows the use of an algorithm to analyze the duration of the peaks

as shown in Fig. 6. The detected segment or border of an interval occupied with the checked fragment is added to set of borders and segments that have been found before.

## 6. Boundaries Adjustment

The repeated processing of the values of distances between PSDs leads to a shifting of maximum peaks and the need to adjust the positions of boundaries of detected segments. The speech signal does not undergo significant changes for up to 20–25 ms, therefore, its spectral composition does not undergo significant changes either.

Then, the estimated boundary can occupy the position as at the beginning of the quasi-stationary section, and in the end. Therefore it is necessary to adjust boundaries in the range $[t_c - 20 \text{ ms};\ t_c + 20 \text{ ms}]$, where $t_c$ – the current position of the border segment. The position is accepted as more refined, if the following equation is fulfilled:

$$r_i^{new} > r_i^{old}, \tag{16}$$

where: $r_i^{new}$ – distance between PSDs in the new position of the boundary, $r_i^{old}$ – distance between PSDs in the old position of the boundary. Distances are calculated according to expression (12).

## 7. Experiments

The experimental validation of the algorithm was conducted to determine the probability of correct segmentation.

Table 1
The results of the segmentation algorithm

| Property [%] / $W_B$ [sample] | $P_{right}^{8000}$ | $P_{wrong}^{8000}$ | $P_{right}^{11250}$ | $P_{wrong}^{11250}$ | $P_{right}^{16000}$ |
|---|---|---|---|---|---|
| $64_{8\text{ bit}}$ | 39,1 | 54,2 | 31,7 | 59,4 | – |
| $64_{16\text{ bit}}$ | 41,5 | 48,2 | 36,3 | 57,2 | – |
| $128_{8\text{ bit}}$ | 49,9 | 21,3 | 51,2 | 15,4 | 41,1 |
| $128_{16\text{ bit}}$ | 75,7 | 11,8 | 56,4 | 17,8 | 49,6 |
| $256_{8\text{ bit}}$ | 43,4 | 25,3 | 46,8 | 23,4 | 62,2 |
| $256_{16\text{ bit}}$ | 50,1 | 18,4 | 63,3 | 19,5 | **89,4** |
| $512_{8\text{ bit}}$ | – | – | 38,1 | 26,4 | 41,8 |
| $5124_{16\text{ bit}}$ | – | – | 41,4 | 19,1 | 51,4 |
| Property [%] / $W_B$ [sample] | $P_{right}^{8000}$ | $P_{wrong}^{8000}$ | $P_{right}^{11250}$ | $P_{wrong}^{11250}$ | $P_{right}^{16000}$ |
| $64_{8\text{ bit}}$ | – | – | – | – | – |
| $64_{16\text{ bit}}$ | – | – | – | – | – |
| $128_{8\text{ bit}}$ | 31,7 | 57,1 | 31,4 | – | – |
| $128_{16\text{ bit}}$ | 27,4 | 63,1 | 28,7 | – | – |
| $256_{8\text{ bit}}$ | 15,7 | 53,1 | 25,8 | 58,8 | 24,4 |
| $256_{16\text{ bit}}$ | **9,2** | 68,4 | 21,1 | 67,1 | 19,9 |
| $512_{8\text{ bit}}$ | 30,7 | 59,7 | 16,3 | 73,8 | 11,5 |
| $5124_{16\text{ bit}}$ | 24,1 | 67,1 | 12,1 | **90,7** | **8,9** |

Denote the following notations:

- $P_{right}^{F}$ – percentage of correctly exposed boundaries at the sampling rate equal to F [Hz]:

$$P_{right}^{F} = \frac{N_{right}}{N_{total}} 100\%, \tag{17}$$

where: $N_{right}$ – number of correctly exposed borders, $N_{total}$ – total count of boundaries which is calculated at the beginning of experiments.

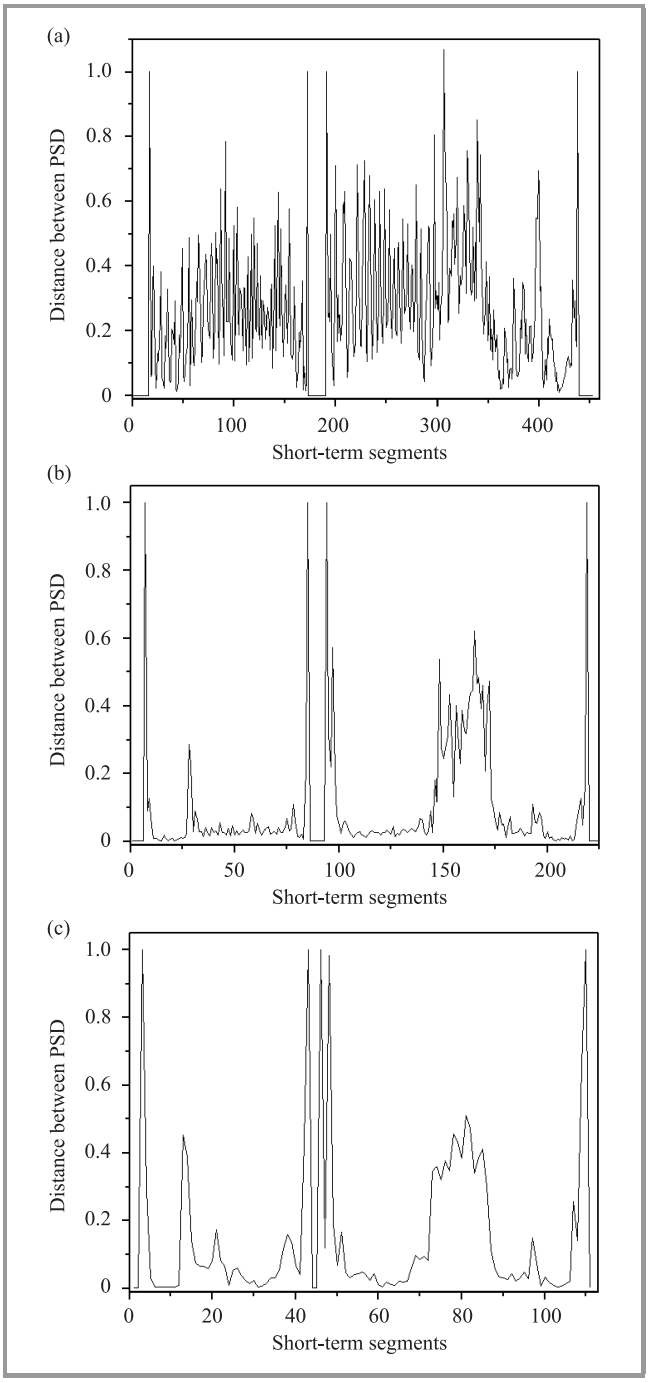



*Fig. 8.* The distance between PSDs depending on the length of the window at the sampling rate $F_s = 800$ Hz: (a) 64 samples, (b) 128 samples, (c) 256 samples.

- $P^F_{wrong}$ – the relative number of wrongly exposed borders:

$$P^F_{wrong} = \frac{N_{wrong}}{N_{total}} 100\%, \tag{18}$$

where: $N_{wrong}$ – number of wrongly exposed borders.

- $P^F_{missed}$ – percentage of missed boundaries defined by the following expression:

$$P^F_{missed} = 1 - P^F_{right} \tag{19}$$

- $W_B$ – length of analyzable short-term segment in samples with the $B$ bits per sample ratio.

The experimental digital speech signals have a sampling frequency in the range from 8000 Hz to 32000 Hz. If it goes below this range, the completeness of the presentation of the speech signal is not achieved; and above this range, there is a redundancy of representation, and excessive increase in computational complexity.
The analysis of the results of the experiment (Table 1) shows that to achieve the best segmentation results it is necessary to choose a combination of settings. This fact is explained by the excessive reduction of the window length that leads to noise pollution characteristics of the distance between PSDs (Fig. 8a), which causes the appearance of false boundaries. That's why calculations of the values in the lower left side of the table were not carried out. Otherwise, a decreased temporal localization of spectral changes appears (Fig. 8b), which reduces the accuracy of segmentation.
In the course of the experiment 50 samples of the speech signal were tested which priori contained $N_{total}$ boundaries. The samples were uttered by male and female speakers.

## 8. Conclusion

The proposed algorithm for segmentation of digital speech signals is based on the analysis of the normalized power spectral density which allows determination of the phonemic structure of pronounced speech with probability up to 90% without any priori information about signal, acoustic models of phonemes, or speaker individual characteristics which allows performing segmentation task of speech pronounced by different speakers.

## References


[1] A. Saheli and A. Abolfazl, "Speech recognition from PSD using neural network", in *Proc. Int. MultiConf. Engin. Comp. Scient. IMECS 2009*, Hong Kong, 2009, vol. 1, pp. 174–176.

[2] B. Gajic and K. Paliwal, "Robust parameters for speech recognition based on subband spectral centroid histograms", in *Proc. 7th Eur. Conf. Speech Commun. Technol. EUROSPEECH 2001*, Aalborg, Denmark, 2001.

[3] C. Espy-Wilson and S. Manocha, "A new set of features for text-independent speaker identification", in *Proc. Int. Conf. Spoken Lang. Proces. INTERSPEECH 2006*, Pittsburgh, USA, 2006.

[4] P. Labutin and S Koval, "Speaker identification based on the statistical analysis of f0", in *Proc. 16th Annual Conference IAFPA 2007*, Plymouth, UK, 2007.

[5] T. Becker and M. Jessen, "Forensic speaker verification using formant features and gaussian mixture models", in *Proc. Int. Conf. Spoken Lang. Proces. INTERSPEECH 2008*, Brisbane, Australia, 2008.

[6] E. H. Kim and K. H. Hyun, "Robust emotion recognition feature, frequency range of meaningful signal", in *Proc. IEEE Int. Worksh. Robot Human Interact. Commun.*, Nashville, USA, 2005.

[7] M. A. Al-Alaoui and L. Al-Kanj, "Speech recognition using artificial neural networks and hidden Markov models", *IEEE Multidiscipl. Engin. Educ. Mag.*, vol. 3, no. 3, 2008.

[8] N. Bhatnagar, "A modified spectral subtraction method combined with perceptual weighting for speech enhancement", M.Sc. thesis, The University of Texas, Dallas, August 2002.

[9] M. S. Medvedev, "Ispolzovanije vejvlet-preobrazowanija dla postrojenia modelej fonem russkowo jazyka", *Wiestnik Sibirskogo Federalnogo Universiteta*, no. 9, p. 198, 2006 (in Russian).





**Dzmitry Pekar** was born in Svisloch, Belarus, in 1987. He received a specialist degree in radiophysics in 2010 from Department of Intelligent Systems and Networks, Belarusian State University, Belarus. Since then he has been working on a Ph.D. thesis concerning speech segmentation and speech recognition. His research interests includes GMM based and wavelet based Russian speech processing.
email: pekar.dima@gmail.com
Belarusian State University
Nezavisimosti av. 4
Minsk, Belarus, 220030





**Siarhei Tsikhanenka** was born in Gomel, Belarus, in 1983. He received B.Sc. degree, specialist degreed and Ph.D. degrees in 2004, 2005 and 2009 from Department of Intelligent Systems and Networks, Belarusian State University, Belarus. Since 2008 he has taken a lead scientist role at multimedia data processing laboratory at Department of Intelligent Systems and Networks, Belarusian State University, Minsk. His main scientific interests are still image compression, speech processing and enterprise resource planning systems implementation.
e-mail: siarhei.tsikhanenka@gmail.com
Belarusian State University
Nezavisimosti av. 4
Minsk, Belarus, 220030

*Invited paper*

# Relaxing the WDO Assumption in Blind Extraction of Speakers from Speech Mixtures

Włodzimierz Kasprzak, Ning Ding, and Nozomu Hamada

**Abstract—The time-frequency masking approach in blind speech extraction consists of two main steps: feature clustering in a space spanned over delay-time and attenuation rate, and spectrogram masking in order to reconstruct the sources. Usually a binary mask is generated under the strong W-disjoint orthogonal (WDO) assumption (disjoint orthogonal representations in the frequency domain). In practice, this assumption is most often violated leading to weak quality of reconstructed sources. In this paper we propose the WDO to be relaxed by allowing some frequency bins to be shared by both sources. As we detect instantaneous fundamental frequencies the mask creation is supported by exploring a harmonic structure of speech. The proposed method is proved to be effective and reliable in experiments with both simulated and real acquired mixtures.**

***Keywords—blind source extraction, harmonic frequencies, histogram clustering, spectrogram analysis, speech reconstruction, time-frequency masking, W-disjoint orthogonal.***

## 1. Introduction

Blind source separation (BSS) is an approach for estimating source signals by using only mixed signals observed at many input channels [1], [2], [3]. The source reconstruction is performed blindly, without possessing information on each source, such as its location and active time, and having no knowledge about the mixing matrix. Many methods have been proposed for BSS problems, among them the most popular approaches are: independent component analysis (ICA) [2], multichannel blind deconvolution (MBD) [3] and time-frequency masking (TFM) [4]. ICA and MBD rely on statistical independence of the speech sources and that sources are mixed instantaneously or by FIR filters. However, it is difficult for ICA or MBD to solve the underdetermined case in which the source number is greater than the microphone number.

Time-frequency (T-F) masking methods are based on the assumption called *W-disjoint orthogonal* (WDO) (e.g., the DUET method [4], [5]). This assumes a sparse representation of speech in the frequency domain: although the observed signal is a mixture of several sources, most part of its time-frequency (spectrogram) cells contain one of the source signals' component only. Some other assumptions are also proposed in the method known as SAFIA [6], that performs sound source segregation based on estimating the incident angle of each frequency component. The T-F masking methods firstly make histogram (or cluster) analysis in the attenuation- and time delay-space, in order to detect the number of speakers and their characteristics, and secondly they perform source reconstruction via spectrogram masking. These methods work well for anechoic mixtures and significantly different orientations of speakers w.r.t. the microphone set.

Several novel algorithms have been developed recently, such as time-frequency ratio of mixtures (TIFROM) [7], DEMIX [8] and uniform clustering [9], that try to overcome some weak points of basic T-F masking algorithms. These improvements focus on making more efficient clustering in the 2-D attenuation rate- and delay-time space. Other recent research topic line is to provide proper microphone arrangements for T-F masking, e.g., an array of microphones or a triangle of microphones [10], [11]. Some background knowledge about speech signals can also help. In the HS method [12] it is proposed to use harmonic structure as the clustering feature.

Today the T-F masking methods work quite well for anechoic mixtures. However still there are some drawbacks of them. One of them is the error of phase-difference estimation, which is especially large in the low frequency band. It seems that in this range of frequencies, say up to 1 kHz, the assumptions of WDO is not satisfied. But one can not simply filter out these frequency components or skip it during source reconstruction, as they carry crucial information about signal's energy. The problem gets even more complicated if echoic mixtures are processed.

This paper proposes several improvements to both steps in T-F masking, the feature clustering and source reconstruction, that are relaxing the strict WDO assumption.

The paper is organized as follows. In Section 2, the BSS problem is briefly introduced and a basic T-F masking approach is defined. In Section 3 an analysis of the time-delay feature across the whole frequency spectrum is performed. The observation of large errors of time-delay estimation in the low frequency band leads to the first improvement - the use of a restrictive mask based on local and global energy distribution analysis (Section 4). The second problem is to improve the spectrogram masks for source reconstruction. A novel method for multi-valued mask generation is proposed (Section 5). Experimental results verify that crucial improvements in both histogram analysis and source reconstruction has been achieved (Section 6).

# 2. T-F Masking

In the following, we introduce the BSS problem and main processing stages of a T-F masking approach:

- extraction of spectrogram features and their clustering,
- spectrogram mask generation and source reconstruction.

In experiments we focus particularly on a situation where the number of sources $N = 2$, and the number of sensors $M = 2$.

### *2.1. The BSS Problem*

In discrete time domain, suppose that sources $s_1, \ldots, s_N$ are convolved and mixed. This is observed at $M$ sensors

$$x_j(\tau) = \sum_{k=1}^{N} \sum_{l} h_{jk}(l) s_k(\tau - l), \quad j = 1, \ldots, M, \tag{1}$$

where: $h_{jk}(l)$ represents the impulse response from source $k$ at sensor $j$, $N$ is the number of sources, and $M$ is the number of sensors.
The time domain signals $x_j(\tau)$ sampled at frequency $f_s$ are converted to frequency domain into a time-series of vector signals $X_j(t,f)$ by applying a $L$ point STFT to consecutive signal frames:

$$X_j(t,f) = \sum_{r=-L/2}^{L/2-1} x_j(r+tS) win(r) e^{-i2\pi fr}, \tag{2}$$

where: $win(r)$ is a window function, $S$ is the window shift size, $t$ is the integer time frame index, and $f$ is the integer $(0 \sim \frac{L}{2})$ frequency bin.
The time-frequency approach to blind speech separation utilizes instantaneous mixtures at each time frame $t$ and frequency bin $f$:

$$X_j(t,f) \approx \sum_{k=1}^{N} H_{jk}(f) S_k(t,f), \tag{3}$$

where: $H_{jk}(f)$ is the frequency response of the mixing system, and $S_k(t,f)$ is a frequency domain representation of the $k$-th source signal.
In time-frequency domain, signals have the property of sparseness. In mathematical form, this is described as:

$$S_1\ (t,f) \cdot S_2\ (t,f) \approx 0, \quad \forall (t,f). \tag{4}$$

### *2.2. Spectral Feature Clustering*

Currently most T-F masking algorithms utilize two features:

- the delay time calculated from the phase difference between observations,
- the attenuation rate between observations.

We limit our interest only to the delay-time. Due to our experimental setup, where all sources are located at the same distance from the microphone center, the attenuation rate provides no cues for separating among sources.

#### 2.2.1. Delay Time Calculation

The anechoic mixing process can be expressed as

$$\begin{bmatrix} X_1\ (t,f) \\ X_2\ (t,f) \end{bmatrix} = \begin{bmatrix} 1 & 1 \\ e^{-j\frac{2\pi f \delta_1}{L}} & e^{-j\frac{2\pi f \delta_2}{L}} \end{bmatrix} \begin{bmatrix} S_1\ (t,f) \\ S_2\ (t,f) \end{bmatrix}, \tag{5}$$

where: $\delta_i$ ( $i$ =1,2) is the delay between two microphones, and L is the number of STFT points.
Assuming that microphone 1 is the reference point, under the condition of WDO, the mixing model can be simplified to

$$\begin{bmatrix} X_1\ (t,f) \\ X_2\ (t,f) \end{bmatrix} = \begin{bmatrix} 1 \\ e^{-j\frac{2\pi f \delta_i}{L}} \end{bmatrix} S_i\ (t,f). \tag{6}$$

The delay $\delta_i$ is obtained using a phase correlation function [6]:

$$\delta(t,f) = \frac{L}{2\pi f} \phi(t,f), \tag{7}$$

where $\phi(t,f)$ is the phase difference,

$$\phi(t,f) = \angle X_1(t,f) - \angle X_2(t,f). \tag{8}$$

#### 2.2.2. Delay Time Histogram

Since speech signal has sparsity property against both time and frequency, to reconstruct the original signals, time-frequency cells must be clustered into two groups. The delay between observed signals can be an effective feature. Using the estimated delays and creating their histogram, we shall be able to detect two histogram peaks, $\delta_1$ and $\delta_2$, corresponding to two sources.

### *2.3. Spectrogram Masking for Source Reconstruction*

Source reconstruction is performed by binary mask detection for the spectrogram's cell, for each expected source, due to some specific feature, followed by an inverse short time Fourier transform (ISTFT). The binary mask approach depends strongly on the clustering quality of given feature. Though the delay data $\delta(t,f)$ are spread, the peaks can approximately estimate the direction of sources.
In conventional method the clustering is given by drawing the separation line at the middle of two histogram peaks. Then the binary masks are generated by

$$M_1(t,f) = \begin{cases} 1 & \text{if } |\delta(t,f) - \delta_1| < |\delta(t,f) - \delta_2|, \\ 0 & \text{otherwise} \end{cases} \tag{9}$$

$$M_2(t,f) = \begin{cases} 1 & \text{if } |\delta(t,f) - \delta_1| > |\delta(t,f) - \delta_2|. \\ 0 & \text{otherwise} \end{cases} \tag{10}$$

Therefore, the speech mixture signal can be separated by binary masks $M_i(t,f)$, and the separated signals $\hat{S}_i(t,f)$ are given by the following:

$$\hat{S}_i(t,f) = M_i(t,f) X_j(t,f). \tag{11}$$

Finally, by using the ISTFT, the separated signals are transformed in time domain.

# 3. Analysis: the WDO Assumption

### *3.1. Experimental Set-Up*

Some experiments are performed in a conference room to certify our methods. The geometrical arrangement and parameters are shown in Fig. 1, and other parameters are shown in Table 1.

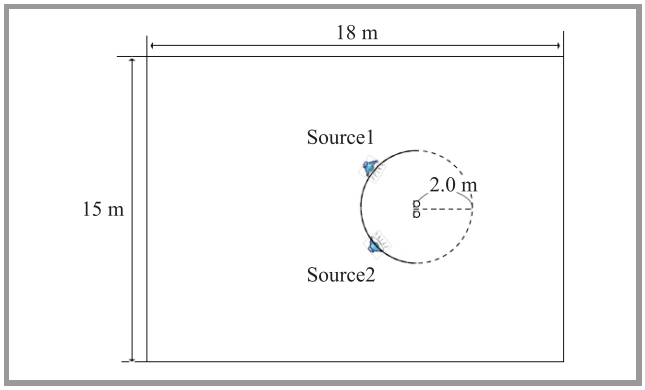



***Fig. 1.*** The arrangement for signal acquisition.

Table 1
Experimental parameter setup

| | |
|---|---|
| Sampling frequency | $f_s = 8000$ Hz |
| Microphone distance | $d = 40$ mm |
| Sound velocity | $c = 340$ m/s |
| Window type | Hamming |
| STFT frame length | $L = 1024$ |
| Frame overlap | $\Delta = 512$ |

We use sources and mixtures coming from the *ASJ continuous speech* corpus [13], available for research work. One source is located at orientation expressed with respect to the normal line of the base line of two microphones. The normal line corresponds to the direction of 0 degree. The first source can be oriented as follows: starting from 0° it can take next orientations with 10-degree increments up to 80 degrees. The other sources can be located from 10° to 90° with 10 degree-increments. Hence the smallest possible orientation difference between two sources is 10 degrees, whereas the largest one is 90 degrees.

### *3.2. Phase Difference Errors*

Although in principle the time-frequency masking based on time delay between microphones is a good method for the BSS problem, in real circumstances there appear large errors of phase difference estimation. For example, the calculated delay time derived from phase difference between two microphone signals should be less than $d/c$, where $d$ is the distance between microphones, $c$ is the sound velocity, but the estimated delays for lower frequencies obviously often violate this restriction. Fig. 2 shows examples of the time delays as a function of individual frequency bins $\{\tau = f(freq)\}$ for real and simulated mixtures.

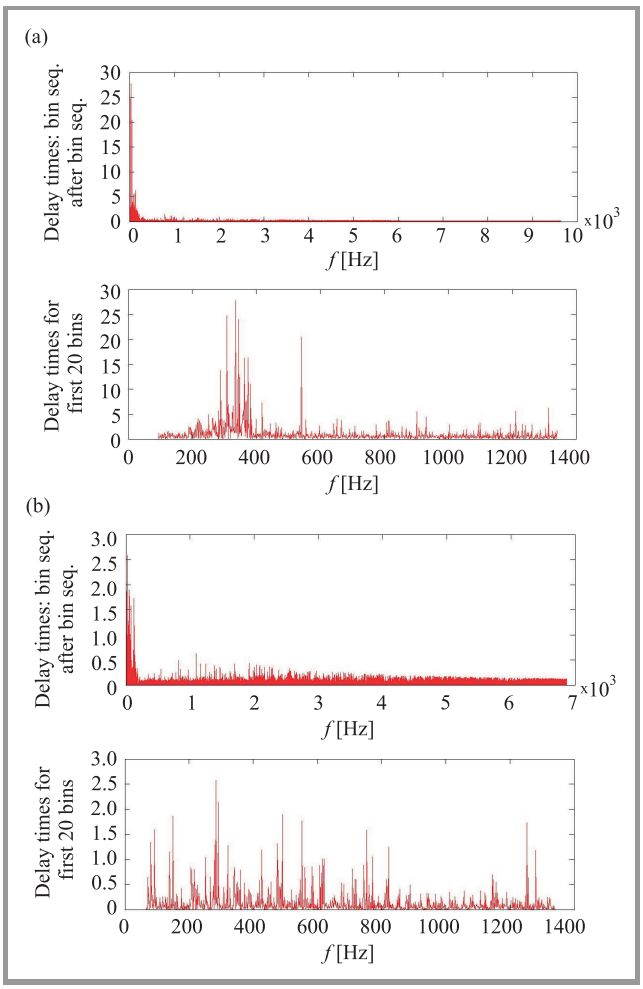



***Fig. 2.*** Time delays as a function of frequency: (a) in real mixtures and (b) simulated mixtures.

Due to large errors, both histogram analysis and cell clustering by the use of delay values will be very difficult in the lower frequency band.

### *3.3. Cut-Off Frequency*

One way to cope with the large estimation error in low frequency band is deleting these lower frequency components when generating binary mask. If we set the cut off frequency very low, the separated signal will still contain the error components. On the other hand, if we set the cut off frequency very high, it will affect the tone quality of separated signals.
The selection criterion for the cut-off frequency is keeping the tone quality of separated signals, at the same time, eliminating error components as much as possible. As demonstrated in reference [12] and by experiments, the cut off frequency need to be set around 400 Hz. In this paper instead of applying a general and simple cut-off we will individually examine each cell according to some energy criteria (Section 5).

# 4. Orientation Histogram Generation and Analysis

### *4.1. 1-D or 2-D Histogram?*

With similar magnitude of sources no significant differences in attenuation rate appear. In Fig. 3a a histogram of symmetric attenuation values (A – 1/A) is shown, computed for already restricted, selected cells. For simulated mixtures there is only one clear maximum, at 0, that corresponds to attenuation ratio, A = 1. The reason for this observation is easy explained as the sources are of normalized amplitude and both mixtures are approximately in the same amplitude range. Hence in the ideal case of simulated data there is no gain of using attenuation ratio. Our element selection rules (that will be explained later) are sufficient to get clear 1-D direction histograms.

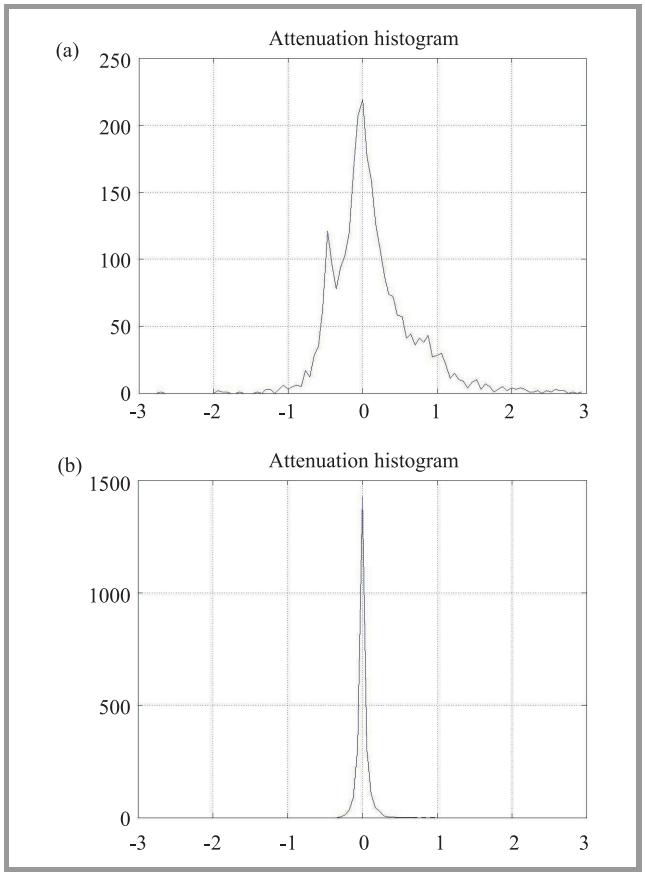



***Fig. 3.*** The 1-D histogram of symmetric attenuation ratio for (a) simulated mixtures and (b) real acquired mixtures.

But is the attenuation ratio helpful for echoic mixtures? In Fig. 3b the attenuation histogram is now deteriorated and it shows a second local maximum, around –0.5. But there is no correspondence of this maximum to any of the sources. By checking where these values come from we conclude that they are the result of a significantly delayed echo, which still is a mixture of sources.

To summarize this discussion: attenuation ratio can eventually help if the mixed signals are of significantly different amplitudes. But these can not be echoic mixtures.

### *4.2. Orientation Instead of Delay Time*

In our approach we compute a histogram of orientation angles instead of delay times. For this feature the histogram bins are linearly matching the angle scale, e.g., the difference of, say, 10 degrees corresponds to the same number of bins when $\theta$ is nearly 90 degrees or near 0 degrees. But in the histogram of delay times, the linear decomposition of histogram bins in the time space will correspond to a nonlinear scale in the orientation space, due to the mapping by the sin() function.
In fact, for the arrangement given in Fig. 1, where two sources are located at the same distance of 2 m from the center of two microphones, we can write:

$$\theta(t,f) = \arcsin(\delta(t,f)\cdot c/d), \tag{12}$$

where: $c$ is the average speed of sound and $d$ – the base distance between two microphones.

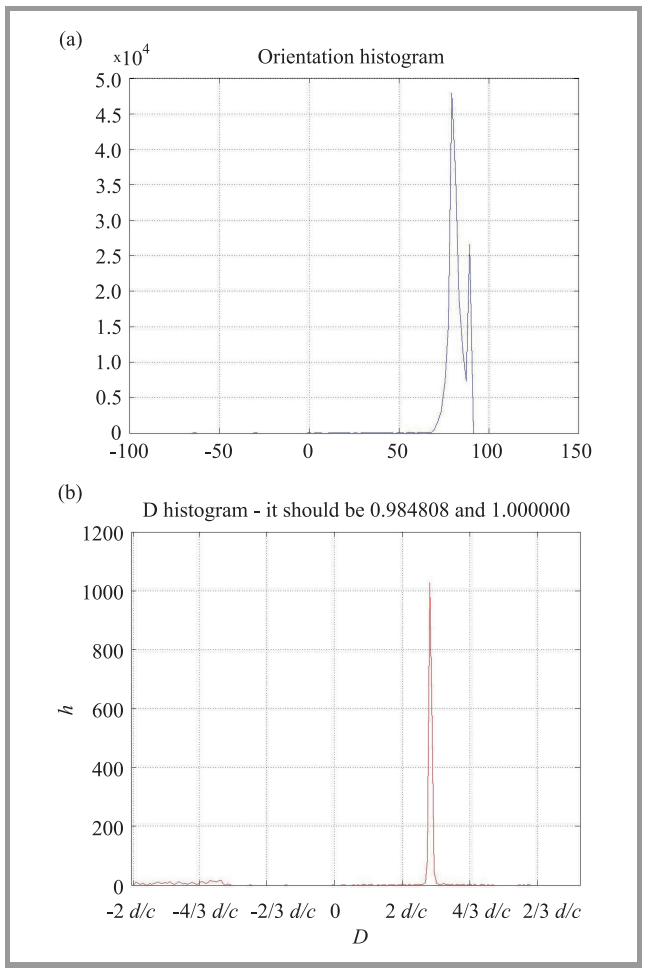



***Fig. 4.*** A difficult separation case of two sources at orientations of 80 and 90 degrees with respect to the normal to base line of microphones: (a) the orientation histogram succeeds, whereas (b) the delay-time histogram fails

The delay time $\delta(t,f)$ can be measured from the mixture spectrogram according to Eqs. (7) and (8). From Eq. (12) in turn we observe that the delay time is nonlinearly dependent on the orientation angle. We can write:

$$\delta(\theta) = \frac{d}{c}\sin(\theta). \tag{13}$$

Let us observe the Fig. 4, which illustrates the most difficult case in T-F based speech separation when both sources are oriented very closely and at 80 and 90 degrees with respect to the normal to base line of microphones, i.e., nearly in-line with this base line. Still two clear local maxima are present in the orientation histogram, but not in the time delay histogram. In the latter case the time delays are nearly the same and they fall into a single histogram bin.

### *4.3. Confidence of Time Delays versus Energy-Based Selection*

It is already well recognized that particular F-T cell's provide features with different quality or error, as in practice the WDO principle is often violated. The recently proposed methods, called TIFROM and DEMIX, use a "confidence measure" to select elements of the T-F signal (mixture) representation, which are with high probability "produced" by a single source only. The "confidence" is based on multiple PCA analysis in the attenuation-delay space for samples coming from the local neighborhood (say $3\times3$) of given element in the T-F space. The principal PCA-based axis is determined for each T-F cell and a confidence value is established that reflects the eigenvalue related to such principal eigenvector. The confidence value plays the role of a weight and allows to generate a weighted histogram. In our experiments, where both sources have similar amplitude, this approach performs worse.

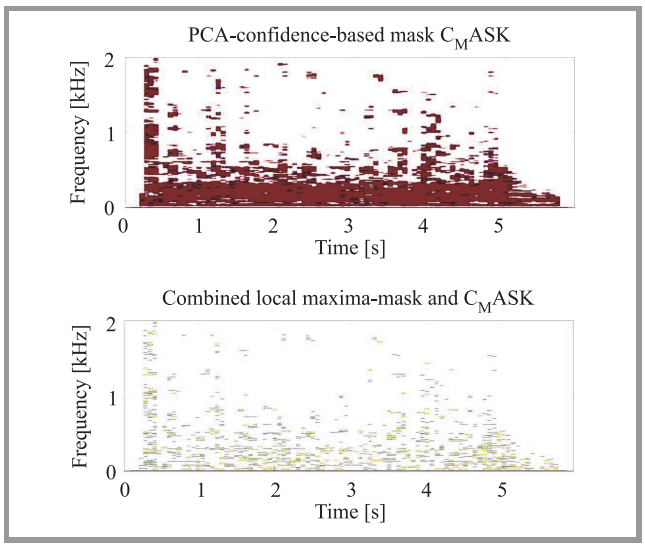



***Fig. 5.*** Selection mask for spectrogram cells based on confidence value (top), compared with energy-based selection mask (bottom) (for real mixtures).

As it is seen in Fig. 5, if we follow the DEMIX approach and allow only highly confident elements (with confidence value $>90$, we still enable most of the high energy, low-frequency elements to contribute to our direction histogram. As we have already shown, the delay information at low frequency bins is deteriorated by large errors. Applying our selection scheme we are able to concentrate on the relatively error-free information. This is further validated by results provided in Section 6.

### *4.4. Energy-Based Selection Criteria*

Instead of computationally expensive approaches in TIFROM or DEMIX, our proposition is to use a restrictive cell selection rule, considering two criteria.

1. The local maxima along each frequency-indexed column (Fig. 6).
2. Near global maximum cells along the time axis for each frequency bin (Fig. 7).

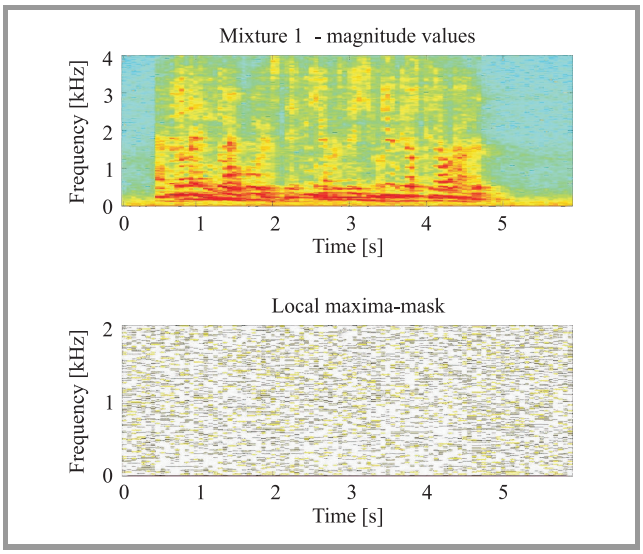



***Fig. 6.*** The spectrogram of first mixture and a local maximum-based cell selection mask.

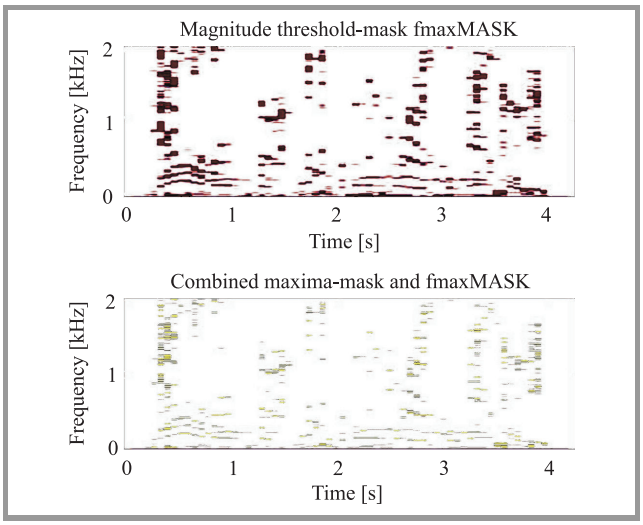



***Fig. 7.*** Global-maximum-based cell selection mask (top) and the combined local- and global-based selection mask.

# 5. Many-Valued Masking

### *5.1. Classification Rule*

The simple criterion for T-F mask generation, presented in Section 2, which classifies a spectrogram cell according to the smallest distance of its feature value to a histogram peak of this feature, is not a proper solution. We propose a more restrictive criterion that allows a feature peak-based classification only for cells where the distances-to-peak are relatively small. Then the appropriate mask is filled with value 1 and the other with 0 at given cell. If the feature value differs to much from all the detected histogram peaks, the masks are filled with some values from the interval [0, 1], computed by some frequency distance functions: $A_i(t,f), i = 1,2$. The rule for creation of the two spectrogram masks is as follows:

$$M_1(t,f) = \begin{cases} 1 & \text{if } |\theta(t,f) - \theta_1| < \theta_{max} \\ 0 & \text{if } M_2(t,f) = 1 \\ A_1(t,f) & \text{otherwise} \end{cases}, \quad (14)$$

$$M_2(t,f) = \begin{cases} 1 & \text{if } |\theta(t,f) - \theta_2| < \theta_{max} \\ 0 & \text{if } M_1(t,f) = 1 \\ A_2(t,f) & \text{otherwise} \end{cases}. \quad (15)$$

The normalized frequency distance functions are:

$$A_1(t,f) = \frac{W_1(t,f,f0_1(t))}{W_1(.) + W_2(.)}, \quad (16)$$

$$A_2(t,f) = \frac{W_2(t,f,f0_2(t))}{W_1(.) + W_2(.)}, \quad (17)$$

where the $f0_i(t) - s$ represents the fundamental frequency of source $i$ in window $t$. The distance function $W_i(t,f,f0_i)$ gives a weight in proportion to two distances of cell's frequency $f$ to the two nearest harmonic frequencies of given source $(n_L f0, n_H f0)$.

### *5.2. Harmonic Frequencies*

The next results illustrate processing steps for the detection of two fundamental frequencies and their common multiple frequency. Even in a general overview of the total energy distribution along frequency bins we can already distinguish local maxima that corresponds to fundamental frequencies of both speakers and to magnifications around common multiple frequency (Fig. 8).

As the fundamental frequency can change during the speech the energy measurements are repeated every several consecutive frames. At first the gradient function is computed from the energy function along frequency axis (Fig. 9). Then the clearly visible local maxima peaks are detected and their harmonic structure is analyzed in order to select the fundamental frequencies and their common multiple frequency (Fig. 10).

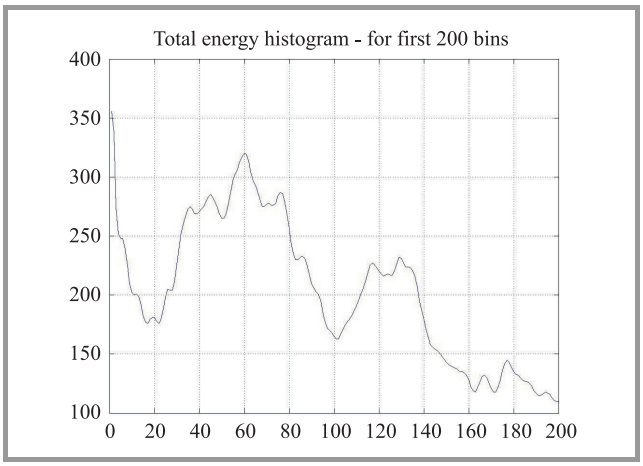



***Fig. 8.*** Total energy distribution per frequency bins for real mixtures.

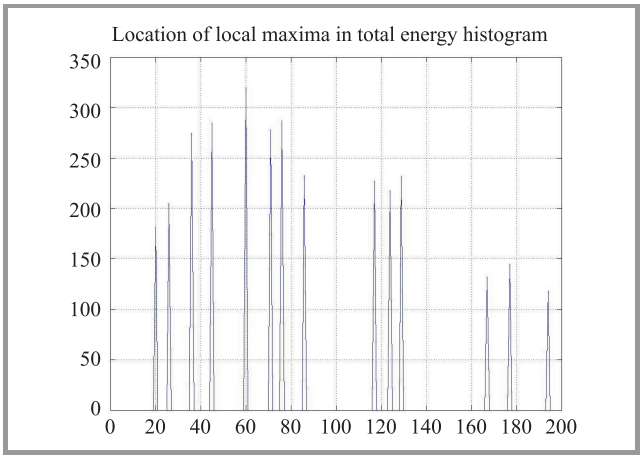



***Fig. 9.*** The energy gradient along the frequency axis computed for real mixtures.

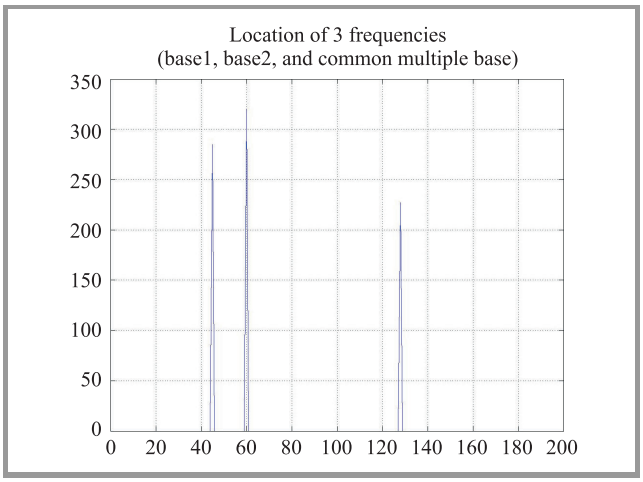



***Fig. 10.*** The locations of two fundamental frequencies and a common multiple frequency for real mixtures.

# 6. Results

### *6.1. Histogram Analysis – Experiments*

In experiments it turned out that the most difficult case is to distinguish between orientations of 80–90 degrees. This was the reason while we prefer to use direction feature clustering instead of the time delay one. The direction histogram for real data is not a simple mixture of two Gaussians, centered at speaker directions, as the interference is so that the second signal seems to "generate symmetric peaks" around the center of the first signal (Fig. 11). This interference effect could also be responsible for lower histogram peaks of the second source.

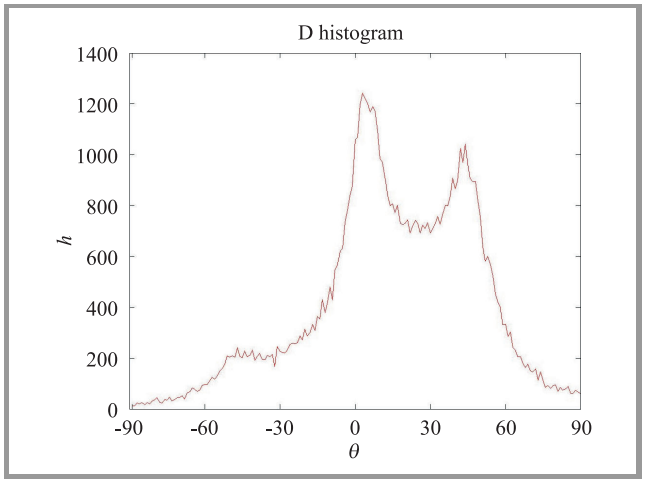



***Fig. 11.*** Example of orientation histograms obtained for mixtures of two sources.

Table 2
The estimated orientations $\theta_1$ and $\theta_2$ (in top and bottom rows) based on the orientation histogram for two real acquired mixtures

| $s_2$ at: | 20° | 30° | 40° | 50° | 60° | 70° | 80° | 90° |
|---|---|---|---|---|---|---|---|---|
| $s_1$ at | 13 | 15 | 13 | 13 | 13 | 12 | 12 | 12 |
| 10° | 22 | 29 | 41 | 46 | 55 | 73 | 78 | 87 |

Table 2 shows how well the clustering can be done for the whole representative range of orientation angles $[0°, 90°]$. The detection of orientations for both speakers has improved, especially in the most difficult range of orientations: $[80°, 90°]$.

### *6.2. Source Reconstruction*

The performance of the spectrogram masking step will be evaluated in terms of the WDO coefficient (measure of W-disjoint orthogonality) [4]. This coefficient is computed for given useful destination source and interference signal. Related criteria are: the preserved-signal ratio (PSR) and the signal-to-interference ratio (SIR). The definitions are as follows:

$$WDO(d,i) = \frac{||M_d(t,f)S_d(t,f)||^2 - ||M_d(t,f)S_i(t,f)||^2}{||S_d(t,f)||^2} = PSR - \frac{PSR}{SIR}, \quad (18)$$

$$PSR = \frac{||M_d(t,f)S_d(t,f)||^2}{||S_d(t,f)||^2}, \quad (19)$$

$$SIR = \frac{||M_d(t,f)S_d(t,f)||^2}{||M_d(t,f)S_i(t,f)||^2}, \quad (20)$$

where: $S_d(t,f)$ is the desired source signal, $M_d(t,f)$ is the spectrogram mask for source $d$, and $S_i(t,f)$ is the interfering signal. The range of WDO values is: $0 \leq WDO \leq 1$. For ideal source reconstruction it would be $WDO = 1$.

The results in Table 3 clearly illustrate the statement that a binary spectrogram mask does not allow a proper extraction of speech sources from real echoic mixtures. The WDO coefficients have low values within the range of $[0.26, 0.66]$.

Table 3
The WDO(1,2) and WDO(2,1) coefficients (in top and bottom rows) for source reconstruction with ordinary binary spectrogram masks (according to Eqs. (9)–(10))

| $s_2$ at: | 20° | 30° | 40° | 50° | 60° | 70° | 80° | 90° |
|---|---|---|---|---|---|---|---|---|
| $s_1$ at | 0.39 | 0.46 | 0.66 | 0.59 | 0.50 | 0.49 | 0.29 | 0.26 |
| 10° | 0.27 | 0.33 | 0.52 | 0.40 | 0.30 | 0.27 | 0.28 | 0.26 |

The results in Table 4 have been achieved by applying the multi-valued mask for source extraction, proposed in this paper. Here we focus on most difficult situations, when the sources are close to each other and their orientations w.r.t. the microphones are ending towards 90°. The results are significantly better than in the binary mask case. With the multi-valued mask a sufficiently good source extraction is

Table 4
The WDO(1,2) and WDO(2,1) coefficients (in top and bottom rows) for source reconstruction with multi-valued spectrogram masks (according to Eqs. (14)–(17))

| $s_2$ at: / $s_1$ at | 60° | 70° | 80° | 90° |
|---|---|---|---|---|
| 50° | 0.92 | 0.91 | 0.90 | 0.88 |
| | 0.90 | 0.89 | 0.87 | 0.85 |
| 60° | – | 0.90 | 0.90 | 0.88 |
| | – | 0.89 | 0.88 | 0.85 |
| 70° | – | – | 0.81 | 0.68 |
| | – | – | 0.75 | 0.58 |
| 80° | – | – | – | 0.45 |
| | – | – | – | 0.27 |

possible even for orientations in the range of 80° (and to some part even to 90°), where the binary mask definitely failed.

## 7. Conclusion

This paper introduces several improvements to the time-frequency masking approach to blind speech separation, that relax the strict DOA assumption. After experiments with anechoic (simulated) mixtures and echoic (real) mixtures of speech sources, acquired by two microphones, we worked out methods that improve two steps of such conventional approach – orientation histogram analysis and T-F mask creation. The creation of an orientation histogram is efficiently performed by considering the phase-difference data of reliable cells only. For this we combine an energy local maximum criterion along the frequency axis (for every time frame) with near global maxima intervals along the time axis (for each particular frequency bin).
Next improvement is due to the use of many-valued spectrogram mask. Thus we relax the strict WDO assumption, that seems not to hold perfectly in practice. The clustering feature is now only responsible for selecting cells with obviously perfect behavior. Otherwise the harmonic frequencies are applied as a new selection criterion. The WDO coefficients for the reconstructed sources document a significant improvement, even for close sources and for large orientation angles – reaching nearly 90°.

## Acknowledgments

Włodzimierz Kasprzak was supported by the Polish Ministry of Science and Higher Education within the grant N N514 1287 33.

## References

[1] S. Makino, T.-W. Lee, and H. Sawada, *Blind Speech Separation*. Berlin: Springer, 1997.

[2] A. Hyvärinen, J. Karhunen, and E. Oja, *Independent Component Analysis*. New York: Wiley, 2001.

[3] A. Cichocki and S. Amari, *Adaptive Blind Signal and Image Processing.* Chichester: Wiley, 2003.

[4] O. Yilmaz and S. Rickard, "Blind separation of speech mixtures via time-frequency masking", *IEEE Trans. Sig. Proces.*, vol. 52, no. 7, pp. 1830–1847, 2004.

[5] S. Rickard, "The DUET blind source separation algorithm", in [1], pp. 217–237.

[6] M. Aoki, M. Okamoto, S. Aoki, H. Matsui, T. Sakurai, and Y. Kaneda, Sound source segregation based on estimating incident angle of each frequency component of input signals acquired by multiple microphones. *Acoust. Sci. & Tech*, vol. 22, pp. 149–157, no. 2, 2001.

[7] F. Abrard and Y. Deville, "A time-frequency blind signal separation method applicable to underdetermined mixtures of dependent sources", *Sig. Proces.*, vol. 85, pp. 1389-1403, 2005.

[8] S. Arberet, R. Gribonval, and F. Bimbot, "A robust method to count, locate and separate audio sources in a multichannel underdetermined mixture", *IEEE Trans. Sig. Proces.*, vol. 58, no. 1, pp. 121–133, 2010.

[9] Z. He, A. Cichocki, Y. Li, S. Xie, and S. Sanei, "K-hyperline clustering learning for sparse component analysis", *Sig. Proces.*, vol. 89, pp. 1011–1022, 2009.

[10] S. Makino, H. Sawada, R.Mukai, and S. Araki, "Blind source separation of convolutive mixture of speech in frequency domain", *IEICE Trans. Fundament.*, vol. 88, no. 7, pp. 1830–1847, 2004.

[11] A. Araki, H. Sawada, R. Mukai, and S. Makino, "Underdetermined blind sparse source separation for arbitrarily arranged multiple sensors", *Sig. Proces.*, vol. 87, pp. 1833–1847, 2007.

[12] H. Ouchi and N. Hamada, "Separation of speech mixture by time-frequency masking utilizing sound harmonics, *J. Sig. Proces.*, vol. 13, no. 4, pp. 331–334, 2009.

[13] T. Kobayasi, S. Itahashi, S. Hayamizu, and T. Takezawa, "ASJ continuous speech corpus for research, *J. Acoust. Soc. Japan*, vol. 48, no. 12, pp. 888–893, 1992 (in Japanese).



**Włodzimierz Kasprzak** received the German Dr.-Ing. in pattern recognition (1996) from University of Erlangen-Nuremberg, and the Polish Ph.D. in computer science (1987) and D.Sc. in automation control (2001) from Warsaw University of Technology (WUT). From 1997, he is with the Institute of Control and Computation Engineering at WUT, currently appointed as Associate Professor. His main interests are in pattern recognition and artificial intelligence, and their applications in image and speech analysis. Among others, he has conducted research at the University of Erlangen-Nuremberg (1989–1991), the Bavarian Research Center FORWISS, Erlangen, Germany (1992–1995), and the RIKEN Institute, Wako-shi, Japan (1995–1996). He is Member of the IAPR and Polish TPO societies.
e-mail: W.Kasprzak@ia.pw.edu.pl
Institute of Control and Computation Engineering
Warsaw University of Technology
Nowowiejska st 15/19
00-665 Warszawa, Poland



**Ning Ding** received the B.Sc. and the M.Sc. degrees in electronic engineering from Xi'an Jiaotong University, Xi'an, China, in 2005 and 2008, respectively. He is currently working toward the Ph.D. degree in signal processing at Keio University. His main research interest is the speech signal processing, Direction of Arrival estimation and speech separation.
e-mail: ding@hamada.sd.keio.ac.jp
Signal Processing Lab.
School of Integrated Design Engineering
Keio University
3-14-1 Hiyoshi, Kohokuku
Yokohama 223-8522, Japan



**Nozomu Hamada** received the B.Sc., the M.Sc. and Ph.D. degrees in electrical engineering from Keio University. In 1974, he became an Instructor in electrical engineering at Keio University. He has been a Professor there in the Department of System Deign Engineering since 1991. He was the visiting researcher in Australian National University in 1982. His research interests include circuit theory, stability theory of dynamical system, design of one or multi-dimensional digital filters, and image processing. One of his resent research fields is a realization of human interface system using microphone array. The main topics in this study are acquisition of audio signal from spatially distributed sound sources and the separation of multiple speech signals. He is the author of "Linear Circuits, Systems and Signal Processing" (Chapter 5) (Marcell Dekker Inc. 1990), "Two-Dimensional Signal and Image Processing" (SICE, 1996), "Introduction of Modern Control Systems" (Corona Pub. Inc. 1997). He was the chair of IEEE Signal Processing Society in Japan Chapter (2004) and editorial board in Journal of Signal Processing. He was guest editor of special issues relevant to multi-dimensional signal processing: its application and realization technique in the IEICE (2000), Signal Processing (2002, 2003, 2006, 2007).

e-mail: hamada@sd.keio.ac.jp
Signal Processing Lab.
School of Integrated Design Engineering
Keio University
3-14-1 Hiyoshi, Kohokuku
Yokohama 223-8522, Japan

*Paper*

# Hidden Signature for DTW Signature Verification in Authorizing Payment Transactions

Joanna Putz-Leszczyńska and Michał Kudelski

**Abstract—Traditional use of dynamic time warping for signature verification consists of forming some dissimilarity measure between the signature in question and a set of "template signatures". In this paper, we propose to replace this set with the hidden signature and use it to calculate the normalized errors of signature under verification. The approach was tested on the MCYT database, using both genuine signatures and skilled forgeries. Moreover, we present the real-world application of the proposed algorithm, namely the complete biometric system for authorizing payment transactions. The authorization is performed directly at a point of sale by the automatic signature verification system based on the hidden signature.**

***Keywords—dynamic time warping, hidden signature, payment transactions, signature verification.***

## 1. Introduction

Despite an abundance of computers systems, handwritten signature is still widely used in everyday life. It is still essential for most of the financial transactions, notarial deeds, etc. However, there exists a considerable risk that someone can forge a signature. Forensic handwriting experts, whose discipline is forensic document examination, can tell a forgery from a genuine with almost 100% accuracy, but it is hard to imagine to seat a graphologist next to every bank worker. What is more, a forensic document examiners usually needs more than few seconds to perform his procedures. In spite of problems related to handwritten signature, it is still irreplaceable in many places. That is why there exists a great demand for automatic signature verification systems.

The first article describing automatic signature verification systems was published in 1979 by Herbst [1], that is more than 30 years ago. Even if the idea of automatic signature verification systems is a very good solution for banks and other places, it is still not commonly used in everyday life. This has a number of causes. First, in the very beginning, this idea was more like science fiction, and paying with plastic cards instead of paper money was not very common outside the US. The high cost and low quality of digitalizing tablets also curbed the on-line verification development. But it changed. Now, there exist tablets designed for signature verification. They are easy to get and their prices are not too steep.

The digitalizing tablets are important, because they alow for on-line signature capture. It means that the signature is represented as a sequence in time. The characteristic feature of signatures is that the instances of one signature can differ strongly between each other. This can be caused both by natural fluctuations and physical or emotional states. Two given instances can differ in amplitudes or values at certain points, but also in the signature dynamics. In other words, the time scales of two signatures can be different, which makes direct (point by point) comparisons impossible. For that issue, a nonlinear sequence alignment method is needed.

One approach that can be employed is dynamic time warping (DTW). The basic idea behind DTW is to "warp" the time dimension of two varying time sequences in a nonlinear manner. This allows for measuring dissimilarity between them.

We successfully used this approach in our original signature verification method. That is why in this paper we concentrate on methods based on DTW. These methods may differ in many important aspects. For example, DTW can be used only to compute the final score or to construct a classifier. However, what all such systems have in common is the template creation stage, in which either several (e.g., three, [2], [3]) or a single signature ([4], [5], [6]) is selected to represent the training set. While this selection has a big impact on the system quality, it is often arbitrary, and typically consists of choosing the signature(s) being closest (in some metric) to the rest of the training signatures. Some authors selected the template signature at random, or without elaboration on the statistic used for the selection [6].

The diversity of solutions shows two important problems associated with it. Both of these problems lie in template creation, which is seemingly very easy, and generally is limited to selecting a subset of a persons signatures. However – and this is the first problem – we do not know how many signatures should be selected for the template. Second, even if we resolve the first problem, we still do not know, which of the signatures should be selected.

The approach presented in this paper makes an effort to meet the mentioned needs. It is based on the well known property of least square solutions, namely, that the average minimize the squared norm distance to all averaged elements. While a direct averaging of signatures is not possible (even signature instances of the same person may have different length and may differ not only in magnitudes of the measured quantities but also in their "local speed"), we may transform all training signatures to a common space

by warping and then find their average in this space. This approach carries on various properties of least squares, while bringing in a possibility of local variation by proper warping. The averaged warped signature is called the hidden signature. It is an artificial signature which has a feature of minimizing the mean dissimilarity between itself and the signatures from the training set. By the limit laws of probability theory, the hidden signature is "close" to the mean value of the real signature after warping, thus assuring the proper level of invariance with the training set. The hidden signature can be calculated numerically.

Authorizing payment transactions is a straightforward area of application for signature verification systems. People are accustomed to using their signature to confirm the identity when paying with debit cards or credit cards. However, the verification method that is commonly used, namely the visual comparison performed by a cashier, leaves much to be desired. The sales staff is usually barely educated in the field of signatures verification and the risk of both false acceptance and false rejection may be significant. The above facts argue for the need of developing the automatic signature verification system for the purpose of authorizing payment transactions.

The paper is organized as follows. In the next Section 2, we show the proposed hidden signature estimation methodology. In Section 3, we introduce the verification algorithm that employs hidden signature for verification. Section 4 presents the application of the presented algorithm in authorizing payment transactions. Section 5 concludes the paper.

## 2. Hidden Signature Estimation

The hidden signature approach extends the least squares approach. In the least squares approach, the least squares model approaches the expected value as the number of observations increases to infinity, and for independent data it is also the most effective (in the statistical sense, e.g., it leads to a given error variance) for the least sample size. Since our approach is a conjunction of the least square modeling and optimal warping, the resulting model has some optimality properties, and it approaches the expected value of the signature, calculated with the use of warping. In other way, it approaches a form which is independent on particular signature instances. It is also independent on sampling frequency (which is not true for other warping models). This is why we may call it a }perfect instance of a signature.

In [7], we proposed two main directions for hidden signature estimation. First, *iterative point-by-point averaging*, and second, *evolutionary algorithms*. Each of those algorithms is constructed as an iterative procedure that alternates the warping steps and the averaging steps. Let $(G_1,\ldots,G_N)$ be a set of training signatures (the results presented here were calculated for $N = 10$). At each iteration stage, the algorithm corrects its approximation of the hidden signature $G^* = g^*(t),\ \ t = 1,2,\ldots,M_{G^*},\ \ M_{G^*} \in \mathbb{N}$, by minimizing the quality index:

$$\mathscr{V}_{\mathscr{H}} = \min_{j=1\ldots H} \sum_{i=1..N} \hat{\mathscr{D}}(G^*{}_j, G_i)\,, \tag{1}$$

where $\hat{\mathscr{D}}(G_1,G_2)$ denotes the dissimilarity measure between two signatures $G_1$ and $G_2$ (wide Eqs.(6), (7)).

Regardless of the estimation method, the resulting hidden signature depends on the coordinates selected to compute the dissimilarity measure and the warping path. This method was first proposed in [7] together with several others. All those methods yielded comparable results; the method presented was chosen as preferred because of its computation times, which were considerably shorter than the times of the other methods.

### *2.1. DTW Background*

Dynamic time warping is used to compute the optimal alignment, written in a form of *warping path w*, between variable length discreet sequences:

$$R = r(t),\ \ t = 1,2,\ldots,M_R,\ \ M_R \in \mathbb{N} \tag{2}$$

and

$$G = g(\tau),\ \ \tau = 1,2,\ldots,M_G,\ \ M_G \in \mathbb{N}\,. \tag{3}$$

A warping path $w$ is a parametric discrete curve parameterised by $\ell$, that aligns $R$ and $G$ via a point-to-point mapping (Fig. 1). The warping path can be defined as:

$$w(\ell) = [w_t(\ell)\ \ w_\tau(\ell)]^T;\ \ell \in 1,\ldots,L_w\,, \tag{4}$$

where

$$\begin{aligned} w_t(\ell) &\in \{t\},\ \ t = 1,\ldots,M_R\,,\\ w_\tau(\ell) &\in \{\tau\},\ \ \tau = 1,\ldots,M_G\,, \end{aligned} \tag{5}$$

where $w_t(\ell)$ and $w_\tau(\ell)$ are two parameterised by $\ell$ functions of aligned sequence indexes. Thus, $w(\ell)$ maps successive steps $\ell$ to a points $(t,\tau)$ where $1 \le t \le M_R$ and $1 \le \tau \le M_G$. The warping path length $L_w$ is a consequence of the minimisation process that minimise the overall distortion Eqs. (6 and 7).

The overall distortion $\mathscr{D}(R,G,w)$:

$$\mathscr{D}(R,G,w) = \sum_{\ell=1}^{L_w} d\left(r(w_t(\ell)), g(w_\tau(\ell))\right) \tag{6}$$

is based on the sum of local distances $d\left(r(w_t(\ell)), g(w_\tau(\ell))\right)$ between sequences elements at points, belonging to warping path. In the set of possible warping paths $w \in \mathrm{W}$, we can find the *optimal warping path* $\hat{w}$, such that its associated distortion (*dissimilarity measure*) is at a minimum:

$$\begin{aligned} \hat{\mathscr{D}}(R,G) &= \min_{w\in\mathrm{W}} \mathscr{D}(R,G,w)\,,\\ \hat{w} &= \arg\min_{w\in\mathrm{W}} \mathscr{D}(R,G,w)\,. \end{aligned} \tag{7}$$

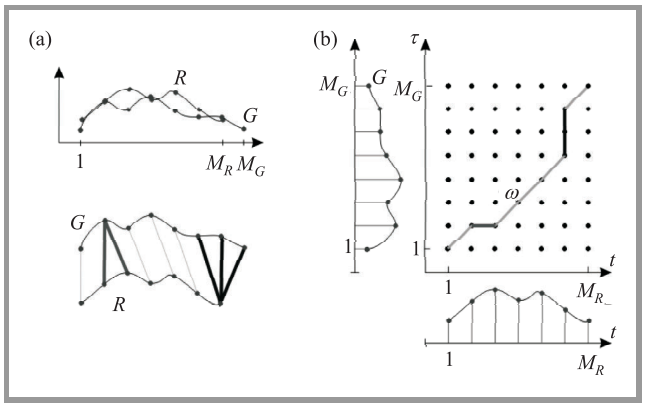



***Fig. 1.*** The compared curves, and its alignment visualization (a). Warping path on a two-dimensional grid: discrete warped time $\tau \in [0, M_G]$ versus discrete reference time $t \in [0, M_R]$ (b). The compared curves are also plotted (the reference curve is plotted vertically).

Allowing for differences caused by time warping, we still encounter differences in sequence values at the aligned time instances $d\,(r(w_t(\ell)), g(w_\tau(\ell)))$, where $t = 1, \ldots, M_R$ and $\tau = 1, \ldots, M_G$. These differences can be measured at the aligned times using typical distance metrics, for instance with the $L_2$ distance but it is not limited to:

$$d(a,b) = \|a-b\|_2 . \tag{8}$$

### *2.2. Transformation from One Time Space to Another*

The iterative point-by-point averaging (IPPA) is a method built on the concept of an average signature. If the signatures from a training set were the same length, the hidden signature could be simply computed as an average in each point.

However, like it was presented before, in each of the persons signature realizations the time flows differently. To overcome this problem we proposed the algorithm that transforms the training signatures into one common time space, thus obtaining signatures of the same length. The common time space, means that both signatures have the same length and the optimal warping path has only diagonal steps, which means that alignment between signatures is liner (Fig. 2b). The result of proposed transformation is presented in Fig. 2.

We assumed that we need a simple and fast transformation from one signature time space to a time space of second signature. Therefore we proposed the following procedure, transforming signature $G$ into a time space of signature $R$, thus obtaining $G' = g'(t), \;\; t = 1, \ldots, M_R$:

$$\hat{w} = \arg\min_{w \in \mathrm{W}} \mathscr{D}(R, G, w)\,, \tag{9}$$

$$g'(t;\hat{w}) = \frac{\sum_{\ell: w_t(\ell)=t} g(\hat{w}_\tau(\ell))}{|l : w_t(l) = t|}, , \quad t = 1, .., M_R\,,$$

where $t$ denotes the time of $R = r(t), \;\; t = 1, \ldots, , M_R$ and operator $|.|$ set cardinality. The illustration of this transformation is presented in Fig. 3.

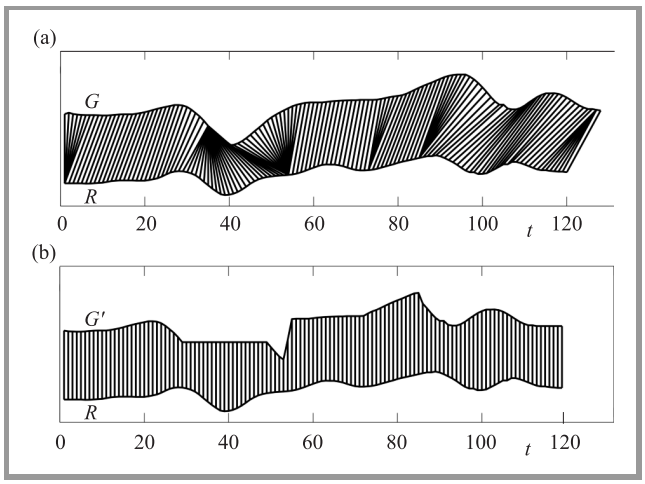



***Fig. 2.*** A transformation into a different signature time space: (a) the DTW nonlinear alignment of two time sequences $R$ and $G$; (b) the linear alignment between $R$ and $G'$ ($G$ in $R$ time space).

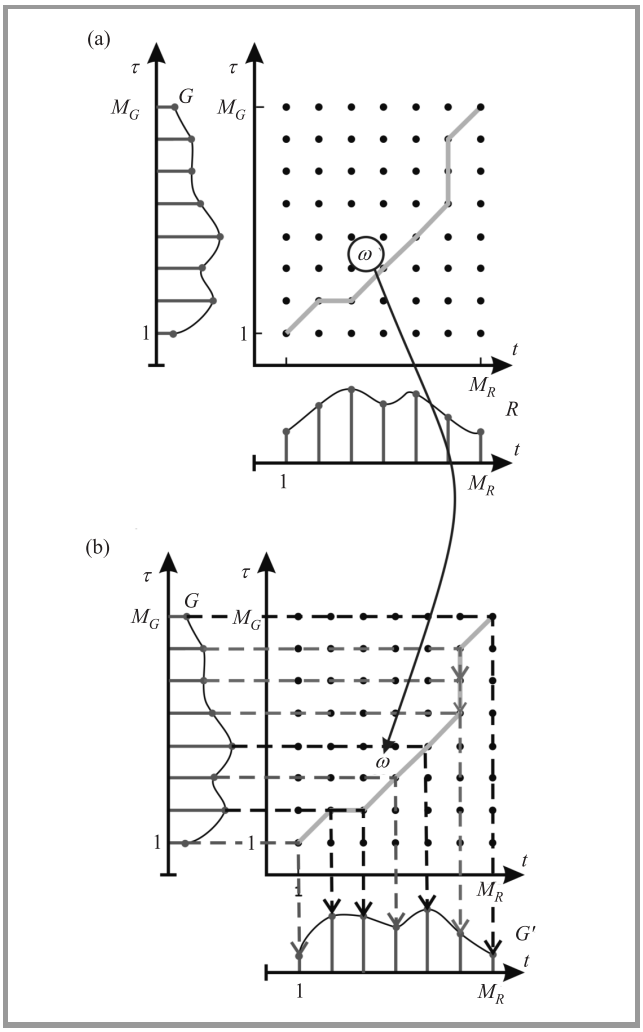



***Fig. 3.*** A transformation into a different signature time space: (a) calculation of the warping path $w$ between $R$ and $G$; (b) $G$ is transformed into the time space of $R$ using the warping path $w$.

### *2.3. Iterative Point-by-Point Averaging*

The Eq. (9) denotes a method of a signature transformation into a time space of another signature. However, it is assumed that the target time space is known. We assume that during each successive iteration only one estimation of hidden signature is calculated. Additionally, the hidden signature size is given by $M$, that can be set equally for the all users templates, or individually for each. In the beginning, we proposed to calculate the hidden signature length as the simple average (rounded down) of signatures from the training set, independently for each user, namely:

$$M = \frac{1}{N}\sum_{n=1}^{N} M_n\,. \tag{10}$$

Finally, we assume that in the initial iteration, all training signatures $\{G_1,\ldots,G_N\}$ are linearly graduated into a time length $M$, the assumed size of a hidden signature (Fig. 4).

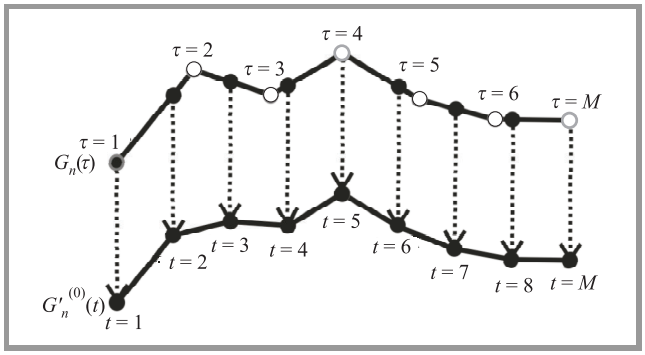



***Fig. 4.*** Linearly graduation into a time length $M$.

As the result we obtain $N$ training signatures of length $M$, namely $\left\{G_1'^{(0)},\ldots,G_N'^{(0)}\right\}$.
Detailed mathematical description has been omitted for the sake of simplicity. Using this starting set, it is possible to calculate the initial estimation of hidden signature $G^{*(0)}$:

$$g^{*(0)}(t) = \frac{1}{N}\sum_{n=1}^{N} g_n'^{(0)}(t), \quad t = 1,\ldots,M\,. \tag{11}$$

Then, in each successive iteration $k = 1,2,\ldots$ newly computed hidden signature approximation is used to calculate $N$ warping paths between itself and the signatures from the training set, thus minimizing the quality index:

$$\hat{w}_n^k = \hat{w}_n^k(G^{*(k-1)},G_n) = \arg\min_{w} \mathscr{D}(G^{*(k-1)},G_n,w), \quad n = 1,\ldots,N \tag{12}$$

then, the warping paths are used to transform the training signatures into the new hidden signature approximation $G^{*(k)}$ space using DTW. As the result of this operation, we obtain $N$ signatures with the lengths equal to the hidden signature length $M$. We can then calculate a new hidden signature approximation as a weighted mean for each point:

$$g^{*(k)}(t) = \frac{1}{N}\sum_{n=1\ldots N} g_n'(t;\hat{w}_n^{(k)}),, \quad t = 1,\ldots,M \tag{13}$$

where $g_n'$ is calculated from Eq. (9). This process repeats in every iteration, because changing warping paths imply changes in the hidden signature approximation. Visualization of this method is presented in Fig. 5.

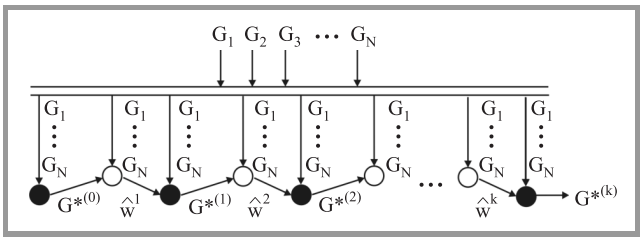



***Fig. 5.*** The iteration process of iterative point by point averaging with one minimization at each stage.

To apply the hidden signature concept, we used it in our two stage approach, which employs the DTW for computing the final score. In [7], we showed that replacing the representative of a training set with the hidden signature allows to achieve better verification results in the existing global on-line system. Here, we extend our approach to exploit the properties of the hidden signature by using its certain statistics.

## 3. Verification Algorithm – Error Signal Approach

Error signals approach is a simple method dedicated for on-line verification with hidden signature.

### *3.1. Template*

For the new approach to DTW, we proposed a parametric template:

$$\lambda_u = \{G^{*(u)},\sigma^{(u)}\}, \tag{14}$$

where $u$ denotes the user identifier. This template consist of the hidden signature $G^*$ and its corresponding standard deviation $\sigma$ sequence. Hidden signature is created form users $u$ training signatures $\{G_1,\ldots,G_N\}$. Standard deviation is calculated as the roots of mean-square errors between the hidden signature and the training signatures at each hidden signature point, independently for each feature $f \in \{\Delta x,\Delta y,p\}$ sequence:

$$\sigma_f(i) = \sqrt{\frac{1}{N}\sum_{n=1}^{N}(g_n'(i,f) - g^*(i,f))}\,, \quad i = 1,\ldots,M, \tag{15}$$

where feature sequences are used according to definition of signature for each feature $f$ sequence as:

$$G_f = g(\tau,f), \quad \tau = 1,2,\ldots,M_G, \quad M_G \in \mathbb{N}. \tag{16}$$

Table 1
On-line systems comparison

| Authors | Year | FAR [%] | FRR [%] | EER [%] | No. of users | Database |
|---|---|---|---|---|---|---|
| Quan *et al.* [9] | 2006 | | | 7 | 100 | MCYT_Online |
| Miguel-Hurtado, O. *et al.* [5] | 2007 | | | 8 | 100 | MCYT_Online |
| Garcia-Salicetti, Dorizzi [10] | 2007 | | | 3.37 | 100 | MCYT_Online |
| Guru, Prakash [11] | 2007 | 9.16 | 5.42 | 5.3 | 100 | MCYT_Online |
| Galbally, Ortega-Garcia [12] | 2007 | | | 3.5 | 330 | MCYT_330 |
| Faundez-Zanuy [13] | 2007 | | | 5.4 | 330 | MCYT_330 |
| Nanni and Lumini [14] | 2008 | | | 5.2 | 100 | MCYT_Online |
| Yanikoglu and Kholmatov [15] | 2009 | | | 7.22 | 100 | MCYT_Online |
| **This work** | 2010 | **1.74** | **1.82** | **1.72** | 100 | MCYT_Online |

The features $f$ were selected, due to our previous experiments, presented in [8]. Additionally, we want the system to work with LCD tablets, and they allows only for $x$, $y$ and pressure acquisition.

### *3.2. Verification Stage*

In the verification stage, a signature in question $G_q$, after the transformation into the hidden signature space $G'_q$ Eq. (9), can be standardized for every feature, by point-by-point subtraction of the hidden signature $G^*$ as the sequence of mean values, and division by the standard deviations $\sigma$, given by Eq. (15):

$$g''_q(i,f) = \frac{g'_q(i,f) - g^*(i,f)}{\sigma_f(i)}\ ,\ i = 1,\ldots,M\,, \qquad (17)$$

where $f \in \{\Delta x, \Delta y, p\}$. The new resulting standardized sequences $G''_{q,\Delta x}$,$G''_{q,\Delta y}$,$G''_{q,p}$, are called the error signals, because they represent a normalize errors between signature in question and hidden signature at each point of a sequence.
We propose the final $SCORE^H$ used in at the verification stage. For the signature in question $G_q$ its selected values are used for the scores calculations:

$$SCORE^H(G''_q; G^*, \sigma) = \frac{1}{\|\left[s^2(G''_{q,\Delta x})\ s^2(G''_{q,\Delta y})\ s^2(G''_{q,p})\right]^T\|_2}\,, \qquad (18)$$

where: $\overline{Q}$ denotes the arithmetic mean value of the $Q$ sequence, $s^2(Q)$ denotes the standard variance of non-zero values of $Q$ sequence, and $\|q\|_2$ denotes the Euclidean norm of $q$.
The logical sense of these scores is that the error signal $G''_f$ is a resulting signal of normalization with a use of the hidden signature. This means, the closer is the verified signature $G^Q$ to the hidden signature, the lower are the values at each point of the error signal.
The final stage of verification is very simple. The signature in question $G_q$ is accepted if following condition is satisfied:

$$SCORE^H \leq \theta\,.$$

The threshold $\theta$ is calculated during the estimation phase.

### *3.3. Testing Methodology and Results*

The tests were conducted on the MCYT on-line database [16] (MCYT_Online for 100 users). Our methodology was to evaluate in simulated real-world conditions. For this goal, the database was divided into two parts. The *estimation data* used for the estimation phase included 40 persons, and the remaining 60 person data was used as the *testing data*. The first part of the data was used for the estimation of the equal error rate (EER) and the corresponding threshold $\theta$. In the testing phase, we checked the repetitiveness of our approach. We used the remaining 60 persons of the 100-people database for testing data, and fixed the same threshold $\theta$ as that adjusted for the first database part. In practice, it was the level of threshold $\theta$ for the EER obtained in the estimating phase. We then obtained false acceptance rate (FAR) and false rejection rate (FRR). If results of FAR and FRR are close to the EER obtained for the first part of the database, then they can be trusted. The presented results were obtained after 1000 different distributions of database for phases:

- estimating phase: $EER = 1.72 \pm 0.25\%$;
- testing phase:
  - $FRR = 1.82 \pm 0.86\%$,
  - $FAR = 1.74 \pm 0.3\%$ (skilled forgeries),
  - $FAR = 0.06 \pm 0.05\%$ (random forgeries).

## 4. Implementation

We propose to employ the automatic signature verification system based on the hidden signature for the purpose of authorizing payment transactions. In NASK Biometric Laboratories, we designed an adaptation of the verification algorithm presented in Section 3 to the platform of payment terminals. Together with the MCX Systems Company that specializes in developing software for payment terminals, we prepared the implementation that can be run directly on the terminal used at a *point of sale* (POS). We also built a prototype of a complete biometric system for authorizing

payment transactions. The system consists of three components: an enrollment stand, a POS terminal with biometric signature verification and a technique for storing biometric templates.
The automatic signature verification system can replace the cashier in performing the verification or can serve as a decision support system for the cashier.

### *4.1. Hardware Platform*

Payment terminals make a demanding programming environment due to a limited computational power, reduced memory and (usually) only a software support for floating point operations.




***Fig. 6.*** Terminal with tablet.

For the purpose of our prototype system, we used the Ingenico i5100 terminal. It is equipped with ARM7TDMI processing unit with computational power of about 30 MIPS. There is 2 MB of RAM memory available, yet only 256 kB can be used by the algorithm due to the requirements of other applications and the operating system. The terminal is connected with the Wacom STU-500 LCD signature tablet that is specially designed to capture digital signatures.
The specification of the terminal, together with time constraints related to payment transactions, required developing a specialized implementation of the verification system.

### *4.2. DTW Implementation*

An effective implementation of the DTW algorithm is crucial from the performance point of view. The comparison of two signatures in a given space is relatively cheap, yet finding the optimal warping path requires solving a dynamic programming task with the quadratic time and space complexity.
In order to overcome the problem, we performed numerous steps that allowed us to reduce time and memory requirements of the DTW algorithm. First, we limited the maximum length of signatures to 500 points (longer signatures were subjected to downsampling). Second, we employed the Sakoe-Chuba Band constraint [17] to limit the number of cells that are evaluated in the cost matrix. We chose the window size according to a trade-off between memory requirements and the verification quality (only a minor increase of the ERR was observed after applying the constraint on the MCYT database). Finally, we reduced the memory requirements by using two moving buffers combined with a bitmap of the cost matrix (Fig. 7).

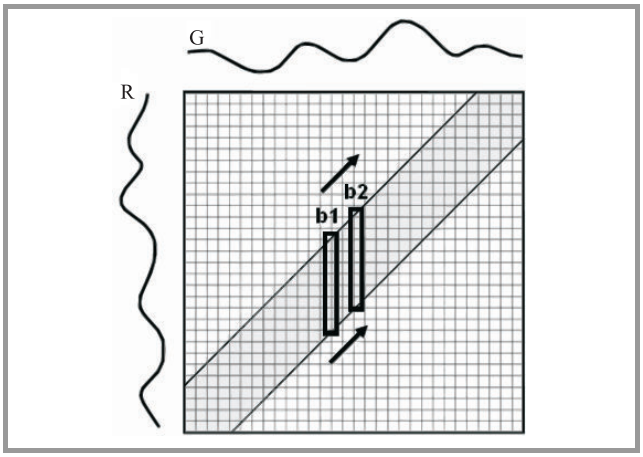



***Fig. 7.*** Visualization of a discrete DTW implementation using a cost matrix. The number of cells that are evaluated is limited and two moving buffers are used to reduce the memory requirements.

The idea of using moving buffers is the following. Dynamic programming algorithms can be visualized as "filling an array". In DTW, each cell of the array represents the total cost of the best warping path reaching this cell (i.e., the partial warping path ending in this cell). The array is filled starting from the lower-left corner. After the whole array is filled, the value in the upper-right corner represents the minimal cost of the whole warping path. It is possible to reconstruct the best warping path by moving back from the upper-right corner, choosing always the neighboring cell with the lowest (partial) cost value (according to the Bellman's Dynamic Programming rule).
However, one may notice that it is not necessary to remember the whole array. Suppose that from a given cell we can move only in three directions (right, up and diagonal upper-right). In order to fill the $n$th column of the array (starting form the lowest cell), we need only to know the $(n-1)$th column. Hence, instead of keeping the whole cost matrix in memory, we may use two moving buffers, each one to store one column. After performing the whole run, we obtain the same optimal cost in the upper-right corner of the matrix. In addition, we need to store the information that would allow us to reconstruct the best warping path. We achieve this by remembering in each cell the information about the previous cell of the warping path. We code such information using two bits (as there are only three directions possible: left, down and diagonal lower-left) and

we store it an additional bitmap. As a result, instead of keeping all values of the cost matrix in memory (usually floating point values with double precision), we need only the memory for two moving buffers and the bitmap.

### *4.3. Performance and Memory Requirements*

A combination of the techniques described in the previous section allowed us to obtain the reasonable memory requirements presented in Table 2.

Table 2
Memory requirements

| Source of the requirement | Memory [kB] |
|---|---|
| Buffer for the raw signature | 18 |
| Buffer for the template | 6 |
| Temporary buffer for an exemplary signature | 3 |
| Temporary buffer for a candidate signature | 3 |
| Buffer for the warping path | 5 |
| Bitmap | 20 |
| Two moving buffers | 1 |
| **Total:** | **56** |

The effective implementation of the DTW algorithm determined the following running times on the terminal:

- verification of a single signature below 5 s,
- template creation (from 5 signatures): 2–5 min, depending on the signatures length.

From the practical applications perspective, only the verification time is important as the enrollment procedure is usually performed outside the terminal. The verification time of 5 s is satisfactory as far as payment transactions are concerned. Moreover, this time may be further reduced using more powerful modern terminals.
The correctness of the results has also been verified: we performed tests on the MCYT database using the terminal for verifications. We observed some minor differences in the obtained scores (resulting from different floating point implementations), yet the verification results were identical on both PC and the payment terminal platforms.

### *4.4. Prototype of the System*

Basing on the running times obtained in Subsection 4.3, we proposed the following operational scenario: signature templates should be created using a specialized enrollment stand (e.g., equipped with a PC computer) and payment terminals should only perform signatures verification during payment transactions. According to this scenario, we build a prototype biometric system for authorizing payment transactions (Fig. 8).
The first element of the system is an enrollment stand. The stand is composed of a PC computer, a smart card reader

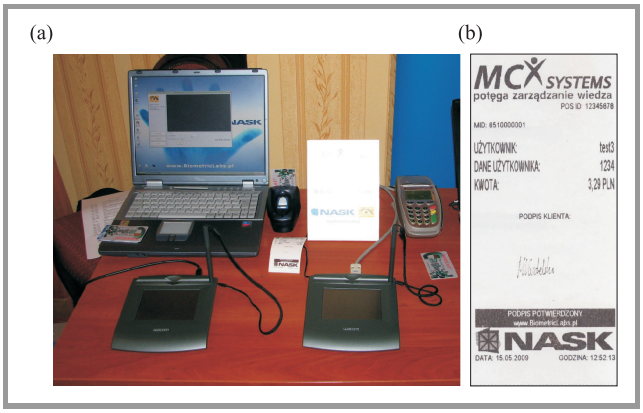



***Fig. 8.*** Prototype of the system: a demonstration stand (a) and a sample printout (b).

and the STU-500 tablet for gathering signatures. The process of enrollment is controlled by a specialized application. The functionality of the application includes:

- creating templates basing on exemplary signatures and performing signatures verification,
- analyzing the template's consistency,
- saving the template in the internal database or on a smart card based on the JavaCard technology,
- making templates accessible via the HTTP protocol,
- visualizing signatures in real-time during writing.

The second element of the system is a payment terminal with a similar STU-500 tablet. The application installed on the terminal allows to simulate a real payment transaction at a point of sale: the cashier scans the customer's credit card, enters the amount of the transaction and waits for the authorization. At the end, the confirmation of the transaction is printed (Fig. 8b).
The transaction is authorized using automatic signature verification. The application running on the terminal downloads the template, captures the signature in question from the tablet and performs the verification. Both magnetic cards and smart cards are supported by the terminal.
The last element of the system is a mechanism of storing and transferring biometric templates. There are two alternative ways of storing templates, namely an *on-line scheme* and an *off-line scheme*. In the on-line scheme, templates are stored in a database and can be accessed via the HTTP protocol. The terminal scans the unique number of the customer's magnetic card, connects to the database using Ethernet connection, downloads the customer's template and performs verification. In the off-line scheme, the customer's template is stored in the customer's smart card. The terminal reads the template directly from the card and performs verification.

## 5. Conclusions

The presented novel approach to signature verification draws from the hidden signature concept and differs from the traditional DTW-based methods. This approach is based on replacing the template signatures with an artificial signature – the hidden signature.

The proposed method can be regarded as model-based, since the hidden signature can be thought as an approximation of a perfect signature of a given person.

The verification algorithm is based on computing error signals between the signature in question and the model. In comparison to results of other systems, obtained on the MCYT database (see Table 1), the complex system for hidden signature results is placed very high.

While the results obtained in tests on the MCYT database are promising, the method is mostly based on an engineering approach to error signals processing, and could benefit from more sophisticated comparison algorithms.

The overall simplicity allows this method to be used in mobile or embedded systems. One of them is the system for signature verification in authorizing payment transactions. The effective implementation of the proposed approach allowed to meet the memory and computational power constraints of payment terminals, as well as the time constraints of payment transactions. The automatic signature verification based on the hidden signature can be performed directly on the payment terminal at a point of sale.

## Acknowledgements

This paper has been financed by the Ministry of Science and Higher Education grant OR00 0026 07 "*A platform of secure biometrics implementations in personal verification and identification*".

## References

[1] N. Herbst and C. Liu, "Automatic signature verification based on accelerometry", in *IBM J. Res. Develop.*, pp. 245–253, 1977.

[2] D. Sakamoto, H. Morita, T. Ohishi, Y. Komiya, and T. Matsumoto, "On-line signature verifier incorporating pen position, pen pressure and pen inicliation trajectories", in *Proc. 3rd Int. Conf. AVBPA 2001*, Halmstad, Sweden, 2001, pp. 318–323.

[3] A. Pacut and J. Putz-Leszczyńska, "Dynamic time warping in subspaces for on-line signature verification", in *Proc. 12th Conf. Int. Graphonom. Soc. IGS 2005*, Salerno, Italy, 2005, pp. 108–112.

[4] A. Kholmatov and B. Yanikoglu, "Identity authentication using improved online signature verification method", *Pattern Recogn. Lett.*, vol. 26, no. 15, pp. 2400–2408, 2005.

[5] O. Miguel-Hurtado, L. Mengibar-Pozo, M. Lorenz, and J. Liu-Jimenez, "On-line signature verification by dynamic time warping and gaussian mixture models", in *41st Ann. IEEE Int. Carnahan Conf. Secur. Technol.*, Ottawa, Canada, 2007, pp. 23–29.

[6] B. Fang, C. Leung, Y. Tang, K. Tseb, P. Kwokd, and Y. Wonge, "Off-line signature verifcation by the tracking of feature and stroke positions", *Pattern Recogn.*, vol. 36, pp. 91–101, 2003.

[7] A. Putz-Leszczyńska, J.and Pacut, "'Hidden signature' – a new solution for on-line verification ", in *42nd Ann. IEEE Int. Carnahan Conf. Secur. Technol.*, Prague, Czech Rebulic, pp. 68–73, 2008.

[8] J. Putz-Leszczyńska, "On-line signature verification using dynamic time warping with positional coordinates", in *Proc. SPIE – Vol. 6347, Photonics Applications in Astronomy, Communications, Industry, and High-Energy Physics Experiments*, R. S. Romaniuk, Ed., 2006 (doi: 10.1117/12.714578).

[9] X.-L. X. M. R. L. Z.-H. Quan, D.-S. Huang and T.-M. Lok, "Spectrum analysis based on windows with variable widths for online signature verification", in *Proc. 18th Int. Confe. Pattern Recogn. ICPR 2006)*, Hong Kong, China, 2006, vol. 2, pp. 1122–1125.

[10] B. L. Van, S. Garcia-Salicetti, and B. Dorizzi, "On using the viterbi path along with hmm likelihood information for online signature verification", *IEEE Trans. Sys. Man. Cybernetics, Part B: Cybernetics*, vol. 37, no. 5, 2007.

[11] D. Guru and H. Prakash, "Symbolic representation of on-line signatures", in *Proc. Int. Conf. Comput. Intellig. Multimedia Appl. 2007*, Sivakasi, India, 2007, pp. 313–317.

[12] J. Galbally, J. Fierrez, M. Freire, and J. Ortega-Garcia, "Feature Selection Based on Genetic Algorithms for On-Line Signature Verification", in *Proc. IEEE Worksh. Autom. Identif. Adv. Technol. 2007*, Alghero, Italy, 2007, pp. 198–203.

[13] M. Faundez-Zanuy, "On-line signature recognition based on VQ-DTW", *Pattern Recogn.*, vol. 40, no. 3, pp. 981–992, 2007.

[14] L. Nanni and A. Lumini, "A novel local on-line signature verification system,", *Pattern Recogn. Lett.*, vol. 29, no. 5, pp. 559–568, 2008.

[15] B. Yanikoglu and A. Kholmatov, "Online signature verification using fourier descriptors", *EURASIP J. Adv. Sig. Proces.*, 2009.

[16] J. Ortega-Garcia, J. Fierrez-Aguilar, D. Simon, J. Gonzalez, M. Faundez-Zanuy, V. Espinosa, A. Satue, I. Hernaez, J.-J. Igarza, C. Vivaracho, D. Escudero, and Q.-I. Moro, "MCYT baseline corpus: a bimodal biometric database", *IEE Proc.-Vis. Image Sig. Process.*, vol. 150, no. 6, pp. 3412–3426, 2003.

[17] H. Sakoe and S. Chiba, "Dynamic programming algorithm optimization for spoken word recognition", *IEEE Trans. Acoust., Speech Sig. Proces.*, vol. 26, no. 1, pp. 43–49, 1978.



**Joanna Putz-Leszczyńska** received her M.Sc. in 2004 from the Faculty of Electronics and Information Technology of the Warsaw University of Technology, Poland. She is presently a Ph.D. student at the Institute of Control and Computation Engineering at the Faculty of Electronics and Information Technology of the Warsaw University of Technology. Since 2003 she works as a Research Assistant at Biometric Laboratory of Research and Academic Computer Network NASK. She is interested in biometrics, identification, security systems and global optimization heuristics.

e-mail: J.Putz@elka.pw.edu.pl
Research and Academic Computer Network (NASK)
Wąwozowa st 18
02-796 Warsaw, Poland

Institute of Control and Computation Engineering
Warsaw University of Technology
Nowowiejska st 15/19
00-665 Warsaw, Poland



**Michał Kudelski** received his M.Sc. in 2005 from the Faculty of Electronics and Information Technology of the Warsaw University of Technology, Poland. He is presently a Ph.D. student at the Institute of Control and Computation Engineering at the Faculty of Electronics and Information Technology of the Warsaw University of Technology. Since 2008, he is also a Research Assistant at the Biometric Laboratory of Research and Academic Computer Network NASK. He is interested in artificial intelligence, adaptive systems, biometrics and related areas.

e-mail: M.Kudelski@elka.pw.edu.pl
Research and Academic Computer Network (NASK)
Wąwozowa st 18
02-796 Warsaw, Poland

Institute of Control and Computation Engineering
Warsaw University of Technology
Nowowiejska st 15/19
00-665 Warsaw, Poland

Paper

# Simulation Model of Biometric Authentication Using Multiagent Approach

Adrian Kapczyński and Tomasz Owczarek

**Abstract—In this article authors present the concept of application of multiagent approach in modeling biometric authentication systems. After short introduction, we present a short primer to multiagent technology. Next, we depict current state of the art related to biometrics combined with multiagent approach. In the next part of the work we present four exemplary simulation models of biometric authentication environments as well as the results of their examination.**

***Keywords—biometrics, multiagent, multimodal, simulation.***

## 1. Introduction

Current level of requirements related to strong authentication mechanisms are either fulfilled by constructing single, strong authentication factor solution or a solution that utilizes the multifactor approach. Analogically, in case of user verification or identification, biometric methods are widely applied as single modal or multimodal systems. Contemporary theoretical and empirical approaches to construct biometric systems focus on converging different authentication factors, algorithms, protocols and equipment in networked environments. This emphasize the emerging role of methods and tools used to model, simulate and analyze networked and more complex then single instance systems. Therefore, the need of performing analysis from different abstraction levels systems can be satisfied by providing apparatus operating not only from micro, but also macro perspective. Complete biometric system models shall combine technical and non-technical (human) element. Such approach can be found in many modeling languages, even in BANTAM (biometric and token modeling language) language, dedicated to biometric domain. BANTAM however does not provide the capability of observing the active, environment of biometric systems. In this article authors propose the use of multiagent systems as simulation tools of biometric authentication systems. The authentication processes are realized between users (agents having the need of being authenticated) and the authentication center. This concept we illustrate by four simulation models of single- and multibiometric authentication environments. In next part of the work we present a primer on multiagent systems.

## 2. Multiagent Systems

Agent-based model can be simply defined as a simulation made up of agents, objects or entities that behave autonomously [1]. The shortest definition of the term agent can be found in [2], where it is described as a proactive object. These two definitions contains two main features of agency:

- proactiveness: agent can take initiative, it does not simply wait for a signal to start acting but it is able to undertake actions in order to fulfill its goals;
- autonomy: agent is an autonomous entity which can operate without direct control.

Apart from these Wooldridge and Jennings [3] provide two more essential agents' properties:

- reactivity: agents respond to signals perceived from their environment;
- social ability: agents interact with each other, they communicate, cooperate and even compete.

According to [4] the indispensable feature of any agent is its (temporally) continuity which means that it is a continuously running process. Franklin and Graesser also propose a taxonomy of agents which at the highest level divides them into biological agents (human and animal), robotic agents and computational agent (computer program). Agent's definition varies and different features are emphasized depending on authors [5], [6]. But they all agree that an agent is situated in some environment and able to make autonomous decisions [7].
As it is pointed out in [6], [4], [8] one cannot talk about agent without environment in which it is situated. According to the definition from [5] an agent is "anything that can be viewed as perceiving its environment through sensors and acting upon the environment through actuators". A schema of an agent interaction with its environment is shown in Fig. 1.

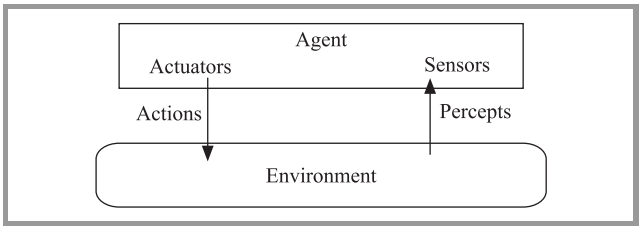



***Fig. 1.*** Agent and its interaction with environment.

The environment determines an agent; placing an agent in a different environment often stops it from being an agent (e.g., a robot with only visual sensors placed in

a dark room) [4]. Single agent environment are very rare. In fact, in multiagent systems community exists a popular slogan that "there's no such thing as a single agent system" [6, pp. 105]. Complexity and unpredictability of real world situations often require a combination of specialized problem solvers (agents) which cooperate in order to find a solution to problems that are far beyond their individual capabilities [9]. When there are more than one agent then we deal with multiagent system and agent's environment is constituted by all other agents. Agent-based models are useful in modeling complex, nonlinear systems. But they can be also treated as generalizations of analytical models [10], especially when the system modeled consists of numerous interacting autonomous objects. This is why we chose agent-based approach to the specified problem. In next part of the work we present the current approaches in combing biometrics with multiagent methodology.

## 3. State of the Art

M. Abreu and M. Fairhurst [11] focus on evaluation of multimodal structures and they investigate how fundamentally different strategies for implementation can influence the degree of choice available in meeting chosen performance criteria. In particular they implement computational architecture based on a multiagent approach which goal is to achieve high performance. In their work authors also propose and evaluate a novel approach to implementation of a multimodal system based on negotiating agents.

R. Meshulam et al. [12] introduced the concept of multiagent framework which works in large-scale scenarios and is capable of providing response in real time. The input for the framework is biometric data acquired at a set of locations and that data is used to point out individuals who act accordingly to pattern defined as "suspicious". Authors present two interesting scenarios in order to demonstrate the usefulness of their framework. In first scenario, the goal of the system is to point to individuals who visited a sequence of airports. In this scenario, face biometrics is applied. The goal in the second scenario is to point out individuals who called a set of phones. In the second scenario the use of speaker biometrics is proposed.

G. Ali, N. Shaikh and Z. Shaikh note that traditional insider threat protection models are not efficient and that there is a need of an autonomous and flexible model against insider threat [13]. In the paper authors present agent-based model that monitors behavior of the authorized users. So, the agents are responsible for recording all actions of the authorized user and deliver all recorded data to the main agent for processing and decision making.

Finally, G. Chetty and D. Sharma present an application of agent technology to the problem of face identification, which is performed robustly in even difficult environmental conditions [14]. Authors apply new composite model consisting of multiple layers that is supported by integration with agent based paradigm. Obtained experimental results are suggesting further investigations in application of agent methodology in building multimodal biometric systems. Other similar approaches can be found in [15], [16].

We can notice that agent-based concept is applied in order to enhance the performance of single instance (but not only single modal) biometric systems or to provide capabilities of detection of inexpedient behavior from security point of view. In this work we proposed complementary approach which relies on use of agent-based paradigm for simulation enabling macro scale analysis of interactions between authenticator and authenticatee. In next part of the paper we present the foundation of agent-based biometric authentication as well as we illustrate it by providing three examples.

## 4. Agent-Based Biometric Authentication

Our models were created in NetLogo, a multiagent programmable modeling environment [17]. This allowed for rapidly implementation of the model's variants and made all results scientifically reproducible. There are three types of agents in proposed model: users, authentication centers, and experts (Fig. 2).

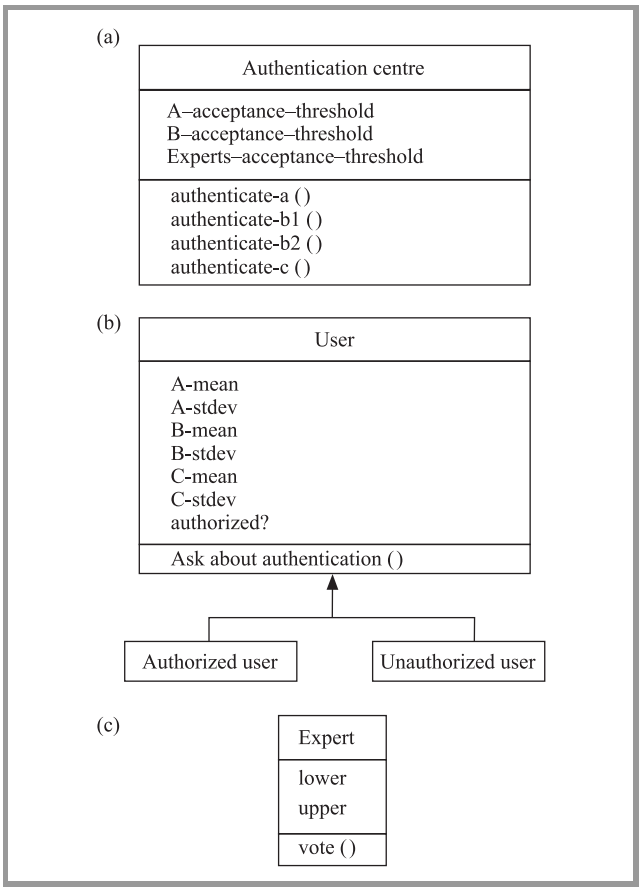



***Fig. 2.*** Agent types: (a) authentication centre, (b) user, (c) expert.

Users (agents being authenticated) are divided into genuine users (authorized) and impostors (unauthorized). Distinction between those agents is performed by use of attribute

`Authorized?` (taking values true or false). Each user has three modalities: A, B and C. Those biometric characteristics are here represented by matching scores, which are an output of comparison module performing action on enrollment and verification templates. The enrollment template is created during the first interaction with biometric system and arises from raw biometric data which is transformed into its mathematical representation. The reference template is created every time the user wants to be authenticated basing on provided raw biometric data. Each agent has for each modality one corresponding matching score described using two attributes: the average and standard deviation. Matching scores are random variables with normal distribution. In addition to the operations shown in Fig. 2 `ask about authentication()`, all users have the instructions also responsible for their movement to and from the authentication center (they are not relevant to the described problem). Authentication centers have attributes which are acceptance thresholds and operation `authenticate()`. The experts occur only in third variant of simulation models. Description of their attributes are presented in further part of this article.

Overall, the simulation process is as follows. After the opening initialization of agents (users stay in randomly deployed in a two-dimensional space, inside which there is a authentication center), any user at random intervals goes to the authentication center. Upon arrival agent delivers its matching score (for each modality the system generate a random value of a random variable). On that basis the center formulates decision: accept or reject. Regardless of the result, the user returns to its initial position and looks forward to the next signal of going to the authentication center.

Basing on formulated above general foundings, four simulation models of biometric authentication systems were constructed.

**Model a**. Multiagent system with given number of authorized and not-authorized agents and with one authentication centre. The authentication centre during authentication process receives from the authenticated agent its matching score of modality A which is compared to global threshold TA. The output of the comparison is the basis of the decision about acceptance (in case the matching score is equal or grater than threshold) or rejection (in case the matching score is lesser than threshold).

**Model b1**. Multiagent system with given number of authorized and not-authorized agents and with one authentication centre. The authentication centre during authentication process receives from the authenticated agent its matching score of two modalities: A and B. The matching scores are compared with appropriate global thresholds TA and TB respectively. The outputs of performed comparisons are the basis of the final decision. The system accepts the users if both matching scores are not lesser than given thresholds (AND rule) else it rejects the user.

**Model b2**. Multiagent system with given number of authorized and not-authorized agents and with one authentication centre. The authentication centre during authentication process receives from the authenticated agent its matching scores of two modalities: A and B. The matching scores are compared with appropriate global thresholds TA and TB respectively. The outputs of performed comparisons are the basis of the final decision. The system accepts the users if at least one matching score is not lesser than given threshold (OR rule) else it rejects the user.

**Model c**. Multiagent system with given number of authorized and not-authorized agents and with one authentication centre. The authentication centre during authentication process receives from the authenticated agent its matching scores of three modalities: A, B and C. The authentication process is carried out by three experts and each expert has its own set of two thresholds (upper limit and lower limit). If matching score is greater or equal than upper limit than user is accepted else if matching score is lesser or equal than lower limit then user is rejected else the decision is inconclusive. Experts has predefined set of thresholds (presented as s triple: expert number, upper limit, lower limit): 1, 0.7, 0.3; 2, 0.5, 0.1; 3, 0.8, 0.7. Each expert generates output: +1 – in case the logical condition related to upper limit is true; -1 in case the logical condition related to lower limit is true; 0 – in case the previous conditions are false. Final decision is based on summed output divided by number of experts which is compared against the expert-acceptance-threshold TE.

Presented models have been implemented and examined in prepared simulation environment.

# 5. Simulation Environment and Simulation Results

All described models have been implemented in NetLogo environment.

### *5.1. Simulation environment preparation*

First, we have implemented:

- initialization procedures (setup-users, setup-centers, setup-experts),
- main procedures reflecting the four models (authenticate-a, authenticate-b1, authenticate-b2, authenticate-c),
- supporting procedures (setup, go, do-plots, etc.).

Next we have prepared the interface which consists of the following input controls:

- setup – which resets the values of environment controls to defaults,
- go – which starts the simulation,

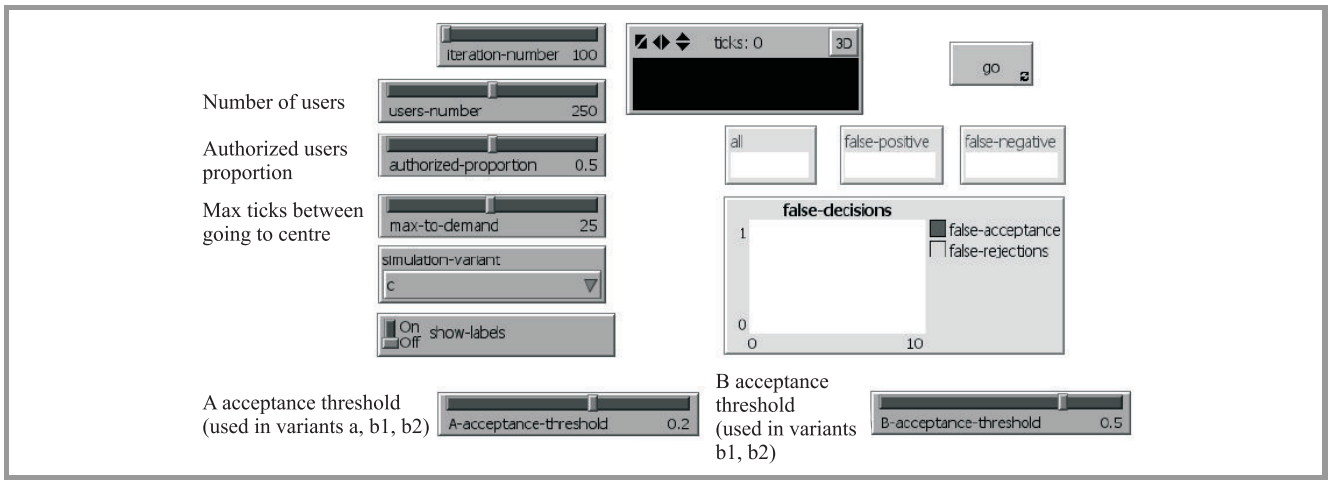



***Fig. 3.*** Simulation environment.

- iteration number – which enables definition of length of simulation (expressed in ticks),
- users number – which enables definition of size of whole population,
- authorized proportion – which enables definition of structure of whole population,
- max-to-demand – which enables definition of the maximum number of ticks between going to authentication center,
- simulation variant – which enables choice of one of four implemented simulation models: a, b1, b2 and c,
- show labels – which enables switching on or off labels of the agents,
- auth-A-mean – which enables definition of average value of matching scores for genuine users using modality A),
- auth-A-stdev- which enables definition of standard deviation of matching scores of genuine users using modality A),
- auth-B-mean – which enables definition of average value of matching scores of genuine users using modality B),
- auth-B-stdev – which enables definition of standard deviation of matching scores of genuine users using modality B),
- auth-C-mean – which enables definition of average value of matching scores of genuine users using modality C),
- auth-C-stdev – which enables definition of standard deviation of matching scores of genuine users using modality C),
- unauth-A-mean – which enables definition of average value of matching scores of impostors using modality A),
- unauth-A-stdev – which enables definition of standard deviation of matching scores of impostors using modality A),
- unauth-B-mean – which enables definition of average value of matching scores of impostors using modality B),
- unauth-B-stdev – which enables definition of standard deviation of matching scores of impostors using modality B),
- unauth-C-mean – which enables definition of average value of matching scores of impostors using modality C),
- unauth-C-stdev – which enables definition of standard deviation of matching scores of impostors using modality C),
- A-acceptance-threshold – which enables definition of threshold for modality A,
- B-acceptance-threshold – which enables definition of threshold for modality B,
- C-acceptance-threshold – which enables definition of threshold for modality C.
- Experts-acceptance-threshold - which enables definition of threshold for preparing the final decision on the basing of votes of the experts.

Moreover we provide the output controls:

- World – which displays the simulation in 2D or 3D,
- Plot – which displays the false acceptance rate and false rejection rate,
- Reporter 1 – which displays number of performed authentications,
- Reporter 2 – which displays number of false acceptance decisions,
- Reporter 3 – which displays number of false rejection decisions.

The simulation environment window which combines enumerated controls is presented in Fig. 3.

### *5.2. Simulation Results*

Each implemented model was executed being previously prepared according to specified values of given controls. During simulations the changes occurring in the environment were easily to be observed and they were logged in a comma seperated values file. Obtained values were used to prepare visualizations.
Here we present the initial values of given controls:

- iteration number = 100,
- users number = 250,
- authorized proportion = 0.5,
- max-to-demand = 25,
- show labels = off,
- auth-A-mean = 1.0,
- auth-A-stdev = 0.5,
- auth-B-mean = 1.0,
- auth-B-stdev = 0.5,
- auth-C-mean = 1.0,
- auth-C-stdev = 0.5,
- unauth-A-mean = –1.0,
- unauth-A-stdev = 0.5,
- unauth-B-mean = –1.0,
- unauth-B-stdev = 0.5,
- unauth-C-mean = –1.0,
- unauth-C-stdev = 0.5.

We conducted four group of simulations:

- First set of simulations were based on simulation model a. We were observing the false acceptance indicator (FA) and false rejection indicator (FR) in three different configurations of threshold TA (TA = 0.3, TA = 0.5 and TA = 0.7).
- Second set of simulations were based on simulation model b1. Again, se were observing the false acceptance indicator (FA) and false rejection indicator (FR) in three different configurations of threshold TA (TA = 0.3, TA = 0.5 and TA = 0.7) and threshold TB (TB = 0.3, TB = 0.5 and TB = 0.7).
- Third set of simulations were based on simulation model b2. Again, se were observing the false acceptance indicator (FA) and false rejection indicator (FR) in three different configurations of threshold TA (TA = 0.3, TA = 0.5 and TA = 0.7) and threshold TB (TB = 0.3, TB = 0.5 and TB = 0.7).
- Fourth set of simulations were based on simulation model c. We were observing the false acceptance indicator (FA) and false rejection indicator (FR) in three different configurations of experts-acceptance-threshold TE (TE = 0.3, TE = 0.5 and TE = 0.7).

In Fig. 4 we present how the FA and FR indicators were changing in simulated environment exploiting model a, for different (discrete) values of threshold TA.

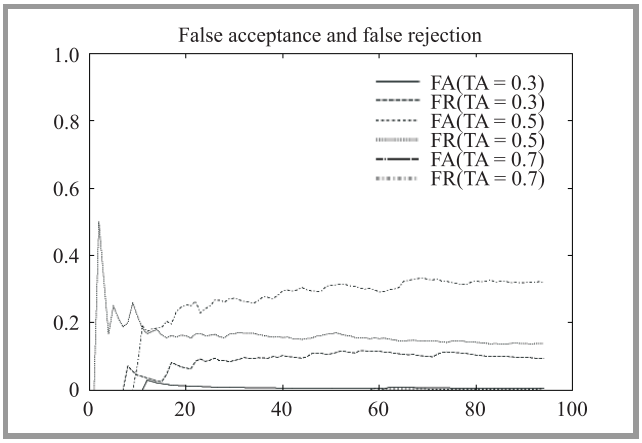



***Fig. 4.*** Simulation results using model a.

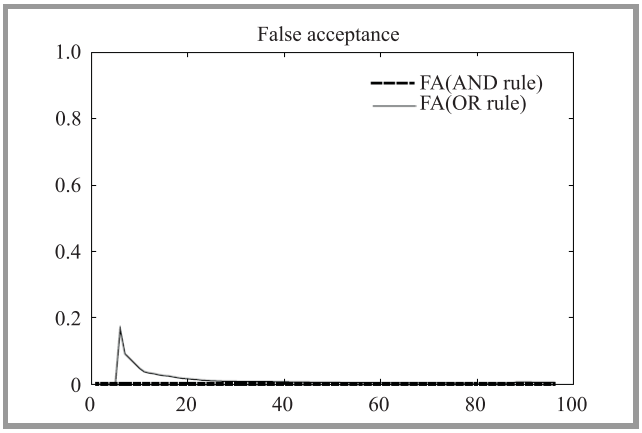



***Fig. 5.*** Simulation results using models: b1 and b2 (FA indicator).

In Fig. 5 we compare FA indicators in simulated environment using models: b1 (AND rule) and b2 (OR rule). We use arbitrary set thresholds:

- TA = 0.3,
- TB = 0.5.

In Fig. 6 we compare the FR indicators in simulated environment using models: b1 (AND rule) and b2 (OR rule). Analogically, we use arbitrary set thresholds presented above.

The last simulation was performed using model c with three arbitrary set experts acceptance thresholds:

- TE = 0.3,
- TE = 0.5,
- TE = 0.7.

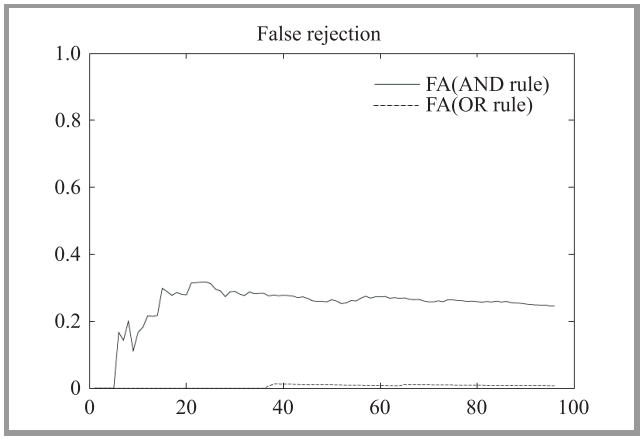



***Fig. 6.*** Simulation results using models: b1 and b2 (FR indicator).

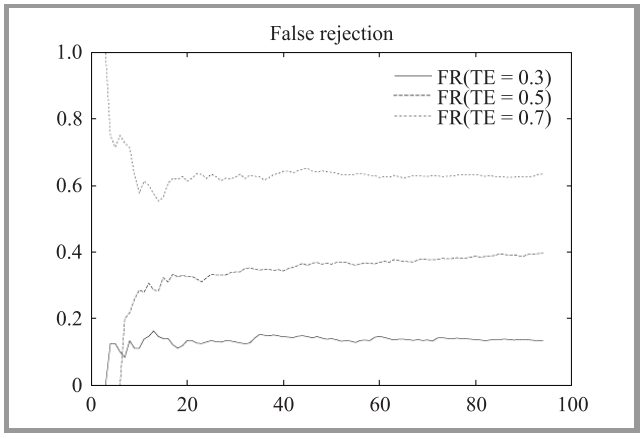



***Fig. 7.*** Simulation results using model c (FR indicator).

The results of last simulation are presented in Fig. 7.

## 6. Conclusions and Further Work

In this paper authors applied multiagent paradigm in order to model single modal and multimodal biometric authentication systems. Four models were implemented using programmable modeling environment for simulating natural and social phenomena. Those models were appropriately parametrized and explored under various conditions. The implemented models enabled observing living environment with agents playing different roles (authenticator, authenticatee and other). The key benefit of proposed approach is the ability of observe how setting different input parameters influences the whole interactive system, as well as watch key performance indicators, i.e., false acceptance rate and false rejection rate. The results of undertaken (preliminary) research task are promising and convinced authors to formulate further research challenges. One of them is an introduction of several (instead of one) authentication centers and represent them in parallel or serial architecture. The second is related to development of learning authentication center exploiting individual instead of global thresholds. The third challenge will be associated with provision of detail parameters of selected biometric method as well as real biometric data.

## References

[1] S. Sanchez and T. Lucas, "Exploring the world of agent-based simulations: simple models, complex analyses", in *Proc. 2002 Winter Simulation Conf.*, San Diego, USA, 2002, pp. 116–126.

[2] H. V. D. Parunak, "Practical and industrial applications of agent-based systems", Environmental Research Institute of Michigan, 1998.

[3] M. Wooldridge and N. R. Jennings, "Intelligent agents: theory and practice", *The Knowl. Engin. Rev.*, vol. 10, no. 2, pp. 115–152, 1995.

[4] S. Franklin and A. Graesser, "Is it an agent or just a program?: a taxonomy for autonomous agents", *Intelligent Agents III*. Berlin: Springer, 1997, pp. 21–36.

[5] S. Russell and P. Norvig, *Artificial Intelligence: Modern Approach*. Prentice Hall, 2002.

[6] M. Wooldridge, *An Introduction to MultiAgent Systems*. Chichester: Wiley, 2002.

[7] C. Macal and M. North, "Tutorial on agent-based modeling and simulation. Part 2: How to model with agents", in *Proc. Winter Simul. Conf. WSC 2006*, Monterey, USA, 2006, pp. 73–83, 2006.

[8] A. Rao and M. Georgeff, "BDI agents: from theory to practice", in *Proc. First Int. Conf. Multi-Agent Sys.*, San Francisco, USA, 1995, MIT Press, pp. 312–319.

[9] K. P. Sycara, "Multiagent Systems", *AI Mag.*, vol. 19, no. 2, Intelligent Agents Summer, pp. 79–93, 1998.

[10] C. Macal and M. North, *Managing Business Complexity. Discovering Strategic Solutions with Agent-Based Modeling and Simulation*. New York: Oxford University Press, 2007.

[11] M. Abreu and M. Fairhurst, "Analyzing the benefits of a novel multiagent approach in a multimodal biometrics identification task", *IEEE Sys. J.*, vol. 3, no. 4, pp. 410–417, 2009.

[12] R. Meshulam, S. Reches, A. Yarden, and S. Kraus, "MLBPR: MAS for large-scale biometric pattern recognition", *Lect. Notes Comp. Sci.* (including Lec. Notes Artif. Int., Lec. Notes Bioinf.), pp. 274–292, 2009.

[13] G. Ali, N. A. Shaikh, and Z. A. Shaikh, "Towards an automated multiagent system to monitor user activities against insider threat", *in Proc. IEEE – Int. Symp. Biometr. Secur. Technol. ISBAST'08*, Islamabad, Pakistan, 2008.

[14] G. Chetty and D. Sharma, "Distributed face recognition: A multiagent approach", Lect. Notes. Comp. Sci. (including Lect. Notes Artif. Int., Lect. Notes Bioinf.), pp. 1168–1175, 2006.

[15] A. Canuto, M. Abreu, A. Medeiros, F. Souza, M. F. Gomes, and V. Bezerra, "Investigating the use of an agent-based multi-classifier system for classification tasks", in *Proc. 11th Int. Conf. Neural Inform. Proces. ICONIP'04*, Calcuta, India, *Lect. Notes Comp. Sci.*, Heidelberg: Springer, 2004, pp. 854–859.

[16] L. Huette, A. Nosary, and T. Paquet, "A multiple agent architecture for handwritten text recognition", *Patt. Recogn.*, vol. 37, no. 4, pp. 665–674, 2004.

[17] U. Wilensky, *NetLogo*. Center for Connected Learning and Computer-Based Modeling, Northwestern University, Evanston, IL, 1999 [Online]. Available: http://ccl.northwestern.edu/netlogo/



**Tomasz Owczarek** received the M.Sc. in management from the Silesian University of Technology in 2005, and the M.Sc. in mathematics from the University of Silesia in 2006. Currently he is a Ph.D. student in the Institute of Economics and Informatics at the Silesian University of Technology. His research interests are game theory, probabilistic graphical models and agent-based modeling and simulation.
e-mail: tomasz.owczarek@polsl.pl
Institute of Economics and Informatics
Silesian University of Technology
Roosevelta st 26-28
41-800 Zabrze, Poland



**Adrian Kapczyński** received the Ph.D. degree in computer science with honors from Silesian University of Technology in 2004. He is experienced in developing and customizing biometric solutions and Well-versed in areas such as database and network designing and administration. A member of IEEE, PIPS, ISACA, ACM and Mensa Polska.
e-mail: adriank@polsl.pl
Institute of Economics and Informatics
Silesian University of Technology
Roosevelta st 26-28
41-800 Zabrze, Poland

Paper

# Probabilistic Issues in Biometric Template Design

Andrzej Pacut

**Abstract—Since the notion of biometric template is not well defined, various concepts are used in biometrics practice. In this paper we present a systematic view on a family of template concepts based on the $L_1$ or $L_2$ dissimilarities. In particular, for sample vectors of independent components we find out how likely it is for the median code to be a sample vector.**

***Keywords—biometrics, sample median, template.***

## 1. Introduction

*Biometric template* is commonly understood as a certain *best representative* of a set of *enrolment data*. This description does not actually makes a definition, since the meaning of 'representative' is only intuitive and the meaning of 'best' is also not defined. In fact, various understanding of those terms lead researchers to quite different transformations of the enrolment data into the template.
In this paper we will sort out several meanings of the term "the best representative" and discuss the resulting methods of template construction.

## 2. Enrolment Measurements as the Sample

We assume that the biometric *enrolment sample* $\mathbf{X}$ for a given subject is a sample of size $n$ in $\mathbb{R}^\ell$, i.e., it consist of a finite sequence of biometric measurements

$$\mathbf{X} = (\mathbf{x}^{(1)}, \ldots, \mathbf{x}^{(n)}), \tag{1}$$

where each measurement $\mathbf{x}^{(i)}$ is represented by a *vector*

$$\mathbf{x}^{(i)} = \begin{bmatrix} x_1^{(i)} \\ \vdots \\ x_\ell^{(i)} \end{bmatrix} \in \mathbb{R}^\ell. \tag{2}$$

To differentiate between the sample elements (the vectors) and vectors elements we call the latter the vector components. In the probabilistic context, we always assume that the sample vectors are independent and have identical distribution (the i.i.d. sample). It is often useful to reinterpret the measurements as finite *sequences* of real numbers $\mathbf{x}^{(i)} = \left(x_1^{(i)}, \ldots, x_\ell^i\right)$; we will use either interpretation. Note that we use upper indexes in parentheses to number the measurements, and reserve lower indexes for their components.

If $\ell = 1$, the sample is called scalar. In particular, one may consider scalar *component samples* that consist of selected components of all sample vectors, namely

$$X_j = (x_j^{(1)}, \ldots, x_j^{(n)}), \qquad j = 1, \ldots, \ell. \tag{3}$$

Elements of the scalar sample can be rearranged in a non-decreasing order

$$x^{[1]} \le x^{[2]} \le \cdots \le x^{[n]}, \tag{4}$$

so that $x^{[1]}$ is the smallest sample element, $x^{[r]}$ is the $r$th smallest, so that $x^{[n]}$ is the largest. The sample can thus be represented by the *ordered sample*

$$(x^{[1]}, \ldots, x^{[n]}) \tag{5}$$

if the original order of sample elements is irrelevant. Note that the ordered representation (5) is in general non unique due to possible repetitions in the sample. This happens in particular if a scalar sample is generated by a discrete random variable whose *support set* is finite $Y = \{y^{(j)}, j = 1, \ldots, M\}$, $y^{(1)} < \ldots < y^{(M)}$. The sample can be then characterized by the support values together with their multiplicities $m^j$, namely, by *the set*

$$\left\{(y^{(j)}, m^{(j)}),\ y^{(j)} \in Y,\ j = 1, \ldots, M\right\}. \tag{6}$$

Certainly $\sum_{j=1}^{M} m^{(j)} = n$. In particular, we will be interested in the *binary case* with $Y = \{0, 1\}$.
The enrolment sets used in biometrics can have more complex structure. For instance, the measurements can be of varying lengths, like in signature biometrics. In those cases, the concepts discussed in this paper must be appropriately modified.

## 3. Template Concepts

We now consider several concepts of the template for the enrolment sample Eq. (1). All concepts employ the notion of dissimilarity $\overline{D}_p$ between the sample and a vector, understood here as the average $p$th power of the $L_p$ distance of the vector to all the enrollment vectors

$$\begin{aligned} \overline{D}_p(\mathbf{x}, \mathbf{X}) &= \frac{1}{n} \sum_{\mathbf{x}^{(i)} \in \mathbf{X}} d(\mathbf{x}, \mathbf{x}^{(i)})^p = \frac{1}{n} \sum_{\mathbf{x}^{(i)} \in \mathbf{X}} \sum_{j=1}^{\ell} \left| x_j - x_j^{(i)} \right|^p \\ &= \frac{1}{n} \sum_{j=1}^{\ell} \sum_{x_j^{(i)} \in X_j} \left| x_j - x_j^{(i)} \right|^p, \end{aligned} \tag{7}$$

where $X_j$ is $j$th component sample. In particular, $\overline{D}_p$ comes down to the average distance ($p=1$) or the average squared distance ($p=2$) between a vector and the enrollment vectors. We often skip the index $p$ when it is obvious from the context.

Within this approach, we distinguish four elementary concepts, each in either $L_1$ or $L_2$ versions, thus making together eight interpretations of the template. The templates obtained by a search for "the best" sample vector will be referred to as template-S (for **s**ample), and those obtained by looking for "the best" vector, not necessarily being a sample vector, will be called template-R (for **r**eal). The result of the approach that combines the two will be called template-RS. Finally, the template intended to minimize the average dissimilarity between the sample vectors and an unknown testing vector will be called template-T (for **t**esting). Certainly, these elementary concepts are far from being exhaustive, and many other, more sophisticated template concepts can be introduced.

The solutions to the underlying minimizations problems we discuss are typically not unique, and by $\mathrm{Arg}\min_{\mathbf{x}\in Z}(\overline{D}_p)$ we denote *the set* of vectors that minimize $\overline{D}_p$ over $x\in Z$.

In the first concept of the template, one of the enrolment vectors is chosen to represent the sample.

**Definition of template-S**. *The template is equal to any enrolment vector that minimizes the average dissimilarity Eq.* (7) *between this vector and the enrolment vectors of the same subject, namely*

$$\mathbf{x}^{*S}\in\mathbf{X}^{*S}\stackrel{\mathrm{def}}{=}\mathrm{Arg}\min_{\mathbf{x}^{(k)}\in\mathbf{X}}\overline{D}_p(\mathbf{x}^{(k)},\mathbf{X})\,. \tag{8}$$

Note that template-S is in general not defined uniquely; it is even possible that all the enrolment vectors fulfil the definition condition. Certainly, all $\mathbf{x}^{*S}\in\mathbf{X}^{*S}$ lead to the same minimal average dissimilarity

$$\overline{D}_p^{*S}\stackrel{\mathrm{def}}{=}\overline{D}_p(\mathbf{x}^{*S},\mathbf{X})\quad\text{for all}\quad\mathbf{x}^{*S}\in\mathbf{X}^{*S}. \tag{9}$$

Template-S definition restricts the search to the enrolment vectors. In the next definition the search is extended to the entire $\mathbb{R}^\ell$.

**Definition of template-R**. *The template is equal to any vector that minimizes the average dissimilarity Eq.* (7) *between this vector and all the enrolment vectors of the same subject, namely*

$$\mathbf{x}^{*R}\in\mathbf{X}^{*R}\stackrel{\mathrm{def}}{=}\mathrm{Arg}\min_{\mathbf{x}\in\mathbb{R}^\ell}\overline{D}_p(\mathbf{x},\mathbf{X})\,. \tag{10}$$

The template here *may not belong* to the enrolment sample. Again, the definition does not in general lead to a unique solution. Unlike Eq. (8), definition Eq. (10) can be substantially simplified: by Eq. (7), the minimization in Eq. (10) can be performed separately for the sample components, namely

$$\min_{\mathbf{x}\in\mathbb{R}^\ell}\overline{D}_p(\mathbf{x},\mathbf{X})=\frac{1}{n}\sum_{j=1}^{\ell}\min_{x_j\in\mathbb{R}}\sum_{x_j^{(i)}\in X_j}|x_j-x_j^{(i)}|^p\,. \tag{11}$$

In the result, the definition of template-R, can be expressed in an equivalent form:

*The template is equal to any vector whose components minimize the average dissimilarities between these components and the corresponding components of the enrolment vectors, namely*

$$x_j^{*R}\in X_j^{*R}\stackrel{\mathrm{def}}{=}\mathrm{Arg}\min_{x\in\mathbb{R}}\sum_{x_j^{(i)}\in X_j}|x-x_j^{(i)}|^p=\mathrm{Arg}\min_{x\in\mathbb{R}}\overline{D}_p(x,X_j),\qquad j=1,\ldots,\ell\,. \tag{12}$$

The minimization of template-R in $\mathbb{R}^\ell$ has been in the above formulation replaced by a series of minimizations in $\mathbb{R}$, which may computationally be much simpler. The minimal dissimilarity is the sum of the component dissimilarities, namely

$$\overline{D}_p^{*R}\stackrel{\mathrm{def}}{=}\overline{D}_p(\mathbf{x}^{*R},\mathbf{X})=\sum_{j=1}^{\ell}\overline{D}_p(x_j^{*R},X_j) \tag{13}$$

and certainly

$$\overline{D}_p^{*R}\le\overline{D}_p^{*S}\,, \tag{14}$$

hence template-R is "better" than template-S. Note that the minimization in Eq. (8) cannot be decomposed into component sample minimizations due to a dependence between the components of $\mathbf{x}$ induced by the restriction of $\mathbf{x}$ to enrolment vectors.

Simplification in template-R definition comes for the cost of the template being not an element of the enrollment sample. To overcome this, one may in a sense integrate a simplicity of template-R definition with an intuitive need of the template to be a sample element as realized by template-S. In the next template concept, we will be looking for the sample element closest to the reference vector calculated according to the definition of template-R.

**Definition of template-RS**. *The template is equal to any enrolment vector that minimizes the distance to template-R, namely*

$$\mathbf{x}^{*RS}\in\mathbf{X}^{*RS}=\arg\min_{\mathbf{x}^{(i)}\in\mathbf{X}}d(\mathbf{x}^{(i)},\mathbf{x}^{*R})\,. \tag{15}$$

Note that the dissimilarity comes down here to $p$th power of the distance in $L_p$, thus the minimization just calls for a minimization of the distance. The value of the $\overline{D}_p$ for template-RS

$$\overline{D}_p^{*RS}=\overline{D}_p(\mathbf{x}^{*RS},\mathbf{X}) \tag{16}$$

certainly fulfils the inequality

$$\overline{D}_p^{*R}\le\overline{D}_p^{*S}\le\overline{D}_p^{*RS}\,, \tag{17}$$

so it is the worst, *in the sense of the average dissimilarity*, of the three templates considered so far.

All the concepts outlined above define the template as a certain representation of the sample, with formally sound but arbitrary meanings of 'representation'. This raises a question whether 'representation' could not be defined less arbitrarily. We thus propose a template concept based on the

template use, introducing an unknown test vector $\mathbf{x^0}$ the template will be compared to. Let $\mathbf{g}$ be a (Borel) function that maps a sample (of a fixed size) in $\mathbb{R}^\ell$ into a vector in $\mathbb{R}^\ell$, i.e., $\mathbf{g}(\mathbf{X}) = \mathbf{g}(\mathbf{x}^{(1)},\ldots,\mathbf{x}^{(n)}) \in \mathbb{R}^\ell$, and denote by $\mathcal{G}$ the family of all such functions.

**Definition of template-T**. *The template vector is equal to the value of any vector-valued (Borel) function of the enrolment data that minimizes the conditional expected distance to the (unknown) test vector given the enrollment data, namely*

$$\mathbf{x}^{*T} = \mathbf{g}^*(\mathbf{x}^{(1)},\ldots,\mathbf{x}^{(n)}), \quad \text{where}$$
$$\mathbf{g}^* \in \operatorname{Arg}\min_{\mathbf{g}\in\mathcal{G}} \mathcal{E}_{\mathbf{X}} d\big(\mathbf{x^0}, \mathbf{g}(\mathbf{X})\big), \tag{18}$$

*where $\mathbf{x^0}$ is a test vector, and $\mathcal{E}_{\mathbf{X}}$ denotes the conditional expectation given $\mathbf{X}$.*

Here the minimization is performed over all (Borel) vector *functions* of the template. The result is obviously not necessarily one of the template elements. The minimization in Eq. (18) can be performed separately for each component of the vector function $\mathbf{g}$, similarly to what we did for template-R. Consequently, the definition of template-T, can be replaced by the following equivalent concept:
*The template is equal to any (Borel) vector-valued function of the enrolment data whose each component minimizes the expected distance to the corresponding component of the test vector, namely*

$$\mathbf{x}^{*T} = \begin{bmatrix} g_1^*(\mathbf{X}) \\ \vdots \\ g_\ell^*(\mathbf{X}) \end{bmatrix}, \quad \text{where}$$
$$g_j^* \in \mathcal{G}_j^* = \operatorname{Arg}\min_{g_j\in\mathcal{G}} \mathcal{E}_{\mathbf{X}} \big|x_j^0 - g_j(\mathbf{X})\big|^p, \tag{19}$$

*where $\mathcal{G}$ is a family of (Borel) functions that map a scalar sample into a scalar.*

Within Bayesian context, we assume here that the test vector $\mathbf{x^0}$, and the template vectors $\mathbf{x}^{(1)},\ldots,\mathbf{x}^{(n)}$ are independent, identically distributed, parameterized by the same unknown parameter vector $\boldsymbol{\vartheta}$, and moreover, that $\mathbf{x^0}$ and $\mathbf{X}$ are conditionally independent given $\boldsymbol{\vartheta}$. We may further rewrite $\mathcal{G}_j^*$ of Eq. (19) in the form

$$\begin{aligned}\mathcal{G}_j^* &= \operatorname{Arg}\min_{g_j\in\mathcal{G}} \mathcal{E}_{\mathbf{X}} |x_j^0(\boldsymbol{\vartheta}) - g_j(\mathbf{x}(\boldsymbol{\vartheta}))|^p \\ &= \operatorname{Arg}\min_{g_j\in\mathcal{G}} \mathcal{E}_{\mathbf{X}} \mathcal{E}_{\mathbf{X},\boldsymbol{\vartheta}} |x_j^0(\boldsymbol{\vartheta}) - g_j(\mathbf{x})|^p, \end{aligned} \tag{20}$$

so for each sample $\mathbf{X}$ and each (unknown) parameter vector $\boldsymbol{\vartheta}$

$$\mathcal{G}_j^* = \operatorname{Arg}\min_{g_j\in\mathcal{G}} \mathcal{E}_{\mathbf{X},\boldsymbol{\vartheta}} |x_j^0(\boldsymbol{\vartheta}) - g_j(\mathbf{x})|^p. \tag{21}$$

In what follows we analyze some properties of the four above template concepts in $L_1$ and $L_2$ spaces.

# 4. $L_1$ Version of Template-R

As we earlier noticed, calculation of template-R, comes down to a series of minimizations for scalar samples. We will thus remind a classical issue of finding a real number $x^*$ closest on the average to all scalar sample elements, i.e., the one that minimizes the average dissimilarity Eq. (7) specified to $L_1$, namely the average absolute distance

$$\overline{D}_1(x^*;X) \stackrel{\text{def}}{=} \frac{1}{n}\sum_{i=1}^{n} |x^{(i)} - x^*|. \tag{22}$$

Let us first notice that *for scalar samples*, the minimization of $\overline{D}_1$ over $\mathbb{R}$ leads to one of the sample elements, e.a. the minimizations over $\mathbb{R}$ and over $X$ lead to the same result.

**Proposition 1.** *For one-dimensional samples, minimizations of the average $L_1$ distance to the sample elements over all real values (real domain), and over all sample values (sample domain) lead to the identical minimum*

$$\min_{x\in\mathbb{R}} \overline{D}_1(x,X) = \min_{x\in X} \overline{D}_1(x,X) \tag{23}$$

*and the solution set in the sample domain is a subset of the one for the real domain*

$$X_1^{*S} \subseteq X_1^{*R}. \tag{24}$$

**Proof**. The function to be minimized is piecewise linear and bounded from below. Hence the minimum always exists, and can be assumed either at a non-differentiability point, namely one of the sample points $x^{(1)},\ldots,x^{(n)}$, or at the points of the closed segment between two neighboring non-differentiability points (Fig. 1). In the finite support case, the non-differentiability points are just the supporting points $y^{(1)},\ldots,y^{(M)}$ (of non-zero multiplicities).

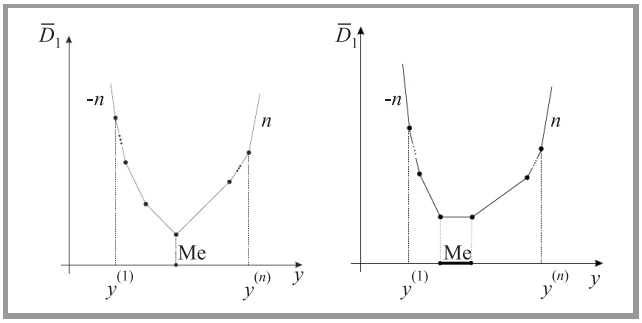



***Fig. 1.*** Two cases of the minimum location of $\overline{D}_1$.

Consequently, for scalar samples $X \subset \mathbb{R}$, the points that minimize $\overline{D}_1$ in the sample domain, minimize also $\overline{D}_1$ in the real domain. In other words, in $L_1$, template-S fulfills the requirements of the definitions of template-R and template-RS.
Minimization of Eq. (22) has a well known solution, which for scalar samples is related to the *sample median*. To formulate it more precisely, we first recall the basic properties of the sample median.

### 4.1. Sample Median

For scalar samples, the sample median $\mathrm{me}(X)$ is understood as any number that "bisects the ordered sample". More precisely, it has the property

$$\begin{aligned} \mathrm{size}\{i : x^{(i)} \le \mathrm{me}(X)\} &\ge n/2\,, \\ \mathrm{size}\{i : x^{(i)} \ge \mathrm{me}(X)\} &\ge n/2\,. \end{aligned} \tag{25}$$

The set of all values that fulfills Ineq. (25) will be called the *median set* $\mathrm{Me}(x)$. If the sample size $n$ is odd then simply

$$\mathrm{me}(X) = x_{[(n+1)/2]}.$$

Note that while in this case the sample median *is defined uniquely*, this value may be taken by more than one sample element: readily, when there are repetitions at the median value, more than one sample element may take the identical value equal to the median; we will call them the *median elements*. This is why the requirements of Ineq. (25) must also allow for sizes greater than the half of the sample size.

If the sample size $n$ is even, any number in the median set

$$\mathrm{Me}(X) = \left[x^{[n/2]}, x^{[n/2+1]}\right], \tag{26}$$

called here the *median interval*, fulfills the requirements of Ineq. (25) hence the definition in this case *may not be unique*. Apart from the values *inside* the median interval, which are not sample values, Ineq. (25) is fulfilled also by the two end points of this interval, which *are* the sample values. If there are repetitions, $x^{[n/2]}$ can be equal to $x^{[n/2+1]}$ so then the median interval shrinks to a single value, and the median is again defined uniquely as the single element of the median interval. Note that both in odd and even sample sizes, *more than one sample element* can be equal to the median.

To make the definition of median unique for any sample size, one often chooses the middle of the median interval as the median in the even sample size case, so then $\mathrm{me}(X) \stackrel{\mathrm{def}}{=} \left(x^{[n/2]} + x^{[n/2+1]}\right)/2$. We are interested in the median as a – non necessarily unique – solution to a minimization problem, so we remain with the definition Eq. (26) for even sample sizes, and often deal with median intervals $\mathrm{Me}(X)$ rather then median values.

Summing up, the median, as an element of the median set, can be equal to one or more sample elements, or be equal to the values which are not sample elements at all (for even sizes).

In the special case of odd-sized binary samples

$$\mathrm{me}(X) = \begin{cases} 0 \text{ if } m^{(0)} > m^{(1)} \\ 1 \text{ if } m^{(0)} < m^{(1)} \end{cases} = \mathbf{1}(m^{(1)} - m^{(0)})\,, \tag{27}$$

where $\mathbf{1}$ denotes the step function, so the sample median is equal to the sample majority value.

In what follows, we will always focus on odd-value samples. It yet straightforward to include also the even-size samples.

### 4.2. $L_1$ Minimization

We are now prepared to minimize Eq. (22) for scalar samples using an elementary reasoning. First we characterize the function to be minimized. To avoid repetitions, sample points $x^{(i)}$ will be represented here by sample suport points $y^{(r)}$.

**Proposition 2** (Average distance). *The following recursive formula applies in the finite support case*

$$\begin{aligned} \delta^{(r)} &= \delta^{(r-1)} + 2m^{(r-1)} \\ \overline{D}_1(y^{(r)}, X) &= \overline{D}_1(y^{(r-1)}, X) + |y^{(r)} - y^{(r-1)}|\,\delta^{(r)}\,, \end{aligned} \tag{28}$$

with

$$\delta^{(0)} = -n\,.$$

The proof of Eq. (28) is immediate and results directly from Proposition 1.

Consequently, the value

$$\min_{x\in\mathbb{R}} \overline{D}_1(x; X) \tag{29}$$

is for odd $n$ uniquely attained by the sample median, and for even $n$ is attained by any point of the median interval.

As seen from Proposition 2, for a given scalar sample, $\overline{D}_1$ is segmentwise linear, with the slopes increasing from some initial negative slope as $x$ increases. Moreover, if $n$ is odd then the function decreases to the left of $y^{((r-1)/2)}$ and increases to right of $y^{((r-1)/2)}$, where

$$\begin{aligned} \delta^{(r)} &< 0\,, \\ \delta^{(r+1)} &> 0\,. \end{aligned} \tag{30}$$

Consequently, the function attains its minimum at $y^{((r-1)/2)}$, which is the sample median. Similarly, if $n$ is even, then $\overline{D}_1$ decreases to the left of $y^{(r/2)}$ and increases to right of $y^{(r/2+1)}$, hence it attains its minimum at all points of the segment $[y^{(r/2)}, y^{(r/2+1)}]$ which is identical to the median set.

### 4.3. Template-R: The Explicit Formula

The above discussion enables to find the vector that fulfills the definition of the template-R in $\mathrm{L}_1$:

**Proposition 3** ( $\mathrm{L}_1$ minimization in $\mathbb{R}^n$). *The template-R is for odd n uniquely given by the vector of sample medians of the component samples*

$$\mathbf{x}^{*R} = \arg\min_{\mathbf{x}\in\mathbb{R}^\ell} \overline{D}_1(\mathbf{x}, \mathbf{X}) = \mathbf{me}(\mathbf{X}) = \begin{bmatrix} \mathrm{me}(X_1) \\ \vdots \\ \mathrm{me}(X_\ell) \end{bmatrix}. \tag{31}$$

*For even n, the solution is not unique and is attained by any vector whose components belong to the median intervals of the corresponding component samples.*

The solution Eq. (31) will be in short called the *median vector* or the *median sequence*, depending on the interpretation. Note that for binary vector samples, the median vector is identical to the *majority code*. The median vector (which fulfills the definition of template-R) *is not* in general a sample vector. In fact, it is easy to see that *for two-dimensional binary samples*, the median vector is always equal to some sample element and thus template-R and template-S are identical. However, there exist 3-dimensional binary samples for which the median vector *is not* an element of the sample. For instance, take $\mathbf{x}^{(1)} = \begin{bmatrix}1\\1\\0\end{bmatrix}$, $\mathbf{x}^{(2)} = \begin{bmatrix}1\\0\\1\end{bmatrix}$ and $\mathbf{x}^{(3)} = \begin{bmatrix}0\\1\\1\end{bmatrix}$. Then $\mathbf{me}(\mathbf{X}) = \begin{bmatrix}1\\1\\1\end{bmatrix}$ is not equal to any sample vector (Fig. 2).

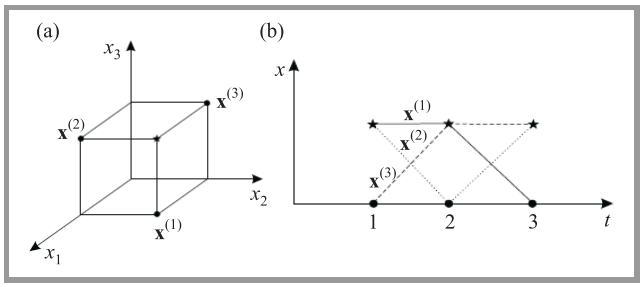



***Fig. 2.*** Example of a 3-dimensional binary median sample for which the median vector is not a sample element. In the vector interpretation (a) the sample elements are marked with filled circles, and the median vector is marked with a star. Similarly, in the sequence interpretation of the vector sample (b) the sample elements are marked with filled circles and joined by different line types; the median sequence points are marked with stars.

# 5. $L_1$ Version of Template-S

In the definition of template-S, the minimization of $\overline{D}_1(\mathbf{X},\mathbf{x})$ is performed over the sample elements, instead of the entire $\mathbb{R}^\ell$. Consequently, this minimization *cannot be decomposed* into independent minimizations in $\mathbb{R}$, hence $\overline{D}_1$ is minimized by a different vector than the one solving (29), and the resulting minimal $L_1$ average distance $\overline{D}_1^*$ is certainly worse.

We will now analyze what is the chance that template-S is identical to template-R for finite support samples. We additionally assume that the finite support sample vectors have *independent components* (not necessarily binary), and derive the probability that the median vector *is* a sample vector. In this order, we first derive the median distribution for finite support samples, and then the distribution of the number of median vectors in the sample.

### *5.1. Sample Median Distribution*

Consider a discrete scalar random variable $\xi$ whose distribution has a finite support $Y = \{y^{(1)},\dots,y^{(M)}\}$. By $P$, $F$, and $S$, we denote its probability function, distribution function, and survival function, respectively, namely

$$\begin{aligned} P(y) &\stackrel{\text{def}}{=} \mathcal{P}(\xi = y)\,,\\ F(y) &\stackrel{\text{def}}{=} \mathcal{P}(\xi < y)\,,\\ S(y) &\stackrel{\text{def}}{=} \mathcal{P}(\xi > y)\ , \end{aligned} \tag{32}$$

for $y \in Y$. Note that some authors use $F(y) = \mathcal{P}(\xi \le y)$, $S(y) = \mathcal{P}(\xi \ge y)$ and then the formulas below would look differently.

To find the distribution of the ordered sample values for finite support i.i.d. sample, we use the result of [1]. The $r$th order statistic is equal to $y$ if there are $u = 0,\dots,r-1$ values less than $y$ and $w = 0,\dots,n-r$ values greater than $y$. The remaining $s = n-u-w$ values must be equal to $y$. Consequently, for $n$-element sample $X$, the probability that the $r$th order statistic is equal to some $y \in \{y^{(1)},\dots,y^{(M)}\}$ is given by

$$\mathcal{P}(x^{[r]} = y) = \begin{cases} \sum\limits_{w=0}^{n-r} \binom{n}{w} P(y)^{n-w} S(y)^w\,, \\ \qquad\qquad \text{for } y = y^{(1)} \\ \sum\limits_{u=0}^{r-1} \sum\limits_{w=0}^{n-r} \binom{n}{u} \binom{n-u}{w} F(y)^u P(y)^{n-u-w} S(y)^w\,, \\ \qquad\qquad \text{for } y = y^{(2)},\dots,y^{(M-1)} \\ \sum\limits_{u=0}^{r-1} \binom{n}{u} F(y)^u P(y)^{n-u}\,, \\ \qquad\qquad \text{for } y = y^{(M)}\,. \end{cases} \tag{33}$$

To simplify the notation, we assume from this moment on that *the sample size is odd*. Derivation for even-size samples must take into account the non-uniqueness of the median value, what makes the formulas a little more complex.

We now can easily find the median distribution for odd sample sizes. Setting $r$ in Eq. (33) to $(n+1)/2$, which corresponds to the median, we obtain the distribution $\mu$ of the median

$$\mu(y) = \mathcal{P}(\mathrm{me}(X) = y) = \begin{cases} \sum\limits_{w=0}^{\overline{n}} \binom{n}{w} P(y)^{n-w} S(y)^w\,, \\ \qquad\qquad \text{for } y = y^{(1)} \\ \sum\limits_{u=0}^{\overline{n}} \binom{n}{u} \sum\limits_{w=0}^{\overline{n}} \binom{n-u}{w} \times \\ \quad \times P(y)^{n-u-w} F(y)^u S(y)^w\,, \\ \qquad\qquad \text{for } y = y^{(2)},\dots,y^{(M-1)} \\ \sum\limits_{u=0}^{\overline{n}} \binom{n}{u} P(y)^{n-u} F(y)^u\,, \\ \qquad\qquad \text{for } y = y^{(M)}\,, \end{cases} \tag{34}$$

where

$$\overline{n} \stackrel{\text{def}}{=} \frac{n-1}{2}\,. \tag{35}$$

For example, for the binary case ($Y = \{0,1\}$, $P(1) = p$, $P(0) = q = 1-p$) we have

$$\mu(y) = \begin{cases} \sum\limits_{w=0}^{\overline{n}} \binom{n}{w} q^{n-w} p^w\,, & \text{for } y = 0 \\ \sum\limits_{u=0}^{\overline{n}} \binom{n}{u} p^{n-u} q^u\,, & \text{for } y = 1\,. \end{cases} \tag{36}$$

While for the i.i.d. samples, the sample median converges with the sample size to the population median if such is uniquely defined, yet for samples sizes typically considered in biometrics, the two quantities may strongly differ (see Fig. 3).

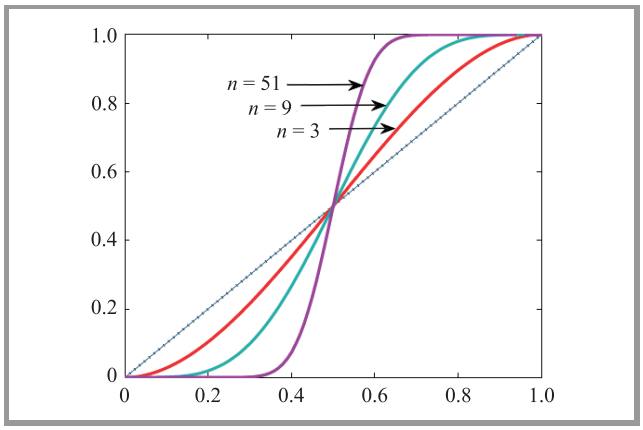



***Fig. 3.*** Probability that the sample median is equal to one versus the probability of success $p$ for the scalar binary i.i.d. sample, for several values of the sample size $n$. The population median is equal to 1 for $p > 0.5$.

### *5.2. Number of Median Elements*

Note that in the scalar sample there is at least one element equal to the median, called here further the *median element*. Typically, for scalar finite-support samples, there are even more than one median element. We will derive the distribution of the number of median elements in a scalar finite support sample, irrespectively of the median value. Denote by $\mathcal{M}$ the number of median elements in $X$ and by $\nu$ its distribution function, i.e.,

$$\nu(z) \stackrel{\text{def}}{=} \mathcal{P}\{\mathcal{M} = z\}, \quad z = 0,\ldots,n. \tag{37}$$

**Proposition 4** (Distribution of the number of median elements in scalar finite-support samples). *The distribution of the number of median elements is given by*

$$\nu(s) = \begin{cases} 0, & \text{for } s = 0 \\ \binom{n}{s} \sum_{m=2}^{M-1} P(y^{(m)})^s \sum_{u=\overline{n}+1-s}^{\overline{n}} \binom{n-s}{u} F(y^{(m)})^u S(y^{(m)})^{n-s-u}, & \text{for } s = 1,\ldots,\overline{n} \\ \binom{n}{s} \sum_{n=1}^{M} P(y^{(m)})^s (1-P(y^{(m)}))^{n-s}, & \text{for } s = \overline{n}+1,\ldots,n \end{cases} \tag{38}$$

**Proof.** The proof is given in Appendix A.

**Corollary 1** (The binary case). For the binary sample we obtain

$$\nu(s) = \begin{cases} 0 & s = 0,\ldots,\overline{n} \\ \binom{n}{s}\left(p^s q^{n-s} + q^s p^{n-s}\right) & s = \overline{n}+1,\ldots,n. \end{cases} \tag{39}$$

As we stressed, for vector samples ($\ell > 1$) the median vector may not be equal to any sample vector. The question arises, how likely it is that the median vector *does* belong to the sample. For i.i.d. finite-support vector samples whose sample vectors *have independent components* we now derive the probability that there exists at least one median vector among $n$ sample vectors. Since we will be dealing here with vector samples in $\mathbb{R}^\ell$ with various $\ell$, we index the samples with their vectors lengths, i.e., $\mathbf{X}_\ell$ denotes a sample of $\ell$-element vectors. Denote by $\mathfrak{M}_\ell$ the number of median vectors in $\mathbf{X}_\ell$ and by $\boldsymbol{\nu}_\ell$ its distribution function, i.e.,

$$\boldsymbol{\nu}_\ell(z) \stackrel{\text{def}}{=} \mathcal{P}\{\mathfrak{M}_\ell = z\}, \quad z = 0,\ldots,n. \tag{40}$$

**Proposition 5** (Distribution of the number of median vectors for finite-support independent component vector samples). *The distribution of the number of median vectors in the sample is for $\ell = 2,3,\ldots$ given recursively by*

$$\begin{aligned} \boldsymbol{\nu}_1(z) &= \nu(z), \quad z = 0,\ldots,n \\ \boldsymbol{\nu}_\ell(z) &= \sum_{z'=z}^{n} \binom{z'}{z} \boldsymbol{\nu}_{\ell-1}(z') \sum_{s=z}^{n-z'+z} \nu(s) \frac{\binom{n-z'}{s-z}}{\binom{n}{s}}, \\ & \qquad z = 0,\ldots,n, \quad \ell = 2,3,\ldots \end{aligned} \tag{41}$$

**Proof**. The proof is presented in Appendix B.

The main problem to overcome is the dependence of random variables $\mathfrak{M}_\ell$ both for different $n$ and for different $\ell$. The former results from a possibility of changing the sample median by any sample vector, and the latter is caused by the dependence between the median vector components.

**Corollary 2** (Binary case). For a binary sample with $Y = \{0,1\}$, $P(0) = q$, $P(1) = p$, Eqs. (41) simplifies to

$$\begin{aligned} \boldsymbol{\nu}_\ell(z) = &\sum_{z'=z}^{\min(n,z+\overline{n})} \binom{z'}{z} \boldsymbol{\nu}_{\ell-1}(z') \\ &\sum_{s=\max(\overline{n}+1,z)}^{n-z'+z} \binom{n-z'}{s-z} \left(p^s q^{n-s} + q^s p^{n-s}\right). \end{aligned} \tag{42}$$

We now can easily calculate the probability of the presence of at least one median vector in the sample, which is equal to $1 - \boldsymbol{\nu}_\ell(0)$.

We illustrate the results for the vector binary case. For binary samples, the number of median vectors strongly depends on $P(1) = p$. Exemplary results are shown in Fig. 4 for a binary sample of a fixed size $n = 15$ and several vector lengths $\ell$. For $0.3 \le p \le 0.7$, the median vectors cease to exist in samples in $\mathbb{R}^\ell$, $\ell > 30$, while for $p = 0.05$ or $p = 0.95$ they still exist with probability $> 0.7$ for $\ell = 50$.

Another view of the same results is shown in Fig. 5, where the probability $1 - \boldsymbol{\nu}_\ell(0)$ that the sample contains any median vectors is plotted versus the vector length $\ell$, for two probabilities of the success: $p = 0.5$ (Fig. 5a) and

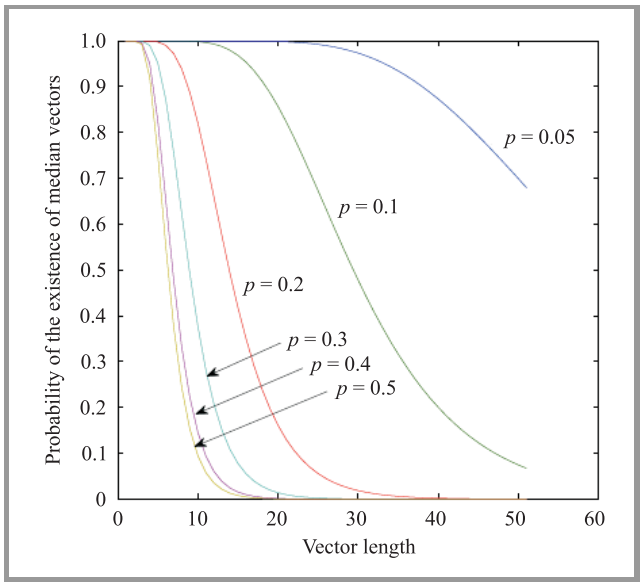



***Fig. 4.*** Probability of the presence of the median vector in a $n = 15$-element binary sample versus the space dimension $\ell$, with the probability of success $p$ as a parameter, $p \in \{0.05, 0.1, 0.2, 0.3, 0.4, 0.5\}$.

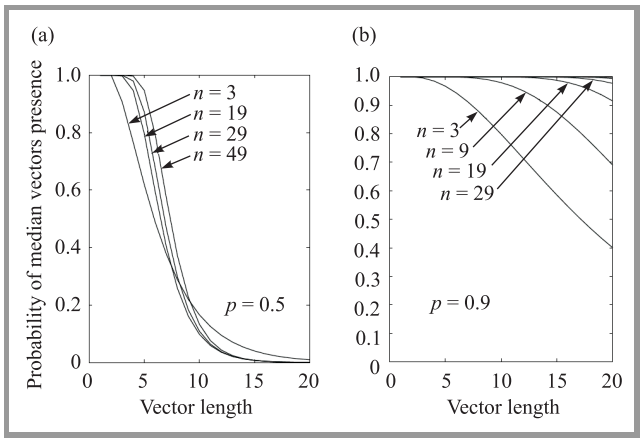



***Fig. 5.*** Probability that there exist median vectors in the binary vector sample versus vector length $\ell$, for the probability of success $p = 0.5$ and sample sizes $n \in \{3, 19, 29, 49\}$ (a), and for $p = 0.9$ and $n \in \{3, 9, 19, 29\}$ (b).

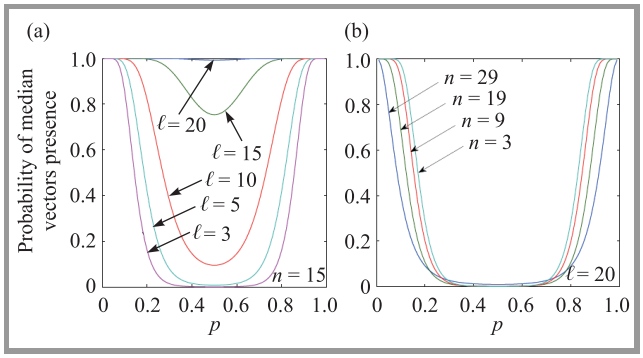



***Fig. 6.*** Probability that there exist median vectors in a binary vector sample versus the probability of success $p$ for the sample size $n = 15$ with vector lengths $\ell \in \{3, 5, 10, 15, 20\}$ (a) and for the vector length $\ell = 20$ and sample sizes $n \in \{3, 9, 19, 29\}$ (b).

$p = 0.9$ (Fig. 5b), each for several sample sizes $n$. The chance that the median vector exists quickly goes to zero with the increase of the vector length, and the velocity of the decrease grows as $p$ get closer to $1/2$.

The last phenomenon is very well visible in yet another visualization of the results (Fig. 6) where the probability $1 - \boldsymbol{\nu}_\ell(0)$ of the median vectors presence in a binary vector sample is plotted versus the probability of success $p$, for a fixed sample size $n = 15$ and several vector lengths $\ell$ (Fig. 6a), and for a fixed vector length $\ell = 20$ and several sample sizes $n$ (Fig. 6b). The probability quickly increases as $|p - 0.5|$ increases, and for each $p$ it increases both with $\ell$ and $n$. Note that the characteristics (Fig. 6b and Fig. 5a) may intersect. This means that the dependence of the probability of $n$ may not be monotonic for lower $n$.

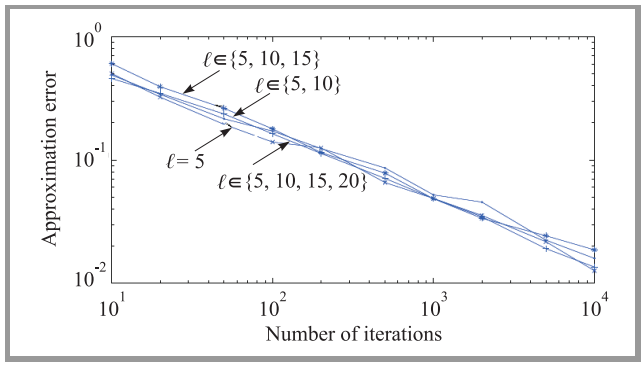



***Fig. 7.*** The absolute difference between the simulated and theoretical distribution values versus number of iterations; logarithmic scales, averaged over $\ell$ and $n$.

An approximation of the discussed distribution can certainly be brought in by direct simulations. In our experiments, it was necessary to use about 10000 repetitions of the entire sample to obtain the simulation error of order of 0.01. The experiments show (Fig. 7) that in logarithmic scales the simulation error decreases almost linearly with the number of sample repetitions, which comes down to an exponential decrease of the simulation error with the number of repetitions. A decrease of the error by an order of one requires the increase in the number of repetitions by order of about one and a half.

## 6. $L_1$ Version of Template-RS

We now consider the definition of template-RS. In $L_1$ one can rewrite the defining formula (15) to the form

$$\mathbf{x}^{*RS} = \arg \min_{\mathbf{x}^{(i)} \in \mathbf{X}} \sum_{j=1}^{\ell} |x_j^{(i)} - \mathrm{me}_j| , \tag{43}$$

where $\mathrm{me}_j = \mathrm{me}(X_j)$ denotes $j$th component of $\mathbf{me}(\mathbf{X})$. We now consider odd-size *binary samples* in $L_1$ and compare

the definitions of template-RS and template-S. The latter is unique and by Eq. (8) equal to

$$\mathbf{x}^{*S} = \arg\min_{\mathbf{x}^{(k)}\in\mathbf{X}} \overline{D}_1(\mathbf{x}^{(k)},\mathbf{X}). \tag{44}$$

Since for binary vectors $\|\mathbf{a}-\mathbf{b}\|_1 = \|\mathbf{a}-\mathbf{b}\|_2^2$ hence $\overline{D}_1(\mathbf{x}^{(k)},\mathbf{X})$ can be rewritten as

$$\begin{aligned}\overline{D}_1(\mathbf{x}^{(k)},\mathbf{X}) &= \frac{1}{n}\sum_{\mathbf{x}^{(i)}\in\mathbf{X}} \left\|\mathbf{x}^{(k)}-\mathbf{x}^{(i)}\right\|_2^2 \\ &= \frac{1}{n}\sum_{\mathbf{x}^{(i)}\in\mathbf{X}} \left\|(\mathbf{x}^{(k)}-\mathbf{me}(\mathbf{X}))-(\mathbf{x}^{(i)}-\mathbf{me}(\mathbf{X}))\right\|_2^2 \\ &= \left\|\mathbf{x}^{(k)}-\mathbf{me}(\mathbf{X})\right\|_1 + \frac{1}{n}\sum_{\mathbf{x}^{(i)}\in\mathbf{X}} \left\|\mathbf{x}^{(i)}-\mathbf{me}(\mathbf{X})\right\|_1 \\ &\quad -\frac{2}{n}\sum_{j=1}^{\ell}(x_j^{(k)}-\mathrm{me}_j)\sum_{\mathbf{x}^{(i)}\in\mathbf{X}}(x_j^{(i)}-\mathrm{me}_j). \end{aligned}\tag{45}$$

Joining the first and the last terms we may write $\overline{D}_1$ in the form

$$\begin{aligned}\overline{D}_1(\mathbf{x}^{(k)},\mathbf{X}) = \sum_{j=1}^{\ell}\left|x_j^{(k)}-\mathrm{me}_j\right|\Big(1-\frac{2}{n}\mathrm{sign}(x_j^{(k)}-\mathrm{me}_j) \\ \sum_{x_j^{(i)}\in X_j}(x_j^{(i)}-\mathrm{me}_j)\Big) + \frac{1}{n}\sum_{\mathbf{x}^{(i)}\in\mathbf{X}}\left\|\mathbf{x}^{(i)}-\mathbf{me}(\mathbf{X})\right\|_1. \end{aligned}\tag{46}$$

Since the last term does not depend on $\mathbf{x}^{(k)}$ we finally may write

$$\mathbf{x}^{*S} = \arg\min_{\mathbf{x}^{(k)}\in\mathbf{X}} \sum_{j=1}^{\ell} w_j^k\left|x_j^{(k)}-\mathrm{me}_j\right|, \tag{47}$$

where

$$w_j^k = 1-\frac{2}{n}\mathrm{sign}(x_j^{(k)}-\mathrm{me}_j)\sum_{x_j^{(i)}\in X_j}(x_j^{(i)}-\mathrm{me}_j). \tag{48}$$

It is easy to show that $\mathrm{sign}(w_j^k)$ is always equal to 1. In fact, since the absolute value of the sum is not greater than the sum of absolute values, and for scalar binary samples there must be more elements equal to the median than those nonequal, we have

$$\begin{aligned}\left|\frac{2}{n}\mathrm{sign}(x_j^{(k)}-\mathrm{me}_j)\sum_{x_j^{(i)}\in X_j}(x_j^{(i)}-\mathrm{me}_j)\right| \\ \le \frac{2}{n}\sum_{x_j^{(i)}\in X_j}|x_j^{(i)}-\mathrm{me}_j| < 1 \end{aligned}\tag{49}$$

hence $\mathrm{sign}(w_j^k)=1$. Considering Eq. (43) and Eq. (47) as linear programming problems with respect to the variables $|x_j^{(k)}-\mathrm{me}_j|$, we see that the solutions of both problems are identical. In the other words, for vector binary samples, the definitions of template-R and template-RS lead to the same template. Note that we did not make any assumptions about independence of the components of sample elements.

# 7. $L_1$ Version of Template-T

In $L_1$, template-T can be by Eqs. (19) and (21) rewritten for each sample $\mathbf{X}$ and each (unknown) distribution parameter vector $\boldsymbol{\vartheta}$ in the form

$$\mathbf{x}^{*T} = \begin{bmatrix} g_1^*(\mathbf{X}) \\ \vdots \\ g_\ell^*(\mathbf{X}) \end{bmatrix}, \quad \text{where}$$

$$g_j^* \in \arg\min_{g_j\in\mathcal{G}} \mathcal{E}_{\mathbf{X},\boldsymbol{\vartheta}}\left|x_j^0(\boldsymbol{\vartheta}) - g_j(\mathbf{X})\right|. \tag{50}$$

The minimum is attained by the (non-random) median of the (conditional) distribution of $x_j^0(\boldsymbol{\vartheta})$, hence $g_j^*(\mathbf{X})$ should approximate this value. We will employ the component sample median $\mathrm{me}_j = \mathrm{me}(X_j)$ to estimate $g_j^*(\mathbf{X})$, and thus take the sample median to estimate $\mathbf{g}^*(\mathbf{X})$, namely

$$\mathbf{x}^{*T} \approx \begin{bmatrix} \mathrm{me}_1 \\ \vdots \\ \mathrm{me}_\ell \end{bmatrix}. \tag{51}$$

Note yet that for dependent components of sample elements, some information about component sample medians is contained also in other component samples, hence the solution (51) is suboptimal. In fact, for independent component vector samples, $\mathbf{x}^{*T}$ obtained here is identical to $\mathbf{x}^{*R}$.

# 8. Template Definitions in $L_2$

### *8.1. Template-R in $L_2$*

As earlier noticed, definition of template-R in $L_p$ comes down to a series of minimization subproblems for scalar samples. However, this feature is not needed to derive $\mathbf{x}^{*R}$ in $L_2$, since here we have just the classical issue of least squares: find a vector $\mathbf{x}^{*R}$ whose average squared distance to all other sample vectors is minimized

$$\mathbf{x}^{*R} = \arg\min_{\mathbf{x}\in\mathbb{R}} \overline{D}_2(\mathbf{x},\mathbf{X}), \tag{52}$$

where

$$\overline{D}_2(\mathbf{x},\mathbf{X}) = \frac{1}{n}\sum_{i=1}^{n}\|\mathbf{x}^{(i)}-\mathbf{x}\|^2. \tag{53}$$

This is solved in a standard way by adding and subtracting the sample average

$$\overline{\mathbf{x}} = \frac{1}{n}\sum_{i=1}^{n}\mathbf{x}^{(i)} \tag{54}$$

to the terms inside the norm, so taking into account that the sum of the product term is equal to zero, we obtain

$$\overline{D}_2(\mathbf{x},\mathbf{X}) = \frac{1}{n}\sum_{i=1}^{n}\|\mathbf{x}^{(i)}-\overline{\mathbf{x}})-(\mathbf{x}-\overline{\mathbf{x}})\|^2$$
$$= \|\mathbf{x}-\overline{\mathbf{x}}\|^2+\frac{1}{n}\sum_{i=1}^{n}\|\mathbf{x}^{(i)}-\overline{\mathbf{x}}\|^2. \quad (55)$$

Readily, for any sample, irrespectively of any assumptions about independence of sample vector components, one obtains

$$\mathbf{x}^{*R}=\overline{\mathbf{x}}. \quad (56)$$

### *8.2. Template-S and Template-RS in $L_2$*

Definition of template-S calls for minimization over the selected points of the vector space, namely

$$\mathbf{x}^{*S}=\arg\min_{\mathbf{x}\in\mathbf{X}}\overline{D}_2(\mathbf{x},\mathbf{X})=\arg\min_{\mathbf{x}\in\mathbf{X}}\|\mathbf{x}-\overline{\mathbf{x}}\|^2. \quad (57)$$

It is certainly unlikely that - even for finite support samples – the sample average is equal to any sample vectors. Note that in $L_2$, template-S Eq. (57) and template-RS: $\mathbf{x}_2^{*RS}\in\operatorname{Arg}\min_{\mathbf{x}\in\mathbf{X}}\|\mathbf{x}-\overline{\mathbf{x}}\|^2$ are equivalent, irrespectively of any independence conditions.

### *8.3. Template-T in $L_2$*

In $L_2$, we rewrite template-T similarly as in $L_1$, namely

$$\mathbf{g}^*=\arg\min_{\mathbf{g}\in\mathcal{G}}\mathcal{E}_{\mathbf{X},\boldsymbol{\vartheta}}\|\mathbf{x}^{\mathbf{0}}(\boldsymbol{\vartheta})-\mathbf{g}(\mathbf{X})\|^2. \quad (58)$$

The minimum is attained by the mean value of the (non-random) mean value of the (conditional) distribution of $\mathbf{x}^{\mathbf{0}}(\boldsymbol{\vartheta})$, so $\mathbf{g}^*(\mathbf{X})$ should approximate this value. Employing the sample average $\overline{\mathbf{x}}$ to estimate the mean value of $\mathbf{x}^{\mathbf{0}}(\boldsymbol{\vartheta})$, we obtain

$$\mathbf{x}^{*T}\approx\overline{\mathbf{x}}. \quad (59)$$

For dependent components of sample vectors, the information about the conditional mean value is contained also in other components, hence the solution can be improved. In other words, the solution (59) is identical to $\mathbf{x}^{*R}$ for independent component samples.

## 9. Conclusions

Our analysis of the art of template creation only touches the problem of choosing "the best representative" of biometric samples. We discussed only the problems characterized by measurements that could be viewed as points of a metric space, and if so, the metric was assumed to be Euclidean. The problem in general touches the notion of information contents of biometric measurement systems. Even if one desires to assume that the biometric measurements lead to Euclidean spaces, there still are various possibilities of choosing the "best representative". Intuitively, such the representative must express some "stable" properties of the measurements for a single subject and as such, it may strongly depend on the biological quantities under scrutiny. Consequently, choosing the template calls for a thorough knowledge of the biological context. On the other hand, apart from this context, one may choose the template on the base of one of "black box" solutions and choose the solution that works best for the given biometric database(s). In the paper we in fact analyzed several "black box" solutions to show their properties and determine their mutual relations.

The concepts we analyzed were based on $L_1$ and $L_2$ distances between the measurements. The possibilities we examined included the template as an enrollment measurement that is on the average closest to all other enrollment measurements (template-S), and a vector (not necessarily any enrollment set vector) closest on the average to all enrollment measurements (template-R). Since the latter is not necessarily the enrollment vector, we may treat it as a reference measurement, and define the template as the enrollment vector closest to the reference (template-RS). Finally, we also introduced the template that aimed into minimization of the distance between the template and a test measurement (template-T). Each of those concepts was analyzed with the use of $L_1$ and $L_2$ distances, so eight versions of the template were investigated.

We investigated closer the $L_1$ concepts, since they are less known. We showed, using independent component binary samples, that template-S differs from template-R, and the difference grows with the dimension of the sample vectors. Also, the difference grows as $p$ approaches 0.5. This suggests that in general, the difference between template-S and template-R is higher for the underlying (population) distributions of higher entropy. We also showed that for binary samples template-RS is identical to template-S.

One may notice that as the enrollment sample size grows, all the concepts considered here may lead to either the subject's theoretical median or the subject's theoretical expected value. The templates based on samples of finite size can be thus treated as various estimators of subject's theoretical characteristics. They may strongly differ from the theoretical characteristics because the enrolment sample size for a single subject can be very low (as low as three measurements). On the other hand, the dimension of the measurements can be very high, since it must have the information contents high enough to differentiate between many subjects of large biometric many-subject databases. Note that if the number of the subjects grow, as in attempts to build universal identity verifiers, the templates considered here must also be based on growing enrollment subject's sample sizes, to be as close as possible to the theoretics subjects' characteristics. The question remains open if these characteristics have sufficient information contents, and even what the underlying theoretical subject and the entire population distributions are. We only hope that the assumption about the very existence of these distributions, or – in other words – on the possibility of describing the biological variability in terms of probabilities, holds.

# Appendix A

## Proof of Proposition 4

To prove Proposition 4 for finite-support scalar samples $X$ we first derive the joint probability function of the sample median $\mathrm{me}(X)$ and the number of median elements $\mathcal{M}$

$$\rho(y,s) \stackrel{\text{def}}{=} \mathcal{P}\{\mathrm{me}(X)=y \wedge \mathcal{M}=s\}, \quad \text{where} \quad y=y^{(1)},\ldots,y^{(M)}, \quad s=0,\ldots,n. \tag{60}$$

Certainly, $\mu(y)=\sum_{s=0}^{n}\rho(y,s)$, for $y=y^{(1)},\ldots,y^{(M)}$ and $\nu(s)=\sum_{m=1}^{M}\rho(y^{(m)},s)$, for $s=0,\ldots,n$. Assuming that the sample size $n$ is odd, we will find the joint probability function $\rho(y,s)$ for $s=0,\ldots,n$, $y=y^{(1)},\ldots,y^{(M)}$.

**Proposition 6** (Joint distribution of the sample median and the number of median elements for scalar samples). *For n-element discrete support scalar sample, the joint probability function of the sample median and the number of median elements is given by*

$$\rho(y,s)=\begin{cases} 0, & \text{for } s=0,\ldots,\overline{n}, \quad y=y^{(1)}\vee y=y^{(M)} \\ \binom{n}{s}P(y)^s\sum\limits_{u=\overline{n}+1-s}^{\overline{n}}\binom{m-s}{u}F(y)^u S(y)^{n-s-u}, & \\ \quad \text{for } s=0,\ldots,\overline{n}, \quad y=y^{(2)},\ldots,y^{(M-1)} & \\ \binom{n}{s}P(y)^s\big(1-P(y)\big)^{n-s}, & \\ \quad \text{for } s=\overline{n}+1,\ldots,n, & \end{cases} \tag{61}$$

where $\overline{n}=(n-1)/2$.

**Proof**. We first rearrange the summation in Eq. (34) to show the influence of the terms related to the number $s=n-u-w$ of the sample elements equal to the median, namely (Fig. 8)

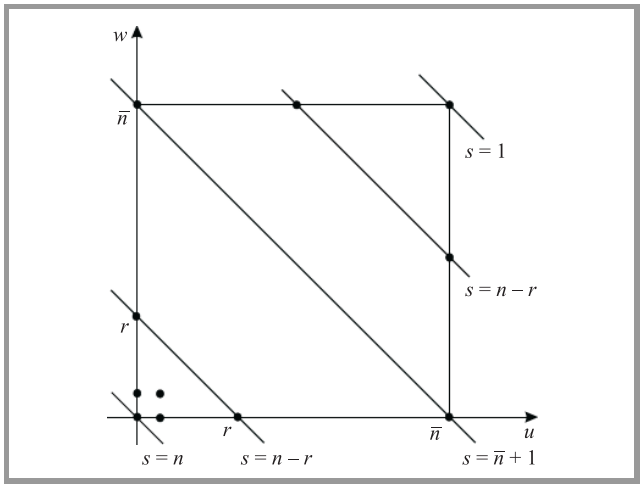



*Fig. 8.* Change of variables in Eq. (62).

$$\mu(y)=\begin{cases} \sum\limits_{s=\overline{n}+1}^{n}\binom{n}{s}P(y)^s S(y)^{n-s}, & \\ \quad \text{for } y=y^{(1)} & \\ \sum\limits_{s=1}^{\overline{n}}\sum\limits_{u=\overline{n}+1-s}^{\overline{n}}\binom{n}{s,u}P(y)^s F(y)^u S(y)^{n-s-u} & \\ +\sum\limits_{s=\overline{n}+1}^{n}\sum\limits_{u=0}^{n-s}\binom{n}{s,u}P(y)^s F(y)^u S(y)^{n-s-u}, & \\ \quad \text{for } y=y^{(2)},\ldots,y^{(M-1)} & \\ \sum\limits_{s=\overline{n}+1}^{n}\binom{n}{s}P(y)^s F(y)^{n-s}, & \\ \quad \text{for } y=y^{(M)}. & \end{cases} \tag{62}$$

We have

$$\rho(y,s)=\binom{n}{s}P(y)^s\begin{cases} S(y)^{n-s}, & \\ \quad \text{for } y=y^{(1)}, \quad s=\overline{n}+1,\ldots,n & \\ \sum\limits_{u=\overline{n}+1-s}^{\overline{n}}\binom{n-s}{u}F(y)^u S(y)^{n-s-u}, & \\ \quad \text{for } y=y^{(2)},\ldots,y^{(M-1)}, & \\ \quad s=1,\ldots,\overline{n} & \\ \sum\limits_{u=0}^{n-s}\binom{n-s}{u}F(y)^u S(y)^{n-s-u}, & \\ \quad \text{for } y=y^{(2)},\ldots,y^{(M-1)}, & \\ \quad s=\overline{n}+1,\ldots,n & \\ F(y)^{n-s}, & \\ \quad \text{for } y=y^{(M)}, \ s=\overline{n}+1,\ldots,n & \\ 0 \quad \text{otherwise}, & \end{cases} \tag{63}$$

hence

$$\rho(y,s)=\binom{n}{s}P(y)^s\begin{cases} \sum\limits_{u=\overline{n}+1-s}^{\overline{n}}\binom{n-s}{u}F(y)^u S(y)^{n-s-u}, & \\ \quad \text{for } y=y^{(2)},\ldots,y^{(M-1)}, & \\ \quad s=1,\ldots,\overline{n} & \\ \big(1-P(y)\big)^{n-s}, & \\ \quad \text{for } y=y^{(1)},\ldots,y^{(M)}, & \\ \quad s=\overline{n}+1,\ldots,n & \\ 0 \quad \text{otherwise}, & \end{cases} \tag{64}$$

so Eq. (61) follows.

**Corollary 3** (The binary case). For a binary sample with $Y=\{0,1\}, P(0)=q$, $P(1)=p$ we have

$$\rho(y,s)=\begin{cases} 0, & s=0,\ldots,\overline{n} \\ \binom{n}{s}q^s p^{n-s}, & y=0, \quad s=\overline{n}+1,\ldots,n \\ \binom{n}{s}p^s q^{n-s}, & y=1, \quad s=\overline{n}+1,\ldots,n. \end{cases} \tag{65}$$

Now, Proposition 4 can be easily proven.

**Proof of Proposition 4**. We can obtain the distribution of the number of median elements by summing up $\rho$ in Eq. (61) over all $y$

$$\nu(s)=\sum_{y\in Y}\rho(y,s). \tag{66}$$

Readily,

$$\nu(s)=\binom{n}{s}\begin{cases}0, & \text{for } s=0\\ \sum\limits_{u=\overline{n}+1-s}^{\overline{n}}\binom{n-s}{u}\sum\limits_{m=2}^{M-1} P(y^{(m)})^s F(y^{(m)})^u S(y^{(m)})^{n-s-u}, & \text{for } s=1,\ldots,\overline{n}\\ \sum\limits_{m=1}^{M} P(y^{(m)})^s (1-P(y^{(m)}))^{n-s}, & \text{for } s=\overline{n}+1,\ldots,n\end{cases} \quad (67)$$

hence we obtain Eq. (38).

# Appendix B

## Proof of Proposition 5

Let $\mathfrak{M}_{1,\ldots,\ell}$ be the number of median vectors in $\ell$-dimensional sample of size $n$, and denote by $\boldsymbol{\nu}_\ell(z)$ its probability function, namely

$$\boldsymbol{\nu}_\ell(z)\stackrel{\text{def}}{=}\mathcal{P}\{\mathfrak{M}_\ell=z\},\quad z=0,\ldots,n \quad (68)$$

for $\ell=1,\ldots..$ We derive the probability function $\nu\nu\nu_\ell$ recursively. Given an $\ell$-dimensional sample $\mathbf{X}_\ell$ we form a $(\ell-1)$-dimensional sample $\mathbf{X}_{\ell-1}{}'$ by removing a single component (say the last one) of each vector in $\mathbf{X}_\ell$. We first calculate the conditional probability that there are exactly $z$ median vectors in $\mathbf{X}_\ell$ given there are exactly $z'$ median vectors in $\mathbf{X}_{\ell-1}{}'$ and $s$ median elements in $X$, namely

$$p_\ell(z|z',s)\stackrel{\text{def}}{=}\mathcal{P}\big(\mathfrak{M}_\ell=z|\mathfrak{M}_{\ell-1}=z'\wedge\mathcal{M}=s\big). \quad (69)$$

It is easy to see that

$$p_\ell(z|z',s)=\begin{cases}\binom{z'}{z}\binom{n-z'}{s-z}/\binom{n}{s}, & \text{for } z\le z' \text{ and } z'-z\le n-s \text{ and } z\le s\\ 0, & \text{for } z>z' \text{ or } z'-z>n-s \text{ or } z>s.\end{cases} \quad (70)$$

Note that $p_\ell(z|z',s)$ is null except for the points in the triangle $z'\ge z,\ s\ge z,\ s+z'\le s+z$ in the $(s,z')$ plane. The distribution $\boldsymbol{\nu}_\ell(z)$ of $\mathfrak{M}_\ell$ can be thus be found by a summation of the conditional distribution Eq. (70) with respect to distributions: $\boldsymbol{\nu}_{\ell-1}$ of $\mathfrak{M}_{\ell-1}$ and $\nu$ of $\mathcal{M}$ of the two independent random variables. Therefore, $\boldsymbol{\nu}_\ell(z)$ can be determined recursively as

$$\begin{aligned}\boldsymbol{\nu}_1(z) &= \nu(z),\quad z=1,\ldots n\\ \boldsymbol{\nu}_\ell(z) &= \sum_{z'=0}^{n}\boldsymbol{\nu}_{\ell-1}(z')\sum_{s=0}^{n}p_\ell(z|z',s)\,\nu(s),\quad z=0,\ldots,n,\\ &\quad \ell=2,3,\ldots\end{aligned} \quad (71)$$

Plugging Eq. (70) into the above we obtain Eq. (41).

For the binary sample with $Y=\{0,1\}$, $P(1)=p$, $P(0)=q$, $\rho(s)$ is given by Eq. (39). Since this is equal to zero for $s\le\overline{n}$, the summation in Eq. (71) narrows down to the triangle $z'\ge z,\ s\ge\overline{n}+1,\ s+z'\le s+z$ for $z\le n$, and the triangle $z'\ge z,\ s\ge z,\ s+z'\le s+z$ otherwise. Consequently, Eq. (71) simplifies to

$$\boldsymbol{\nu}_\ell(z)=\begin{cases}\sum\limits_{z'=z}^{z+\overline{n}}\binom{z'}{z}\boldsymbol{\nu}_{\ell-1}(z')\sum\limits_{s=\overline{n}+1}^{n-z'+z}\binom{n-z'}{s-z}(p^sq^{n-s}+q^sp^{n-s}), & \text{for } z=0,\ldots,\overline{n}\\ \sum\limits_{z'=z}^{n}\binom{z'}{z}\boldsymbol{\nu}_{\ell-1}(z')\sum\limits_{s=z}^{n-z'+z}\binom{n-z'}{s-z}(p^sq^{n-s}+q^sp^{n-s}), & \text{for } z=\overline{n}+1,\ldots,n\end{cases} \quad (72)$$

and Eq. (42) follows.


## Acknowledgements

This paper has been financed by the Ministry of Science and Higher Education grant OR00 0026 07 "A platform of secure biometrics implementations in personal verification and identification".


## Reference


[1] D. L. Evans, L. M. Leemis, and J. H. Drew, "The distribution of order statistics for discrete random variables with applications to bootstrapping", *Informs J. Comput.*, vol. 18, no. 1, pp. 19–30, 2006.





**Andrzej Pacut** holds Ph.D. in electronics and D.Sc. in control and robotics. Since 1969 he is with Warsaw University of Technology, being presently a professor and the head of Biometrics and Machine Learning Group in the Institute of Control and Computation Engineering. Since 2001 he is also the head of Biometric Laboratory of Research and Academic Computer Network NASK. He was Visiting Prof. in the Lefschetz Center for Dynamic Systems at Brown University, Providence, Rhode Island 1980–1981, and Visiting Prof. in the Department of Electrical and Computer Engineering of Oregon State University, Corvallis, Oregon, 1986–1991. He is a senior member of the IEEE and served as the Vice-Chair and then Chair of Poland Section of the IEEE in 2002–2009. He is the head of Technical Committee 309 on Biometrics of Polish Committee for Standardization PKN. He is interested in learning systems, biometrics, identification, bio-inspired modeling and control, and related areas.
e-mail: A.Pacut@ia.pw.edu.pl
Institute of Control and Computation Engineering
Warsaw University of Technology
Nowowiejska st 15/19
00-665 Warsaw, Poland

Biometric Laboratories
Research and Academic Computer Network NASK
Wąwozowa st 18
02-796 Warsaw, Poland

Paper

# The Template Selection in Biometric Systems Based on Binary Iris Codes

Marcin Chochowski

**Abstract—Since the variability of data within readings from the same person is intrinsic property of every biometric system, the problem of finding a good representative – the template – was recognized and present since the beginning of biometrics. This problem was solved differently for different biometric types, yet usually the template somehow averages the collected data samples. However, for the iris type, the template is usually just one or a few samples. In this paper we describe the experiments that suggest that the averaging is also justified in case of iris template creation. This is an important fact, which can significantly improve a performance of biometric template protection methods for iris.**

***Keywords—binary iris codes, biometric template selection, iris biometrics.***

## 1. Introduction

The biometric recognition is based on comparison of the stored representative (the template) for the person in question with the newly acquired biometric sample. The resulting score of such a comparison reflects the similarity (dissimilarity) of the sample to the template. Based on a set threshold the system decides whether this score allows to state that they both originate from the same person or not. Thus it is desirable that the similarity (dissimilarity) between the selected representative and other samples from the same person is above (in case of dissimilarity below) this threshold. This requirement was sufficient for a standard biometric system to be effective.

By standard biometric system we mean a biometric system, where the decision is made upon the direct comparison of the template with the newly acquired sample. In contrast to the standard biometric system are the biometric systems that incorporate technique called the biometric template protection ([1], [2]). In those, the comparison is done not based on the similarity (dissimilarity) of the template and the sample but is an exact match between what is called *pseudonymous identifiers* generated from the template and the sample.

The pseudonymous identifier is a bit string that can be repeatable and with no errors generated from biometric data, possibly with some additional information. It is usually obtained with the help of an error correction mechanism, which might be a quantization scheme, an error-correction code or a secret sharing algorithm. For those algorithms to be efficient (to enlarge the length of the pseudonymous identifier and thus strengthen the security) it is desirable that they need to correct as few errors as possible. This yields for a template that not only will guarantee that the similarity (dissimilarity) will be above (below) some threshold, but also that the similarity (dissimilarity) between template and the samples will be as high (as low) as possible. Thus the problem of selecting the best representative as the template is restated.

## 2. Previous Work

The importance of selecting the best template is often underestimated. It happens that the template is simply any acquired biometric sample with no systematic procedure of its selection. In some cases there is a procedure that selects a sample that is the most similar to other samples of the same person. There are also cases where the template is created as a mean feature vector of collected samples for one person. This is well motivated by the Condorcet rule which states that an estimator (here the template) averaged over many estimators (here each code may be interpreted as an estimator of the ideal code) has smaller variance, thus is better. The question remains how to average. In this section we discuss some known approaches for template selection in different biometric.

### *2.1. Hand*

The hand geometry biometrics uses the features that are very easy to interpret. Those features are the lengths and the widths of the fingers, the widths and the heights of the palm and other geometric features, that are gathered in one fixed-length feature vector $F = [f_1, f_2, ..., f_n]$ that takes the values from $R^n$. As a natural measure of dissimilarity often the Euclidean distance between such a vectors is used. It is a common practice in such a systems that the template is selected as the centroid (the mean vector) of a few samples ([3], [4]). However one must realize an important (though quite simple) fact. If we want to select the point that best represents our set in the sense that it is the closest to all the samples (it minimizes the sum of Euclidean distances between itself and other samples) than it is not the mean vector. The mean vector minimizes the *squared* Euclidean distance and it is not equivalent.

To prove this we have made an experiment with hand geometry system proposed in [4]. We have used the data set

of 149 users with at least 4 hand images each (3 of them were used to create the template and the rest for comparisons). For every user two templates were created – the first one as the mean vector of 3 sample and the second as on of the 3 samples that was closest to other two. For those two templates we calculated genuine and impostor scores (resulting in 179 genuine scores and 45105 impostor ones) using two dissimilarity measures – Euclidean distance (*Euc*) and squared Euclidean distance ($Euc^2$). The results are compared on the basis of the equal error rate (EER) (Table 1) – this maybe a bit simplifying, though it shows an important fact. The method of template selection should be adjusted to dissimilarity/similarity measure, in particular a mean vector is not an appropriate template when using Euclidean distance. For Euclidean distance the best template out of 3 gives better results and for squared Euclidean distance the mean code performs better.

Table 1
EER results for different configurations of template and dissimilarity measure

| EER | *Euc* | $Euc^2$ |
|---|---|---|
| Best [%] | **7.48** | 7.81 |
| Mean [%] | 7.56 | **7.24** |

### *2.2. Fingerprint*

There was much research put into the feature extraction and matching algorithms for fingerprint minutiae but respectively little attention (as in other biometric modalities) was given to the problem of template selection. There were some analysis of different selection of representative fingerprint impression that either best represents the intra-class variations or maximizes the similarity with the rest of the impressions [5]. The results showed that a systematic template selection is much better than random selection. Further work on template creation for fingerprint showed that it is reasonable not to choose a single impression but merge few impressions (mosaicing) of the same fingerprint resulting in bigger coverage of the finger thus better representation.

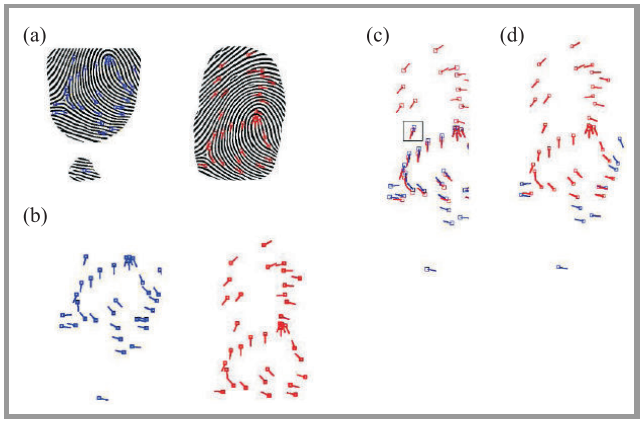



***Fig. 1.*** Fingerprint features mosaicing, (a) two impressions, (b) minutiae extracted from impressions, (c) alignment, (d) mosaicked template [6].

In [6] Ross *et al.* analyzed three different techniques of data merging. The firs was mosaicing on the image level. They aligned the images and merged them using thin plate splines, and then extracted minutiae and performed matching using those minutiae as template. The second approach was to first extract the minutia form two impressions and do the mosaicing on the minutiae level and use the merged minutiae as the template (Fig. 1). The third method was to separately use both impressions (matching two minutiae representations) and fusing the matching scores. The experiments showed that the second method (mosaicing on minutiae level) gives the best results and outperforms single impression matching.

These results are especially important for biometric cryptography (template protection) methods. Most of them that use the fingerprint use the fuzzy vault algorithm (see, e.g., [7]) where a good coverage of fingerprint is one of the most important aspects. This was showed by Nandakumar in his implementation of fuzzy vault for fingerprints ([8]). The usage of mosaiced template improved the results for genuine acceptance by as much as 4% not decreasing the security (false acceptance).

### *2.3. Signature*

Some recent findings in the area of handwritten signatures based on the theory of warped least squares, prove that an template called *the hidden signature* can be defined that greatly improves the performance of matching. This hidden signature can be interpreted as a mean template, but the averaging is done in warped space – for any signature a transform (a warping path) is defined that map it to the space of the warped template were the comparison is done. See [9] for details.

### *2.4. Iris*

As for the iris biometrics, there is no common methodology for template selection. In most cases the template is simply an iris code of acquired image ([10], [11]), or a set of iris codes ([12]). Sometimes like in BiomIris ([13]) the template is chosen as one out of three codes, such that it minimizes the sum of distances to two others. There were also suggestions that average code created as the majority code could be a better representation [14] however that has been argued to have limited use in practice ([15]).

We have to also keep in mind that there are different coding methods for iris recognition that end up with binary code. The question is whether the selection method the best representative would be the same for different algorithms or rather it is algorithm-specific. To address this we propose a few different candidates for the template and verify their effectiveness for two different coding algorithms, namely the OSIRIS implementation of Daugman coding [10] and Czajka's algorithm [16]. OSIRIS is an implementation of

Daugman-like iris texture coding. In our particular realization it produces an binary iris code of length 1974 bits. The Czajka's algorithm represents a different approach to texture coding using Zak-Gabor transform. It produces a binary iris code of length 1024 bits.

# 3. Selection Methods

### *3.1. Notation*

Let us define the following notation that will be valid hereafter:

- $I = \{0,1\}^N$ – space of binary codes of length $N$ (vertices of a unit hyper-cube),
- $A \subset I, A = a_1, a_2, \ldots, a_K$ – set of $K$ available iris codes for particular person, $a$ – iris code,
- $a_{ij}$ – $j^{th}$ bit of $i^{th}$ code for the same person, $i = 1,\ldots,K,\ j = 1,\ldots,N$.

For simplicity, to omit unresolved cases let us assume that the $K$ is odd.

### *3.2. Possible Candidates*

Now we can define different candidates for the iris template. Let us define the average code as

$$\bar{a} = \left(\frac{1}{K}\sum_{i=1}^{K} a_i\right),\ \bar{a} \in R^N,$$

where $a_i$ is an $N$-dimensional iris code $i = 1 \ldots K$.
We can write also,

$$\bar{a} = \arg_{a' \in R^N} \min \sum_{a \in A} \left\| a - a' \right\|^2.$$

This follows from the fact, that the second moment is minimal around the mean value, thus we interpret the code $\bar{a}$ as the real code that minimizes the squared Euclidean distances from all codes from set $A$ – best *represents* them. Let us also define the majority code as

$$a_I^M = \quad \mathbf{Maj}\left(\sum_{i=2}^{K} a_{ij} - \frac{N}{2}\right),\ j = 1,\ldots,N\Bigg),\ a_I^M \in I.$$

This is the code that has $j$th bit equal to 1 if among $K$ codes there were more 1's than 0's on this position and 0 otherwise. Since we assumed $K$ to be odd we excluded the case in which the number of 1's and 0's is equal.
The relation between the code $\bar{a}$ and $a_I^M$ is summarized by the following theorem.

*Theorem 1:* The majority code $a_I^M$ is the nearest code from the subspace $I$ to the average code $\bar{a}$.

*Proof*: The relation between the mean and the median is as follows:

$$\begin{aligned} |m - m_e| = |E(X - m_e)| &\leq E(|X - m_e|), && (1)\\ &\leq E(|X - m|), && (2)\\ &= E(\sqrt{(X-m)^2}),\\ &\leq \sqrt{E((X-m)^2)}, && (3)\\ &= \sigma. \end{aligned}$$

The Eq. (1) inequality comes from the property of sum of absolute values, the Eq. (2) inequality comes from the fact that the median value minimizes the absolute deviation function. The Eq. (3) inequality comes from the Jensen's inequality, for the concave functions (square root function).
Thus the mean $m$ value is less than $\sigma$ from the median $m_e$.

$$|m - m_e| < \sigma$$

what proves the theorem. ■

Yet we know that under taken assumptions ($K$ is odd) we have $\sigma < 0.5$ and $m \in <0,1> - \{0.5\}$. That means that the median code is the closest binary code ($\in I$) to the average code.

$$a_I^M = \arg_{a' \in I} \min \left\| a' - \bar{a} \right\|^2$$

At the same time, from the properties of median, we have

$$a_I^M = \arg_{a' \in I} \min \sum_{a \in A} \left| a - a' \right|.$$

Since $|.|$ and $\|.\|^2$ are equal for the subspace $I$, we see that the $a_I^M$ is an analog of the average code but with constraints to the solution space.

$$a_I^M = \arg_{a' \in I} \min \sum_{a \in A} \left\| a - a' \right\|^2$$

We can also point out two additional codes from the set $A$. The code that is closest to the average code and the code that is closest to majority code. Those are defined as respectively

$$\bar{a}_A = \arg_{a' \in A} \min \left\| a' - \bar{a} \right\|^2 \quad \text{or} \quad (\bar{a}_A = \arg_{a' \in A} \min |a' - \bar{a}|),$$

$$a_A^M = \arg_{a' \in A} \min \left\| a' - a_I^M \right\|^2.$$

There is also a code often used as the template that is defined as

$$a_A^T = \arg_{a' \in A} \min \sum_{a \in A} \left\| a - a' \right\|^2$$

and is the analog of the majority binary code but selected from the set $A$ (set of known sample codes). This is the code previously called the best code.
Intuitively the best representation, contrary to [15], would be the majority code.

# 4. Experiments

According to above discussion we performed a series of experiments to verify the usability of different template selection. In the following experiments we have used part of BATH database (110 eyes with 20 images per eye). For each experiment always the first 10 images were used to create the template and the rest 10 were used as samples for comparisons. The genuine comparisons were performed with the template against 10 genuine samples what makes $1100 = 110 \cdot 10$ comparison in total, and impostors with the template against 10 samples of all other eyes what makes $119900 = 110 \cdot (110-1) \cdot 10$. The experiments were performed for two coding methods - OSIRIS and Czajka's coding. To describe the results we calculated several parameters including false non-match rate (FNMR), false match rate (FMR) and, as suggested in [17], decidability index $d'$. The FNMR was calculated as the rate of positive samples wrongly classified as negative ones, FMR as the rate of negative samples wrongly classified as positive ones and EER as the rate where FNMR and FMR are equal.

### *4.1. OSIRIS Coding*

First we wanted to compare the performance of the recognition algorithms depending on the way the template is created. In particular we compared the performance using the majority code defined as $a_I^M$ the best iris code defined as $a_I^T$ and iris code that is the closest to the majority code $a_A^M$. To compute the majority code we have aligned normalized iris images (in polar format) using 2D correlation, compute the codes for each image and took the median value for each code bit (although the number of samples was even none of the bits for all codes was 0.5). To select the best code we have cross-matched all the 10 codes and selected the one that had the minimal sum of distances to the rest 9.

Next, for different templates, we performed the verification according to the protocol defined above. Figure 2 plots the cumulative distributions (we do not plot the histograms for clarity) of genuine and impostor comparisons for different template selection method. We see that the differences are significant and the best results were obtained for majority code – Table 2 summarizes the results. With the majority code we obtained the perfect separation and good decidability index. Additionally there are plotted results obtained when using as the template one of the 10 codes selected at random labeled as *random iris code*.

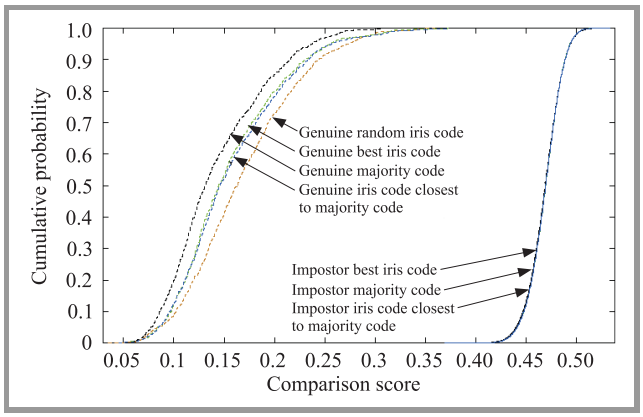



***Fig. 2.*** Cumulative distributions of genuine and impostors scores for different templates (best code, random code, majority code and code closest to majority) for OSIRIS coding algorithm.

Table 2
Summary of verification performance for different template selection method for OSIRIS coding algorithm

| Indexes | Best code | Majority code | Closest to majority code | Mean code |
|---|---|---|---|---|
| EER [%] | 0.0017 | **0** | 0.0017 | 0 |
| FNMR (FMR = 0%) [%] | 0.27 | **0** | 0.18 | 0 |
| $d'$ | 7.82 | **8.84** | 7.62 | 9.06 |

To analyze the averaging property we decided to compare those results with two more possibilities of average template – namely $\bar{a}_A$ (iris code closest to the real-value average) and $\bar{a'}_A$ (iris code closest to the real-value average in $L_1$-norm). The results are plotted in the Fig. 3. Still the majority code outperforms the others, but surprisingly the code $\bar{a}_A$ is as bad as randomly selected code whereas the code $\bar{a'}_A$ is as good as the best code $a_A^T$ and closest to majority code $a_A^M$. There is one more very interesting property worth noticing. The methods that selected one of the iris codes as the template did select different codes thus we cannot infer that some of presented methods are equal. Intuitively we guess that, e.g., $a_A^M$ should be the same as $a_A^T$, but that is not the case.

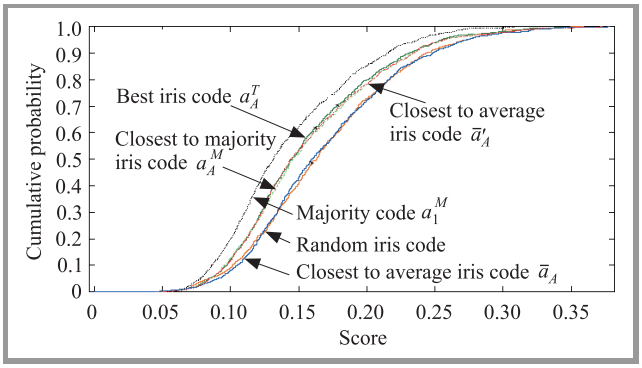



***Fig. 3.*** Genuine comparisons scores (normalized hamming distance) cumulative distributions for different template selection methods

The above described experiment with use of the OSIRIS as the coding algorithm proved the assumption that creating the template code by averaging leads to better performance. The difference between the performance is significant. It is clear that the majority code gives the best results whereas the results with the best code as the template are much worse and with the random code are the worst. This proves

the statement of Davida ([14], the averaging by majority coding as the template creation has influence on the performance and gives much better results.

Noting this fact we decided to go one step further and represent the template as a real-value vector $T_{pr} = [p_1..p_{1974}]$ in which each position $p_i$ represents the rate of this bit was equal one in codes used to create the template. Thus it is a vector with elements from $< 0,1 >$ that could be interpreted as probabilities of 1 on that position in iris code of particular person. This of course makes the template much bigger since it is no longer represented as $N$ bits, but $N$ real numbers, yet the size of it (precision) depends on the number of samples used for template creation and could not be very high. Nevertheless a more complicated problem is the matching algorithm. We can no longer use the ExOR operation and other method should be proposed.

A natural selection of a distance measure is the squared Euclidean distance. The sample codes are the vertices of the 1974-dimension hyper-cube and the templates are points inside this cube. The similarity measure is simply the distance between a vertex and that point. The results on the same data as before with this methodology gives very good results. We obtain perfect separability with decent-looking histograms (Fig. 4).

Yet there could be another similarity measure. If we would extend the $p_i$ to function $P_i$ so that it is a probability function for $i$th position in the code such that $P_i(x = 1) = p_i$ and $P_i(x = 0) = 1 - p_i$ a natural method to verify a new code could be measuring how probable is it, given the template – simply multiply. Of course the probabilities for the elements in the vector are dependent thus multiplying them is not theoretically justified, but the experience in machine learning lets us expect reasonable results. Additionally we have to guarantee that there will be no 0 probabilities to eliminate the effect of zeroing the score (each template element with 0 value, meaning that for all codes used to create the template that particular bit was always 0, was set arbitrary to 0.01). Since there is 1974 bits in the code, calculation of the *pseudo-probability score* ($Pr_{score}$) by multiplying subsequent values is numerically difficult thus we applied log operation and summed the logarithms.

$$Pr_{score}(T_{pr}, a) = \prod_{i=1\ldots1974} (P_i(x{=}a_i)) = \exp\left(\sum_{i=1}^{1974} \log_e(P_i(x{=}a_i))\right),$$

where $T_{pr}$ is the template with probabilities functions $P_i$ for respective positions $= 1 \ldots 1974$. Unfortunately the obtained values of $Pr_{score}$ were of form $\exp(k)$ where $k$ for genuine comparisons was about minus few hundreds

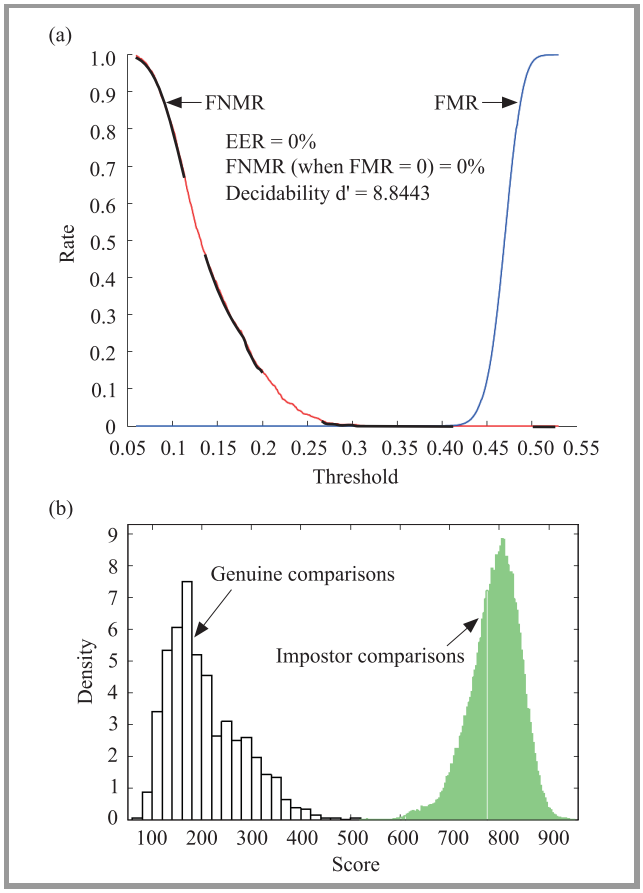



*Fig. 4.* (a) FNMR and FMR graphs, (b) performance rates (left) and comparisons histograms (right) for OSIRIS system with real-valued mean template code and Euclidean distance as similarity measure.

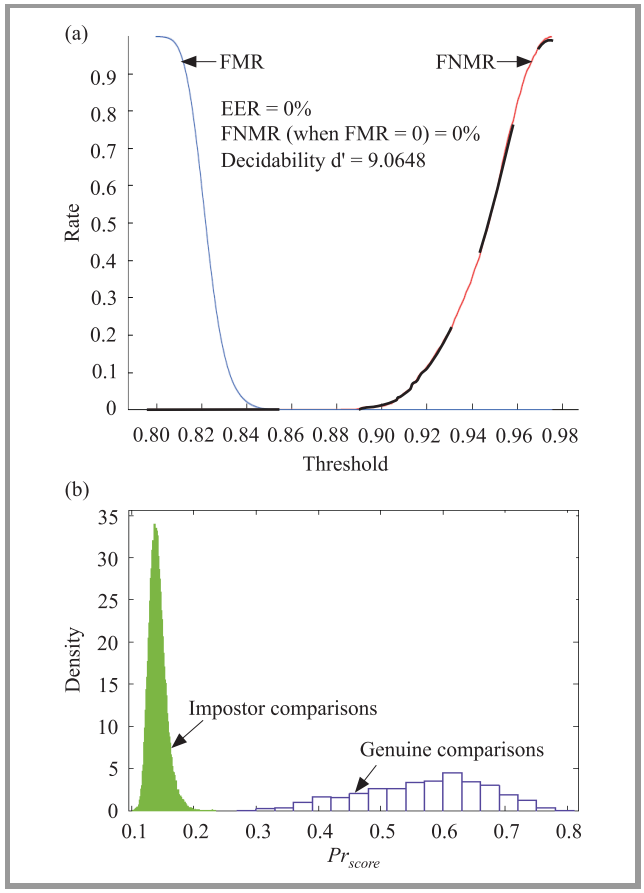



*Fig. 5.* (a) FNMR and FMR graphs, (b) performance rates (left) and comparisons histograms (right) for OSIRIS system with real-valued mean template code and modified matching algorithm.

and for impostors minus few thousands. Thus we decided to divide the $k$ by 1000 to observe the matcher performance. Figure 5 plots the FMR, FNMR for this matcher ($P_{score} = \exp(\sum_{i=1}^{1974} \log_e(p_i)/1000)$).
The results are very promising. With this approach we obtain full separation and the histogram shapes (which have their reflection in $d'$) indicate that such an approach is reasonable and may lead to better results than standard one.

To look into the influence of the template selection on the performance of the iris biometric system we observed also plots representing so-called Dodington-zoo menagerie. This is very helpful for security analysis. It shows whether all irises (with respect to coding algorithm) are equally different or are there some types of irises that either are more similar to others or are less similar to itself. Recent paper from Yager and Dunstone [18] introduced new division and naming for different *behavior* of biometric data depending on mean impostor and genuine scores. We do not want to go into the details of deciding what is a normal behavior and what is not. Instead we want to know whether the different template selection algorithm influences this behavior. Figure 6 plots the menagerie plots for scores obtained with

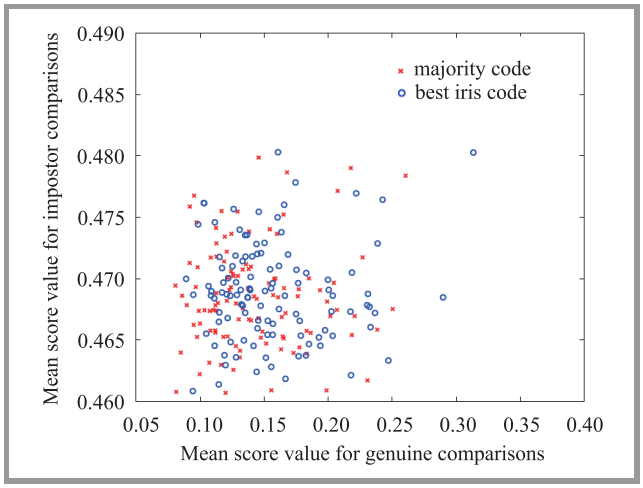



***Fig. 6.*** Menagerie plot for two different template selection methods using OSIRIS coding. Each point represents a single iris showing how well it is on average matched to itself and other irises.

best iris code (circles) and majority code (crosses) as the template. It shows that there is no influence as for the mean impostors scores, what is a good property, and the rightmost mean genuine comparisons for majority code are much smaller what is even better property.

### *4.2. Czajka's coding*

We performed similar experiments using Czajka's coding. Again comparing the genuine cumulative distributions we noticed that the majority code outperforms others giving the best results (Fig. 7). However the behavior of others is significantly different than in case of OSIRIS coding. Here we see that the iris code closest to the majority code gives very poor results (almost as bad as randomly selected iris

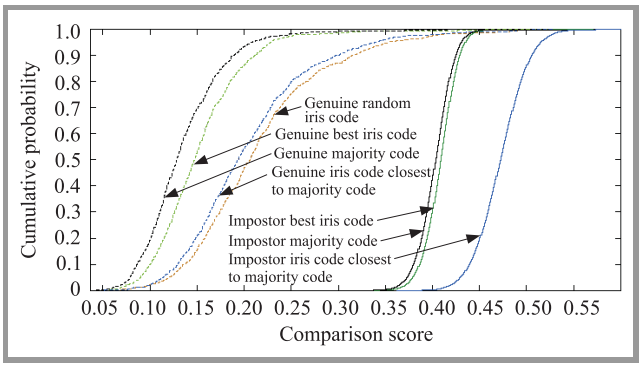



***Fig. 7.*** Cumulative distributions of genuine and impostors scores for different templates (best code, random code, majority code and code closest to majority) for Czajka's coding algorithm.

code). This lets us suspect that the codes created by this algorithm are oddly distributed in the code space, since although majority code estimates the codes well the nearest code does not. Perhaps in this case the majority code does refer any real iris image, but is rather an *virtual object*. The behavior of impostor distributions is even more wired. For OSIRIS, there were no differences for different templates, and here the differences are very significant. We see that the impostor comparisons with majority code are slightly worse (give lower dissimilarity score) and that choosing bad template (iris code closest to majority code gives poor genuine scores) can move the impostors to the right. We can guess that the first observation may be due to not equal distributions of ones and zeros in this type of coding thus averaging may lead to code that better fits different codes (e.g., has more ones). The second observation results from the fact that a bad template in more *noisy*, hence the impostor scores look more random. Both of these facts may prove that this type of coding codes not only the individual characteristics but also some kind of *more global* information. This is quite interesting conclusion and will be a subject for further research.
Table 3 summarizes results of the experiments. We see that we obtained worse results than for OSIRIS but the averaging property of majority code is visible also in this case.

The cumulative distributions from Fig. 7 let us assume that we may expect undesirable distribution changes in menagerie plots. Indeed, Fig. 8 shows that changing the template selection method for majority coding the users tend to be more *wolfy* – mean value of impostor scores gets smaller (different users are more similar).

Table 3
Summary of verification performance for different template selection method for Czajka's coding algorithm

| Indexes | Best code | Majority code | Closest to majority code |
|---|---|---|---|
| EER [%] | 0.662 | **0.542** | 1.60 |
| FNMR (FMR = 0%) [%] | 0.82 | **0.63** | 3.09 |
| $d'$ | 6.87 | **7.69** | 4.99 |

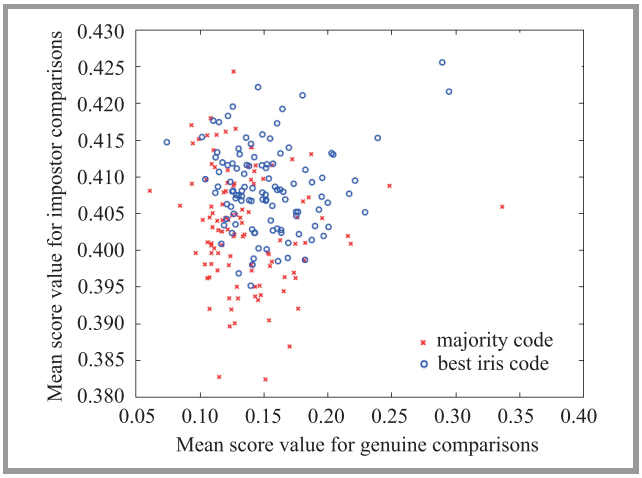



***Fig. 8.*** Menagerie plot for two different template selection methods using Czajka's coding. Each point represents a single iris showing how well it is on average matched to itself and other irises

These experiments prove that a really good understanding of the codes and their properties is needed to propose an bio-encryption algorithm for it.

## 5. Conclusions and further work

Concluding these experiments we claim (in opposite to other authors, e.g., [15]) that for binary iris coding algorithms using the majority code as the template leads to better results. These experiment prove how important is the template selection problem. It was not addressed before in work on biometric template protection, but it seems to be crucial for most of the methods used there. All of them assume that we have a reference code that can be seen as a codeword of error-correcting code and all the query codes lie around it in a distance less than assumed threshold. We showed that depending on the template selection we can obtain different results and that the good understanding of the space of the codes is crucial.

## Acknowledgements

This paper has been financed by the Ministry of Science and Higher Education grant OR00 0026 07 "*A platform of secure biometrics implementations in personal verification and identification*".

## References

[1] ISO/IEC, "CD 24745.2 Information technology – Security techniques – Biometric template protection".

[2] A. K. Jain, K. Nandakumar and A. Nagar, "Biometric template security", *EURASIP J. Adv. Sign. Proces., Spec. Issue Adv. Sign. Proces. Patt. Recogn. Meth. Biometr.*, pp. 1–17, Jan. 2008.

[3] R. Sanchez-Reillo and C. Sanchez-Avila, "Biometric identification through hand geometry measurements", *IEEE Trans. Patt. Anal. Mach. Intell.*, vol. 22, no. 10, pp. 1168–1171, 2000.

[4] L. Stasiak, "Weryfikacja tożsamości poprzez wykorzystanie cech dłoni", M.Sc. thesis, Institute of Control and Computation Engineering, Warsaw University of Technology, Warsaw, 2004.

[5] U. Uludag, A. Ross, and A. K. Jain, "Biometric template selection and update: a case study in fingerprints", *Patt. Recogn.*, vol. 37, no. 7, pp. 1533–1542, 2004.

[6] A. Ross, S. Shah, and J. Shah, "Image versus feature mosaicing: a case study in fingerprints", in *Proc. SPIE, Biometric Technologies for Human Identification III*, Orlando, USA, 2006, vol. 6202, pp. 1–12.

[7] A. Juels and M. Sudan, "A fuzzy vault scheme", in *IEEE Int. Symp. Infor. Theory*, Lausanne, Switzerland, 2002, pp. 408–421.

[8] K. Nandakumar, A. K. Jain and S. Pankanti, "Fingerprint-based fuzzy vault: implementation and performance", *IEEE Trans. Inform. Foren. Secur.*, issue 4, vol. 2, pp. 744–757, 2007.

[9] J. Putz-Leszczyńska and A. Pacut, "Hidden signature – a new solution for on-line verification using DTW", in *Proc. 42nd Ann. IEEE Int.l Carnahan Conf. Secur. Technol. ICCST 2008*, Prague, Czech Republic, 2008, pp. 162–166.

[10] E. Krichen, A. Mellakh, S. Salicetti, and B. Dorizzi, "OSIRIS (Open Source for IRIS) reference system", *BioSecure Project* [Online]. Available: http://www.cilab.upf.edu/biosecure1/public docs deli/BioSecure Deliverable D02-2-2 b4.pdf.pdf

[11] L. Masek and P. Kovesi, *MATLAB Source Code for a Biometric Identification System Based on Iris Patterns*. The School of Computer Science and Software Engineering, The University of Western Australia, 2003.

[12] "VeriEye 2.2 SDK" [Online]. Available: www.neurotechnology.com

[13] "BiomIris SDK" [Online]. Available: www.BiometricLabs.pl

[14] G. I. Davida, Y. Frankel, and B. J. Matt, "On enabling secure applications through off-line biometric identification", in *Proc. IEEE Symp. Secur. Priv. 1998*, Oakland, USA, 1998, pp. 148–157, 1998.

[15] F. Hao, R. Anderson, and J. Daugman, "Combining cryptography with biometrics effectively", Techn. Rep. UCAM-CL-TR-640, University of Cambridge, 2005.

[16] A. Pacut, A. Czajka, and P. Strzelczyk, "Iris biometrics for secure remote access", in *Cyberspace Security and Defense: Research Issues*, J. S. Kowalik *et al.*, Eds. Dordrecht: Springer, 2005, pp. 259–278.

[17] J. Daugman, "High confidence personal identification by rapid video analysis of iris texture", in *Proc. IEEE Int. Carnahan Conf. Secur. Technol.*, Atlanta USA, 1992, pp. 50–60.

[18] N. Yager and T. Dunstone, "The Biometric Menagerie", *IEEE Trans. Patt. Anal. Mach. Intell.*, vol. 32, no. 2, pp. 220–230, 2010.



**Marcin Chochowski** received his M.Sc. in 2004 from the Faculty of Electronics and Information Technology of the Warsaw University of Technology. He is presently a Ph.D. student at the Institute of Control and Computation Engineering at the Faculty of Electronics and Information Technology of the Warsaw University of Technology. Since 2003, he is also a research assistant at the Biometric Laboratory of Research and Academic Computer Network NASK. He is interested security aspects of biometric systems, artificial intelligence, statistics and related areas.

e-mail: Marcin.Chochowski@nask.pl
Research and Academic Computer Network (NASK)
Biometric Laboratories
Wąwozowa st 18
02-796 Warsaw, Poland

Paper

# Impact of Crosstalk into High Resistivity Silicon Substrate on the RF Performance of SOI MOSFET

Khaled Ben Ali, César Roda Neve, Ali Gharsallah, and Jean-Pierre Raskin

**Abstract—Crosstalk propagation through silicon substrate is a serious limiting factor on the performance of the RF devices and circuits. In this work, substrate crosstalk into high resistivity silicon substrate is experimentally analyzed and the impact on the RF behavior of silicon-on-insulator (SOI) MOS transistors is discussed. The injection of a 10 V peak-to-peak single tone noise signal at a frequency of 3 MHz ($f_{noise}$) generates two sideband tones of −56 dBm separated by $f_{noise}$ from the RF output signal of a partially depleted SOI MOSFET at 1 GHz and 4.1 dBm. The efficiency of the introduction of a trap-rich polysilicon layer located underneath the buried oxide (BOX) of the high resistivity (HR) SOI wafer in the reduction of the sideband noise tones is demonstrated. An equivalent circuit to model and analyze the generation of these sideband noise tones is proposed.**

***Keywords— crosstalk, high resistivity Si, mixing products, passivation layer, polysilicon.***

## 1. Introduction

During the last 50 years the performance of silicon integrated circuits (ICs) has been growing exponentially thanks to the continuous downscaling of the transistors and components size. CMOS transistors have greatly increased their operating frequency and today ICs reach a high integration level of analog and digital circuits on the same silicon chip [1]. However, due to such extremely small dimensions the conventional bulk devices are facing many challenges, such as short-channel effects (SCE), junction capacitances and doping fluctuation [2].

In the recent years, silicon-on-insulator (SOI) technology, and more particularly SOI MOSFET, has attracted more attention due to its better scaling capability, higher isolation, reduced parasitic components and higher performance compared to bulk Si technology. As it is shown in Fig. 1, the buried oxide film (BOX) in SOI prevents the latch-up effect observed in bulk technology and reduces the substrate coupling mechanisms at low frequencies [2]. However, SOI technology suffers from parasitic effects such as floating body effect and self heating. These effects can be more or less pronounced depending on the SOI CMOS technology: fully (FD) or partially depleted (PD). The difference between these two families is the thickness of the maximum depletion zone above the buried oxide called the body. The fabrication and modeling of PD devices present substantially lower cost and complexity than those of FD devices, which is essentially related to the higher compatibility between bulk and PD SOI devices processing techniques. One of the main challenges that are associated with the PD SOI technology is the control of the floating body effect, which results from the generation of excess charges in the SOI body, which changes the channel potential, and changes the behavior of the MOSFET [3].

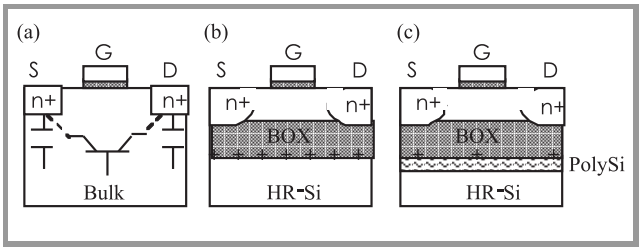



***Fig. 1.*** nMOSFET cross section: (a) bulk technology, (b) HR-SOI, and (c) passivated HR-SOI technology.

The integration of CMOS and non-CMOS technologies into the same chip (systems-on-chip, SoC) is foreseen as a way to reduce the cost of each IC, and to fabricate smaller devices with higher performance and less power consumption [1]. A limiting factor of highly integrated electronics is the coupling through the common silicon substrate between the digital and the analog circuitries. In that sense, the main advantage of SOI compared to bulk Si is its compatibility with the use of high resistivity substrates to reduce substrate coupling and RF losses.

The introduction of high resistivity Si (HR-Si) substrate has converted silicon into a suitable technology for high frequency applications [4]. However, it is known that oxidized high resistivity substrate suffers from parasitic surface conduction (PSC) effect. Positive fixed charges inside the oxide attract electrons to the interface, creating an inversion/accumulation layer at the Si $SiO_2$ interface [5]. This thin highly conductive layer is responsible for the substrate losses. This issue can be overcome by introducing a trap-rich passivation layer between the oxide and the HR-Si substrate which captures the free carriers and locally depletes the HR-Si substrate. Several techniques can be used to generate such trap-rich layer: micromachined structures [6], ion implantation [7], and deposition of an amorphous silicon [8] or polycrystalline silicon layer [4].

In this paper, we investigate the RF behavior of a PD SOI MOSFET when a sinusoidal noise signal is injected through a metallic pad close to the transistor. The introduction of

a trap- rich polysilicon (PolySi) layer at the $Si/SiO_2$ interface of an oxidized HR-Si, as a way to reduce substrate crosstalk, is analyzed.
This paper is structured as follows. Firstly, a theoretical analysis of substrate crosstalk in bulk Si and SOI technologies is presented. Secondly, the impact of PSC effect in oxidized HR-Si and its reduction by using a polySi layer is demonstrated. Next, we describe the CMOS device and technology under analysis, as well as the measurement setup. Finally, we show and discuss the results obtained from the MOS structures lying on HR-Si substrates with and without passivation layer.

## 2. Theoretical Analysis of Substrate Crosstalk in Bulk Si and SOI Substrates

Crosstalk can be defined, in a more general way, as the electromagnetic disturbance induced by a circuit in another one, located nearby. From this general definition it can be deduced that any mechanism that creates such type of interference from a circuit into another falls into the crosstalk definition. In the case of mixed-mode high-frequency ICs, coupling through the substrate – the so-called substrate crosstalk – is one of the main origins of interferences. It is recognized as one of the most limiting factors in the performance of RF ICs [6].
For semiconductor substrates, crosstalk can be divided into two different mechanisms:

– injection into the substrate,

– propagation of this noise signal through the substrate.

It can be easily deduced that regardless of circuit design solutions, two different ways to reduce substrate crosstalk exist:

– substrate isolation,

– substrate coupling reduction.

As it is shown in [6], SOI substrate provides a higher isolation from the lossy Si thanks to its buried oxide layer (BOX) underneath the active layer. In combination with HR-Si it also effectively reduces conductive coupling through the substrate.
An extensive analysis of crosstalk in Si and SOI substrates has been presented in [6]. Hereafter, we briefly introduce some of the considerations related to crosstalk in Si substrates.

### *2.1. Crosstalk in Bulk Si Substrate*

Silicon, as any semiconductor material, exhibits both conductive and dielectric characteristics, which can be translated into a resistive and a capacitive effect, respectively.

**Resistive effect**. At frequencies below a certain crossover frequency ($f_s$) the conductive nature of the semiconductor dominates over the dielectric behavior. Thus, the substrate can be modeled as purely resistive. The conductivity for a doped semiconductor is given by:

$$\sigma = q(p\mu_p + n\mu_n), \tag{1}$$

where $q$ is the electron charge, and $\mu_n$ and $\mu_p$ represent the mobility of the electrons and holes carriers, respectively, and $n$ and $p$ stand for respective carrier densities.
While the effective carrier mobility depends also on the number of carriers (scattering effect), the expression (1) is dominated by the carrier concentration. We can say, as a first-order approximation, that the conductivity is an increasing function of the carrier densities.

**Capacitive effect**. At frequencies above the crossover frequency, the dielectric behavior of the semiconductor can no longer be neglected, thus the substrate must be modeled as a resistive and capacitive network as shown in Fig. 2.

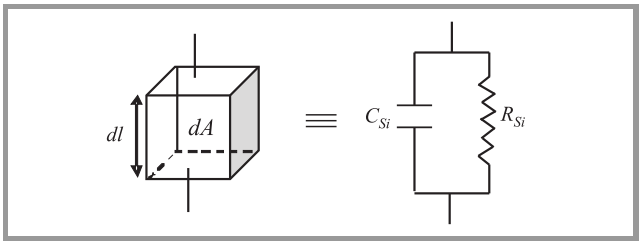



***Fig. 2.*** Equivalent R-C model of a small piece of a homogeneous semiconductor substrate.

In the frequency domain the equivalent admittance $Y_s$ for a piece of substrate is given by:

$$Y_S = \frac{1 + j\omega R_S C_S}{R_S} = \frac{1 + j\omega T_S}{R_S}, \tag{2}$$

where $C_S$ is the associated substrate capacitance, $R_S$ is the substrate equivalent resistance and $T_S$ is the substrate time constant given by:

$$T_S = R_S C_S = \frac{\rho_S dl}{dA}\,\frac{\varepsilon_0 \varepsilon_{Si} dA}{dl} = \frac{\varepsilon_0 \varepsilon_{Si}}{q(p\mu_p + n\mu_n)}, \tag{3}$$

where $\varepsilon_0$ is the vacuum permittivity, $\varepsilon_{Si}$ is the silicon permittivity, $\omega$ is the angular velocity, $dl$ is the elementary length of the small piece of homogenous substrate and $dA$ is its elementary area.
As it can be seen in Eq. (3), $T_S$ is not related to the dimensions of the considered semiconductor volume but only to the substrate electrical properties.
At low frequencies, substrate resistance, $R_S$, is more important and the associated capacitance, $C_S$, can be neglected. As the pulsation $\omega$ increases, the impedance relative to the capacitive effect decreases to become equal to the that of the resistive effect at the crossover frequency, $f_T$, defined by:

$$f_T = \frac{1}{2\pi T_S} = \frac{Q(p\mu_p + n\mu_n)}{2\pi\varepsilon_0\varepsilon_{Si}}. \tag{4}$$

The lumped equivalent circuit representing the coupling between two metallic pads of same size lying on a bulk Si or a SOI substrate is shown in Fig. 3, where $C_{Si}$ and $R_{Si}$ describe the coupling effect between each pad and the back

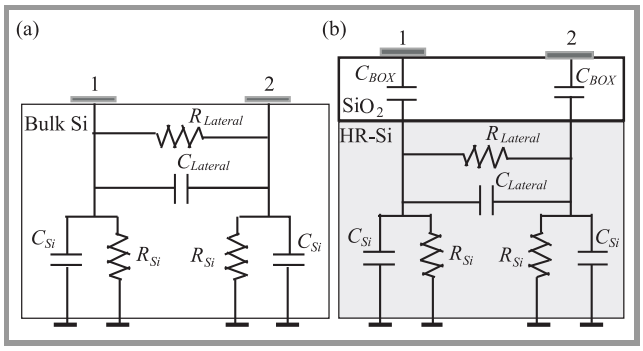



***Fig. 3.*** Lumped equivalent R-C model substrate crosstalk in the case of (a) bulk Si and (b) SOI substrates.

side metallization. The propagation through the substrate is modeled by the elements $R_{Lateral}$ and $C_{Lateral}$. For identical pads with an area of $S_{pad} = W \times W$ and a spacing width $d$, the expressions of $R_{Si}$, $C_{Si}$, $R_{Lateral}$ and $C_{Lateral}$ derived from [6], are given by:

$$R_{Si} = \left[ K \frac{\sigma_{Si} S_{pad}}{t_{Si}} \right]^{-1} \quad [\Omega], \tag{5}$$

$$C_{Si} = K \frac{\varepsilon_0 \varepsilon_{Si} S_{pad}}{t_{Si}} \quad [\mathrm{F}], \tag{6}$$

$$R_{Lateral} = \left[ K \frac{\pi \sigma_{Si}}{4 \ln\left[ \frac{\pi(d-W)}{W+t} + 1 \right]} W \right]^{-1} \quad [\Omega], \tag{7}$$

$$C_{Lateral} = K \frac{\pi \varepsilon_0 (\varepsilon_{Si} + 1)}{4 \ln\left[ \frac{\pi(d-W)}{W+t} + 1 \right]} W \quad [\mathrm{F}], \tag{8}$$

where $K$ is the fringing factor, $t$ is the thickness of the conductors, $t_{Si}$ is the Si substrate thickness and $\sigma_{Si}$ is the Si substrate conductivity.

### *2.2. Crosstalk in SOI Substrate*

Thanks to the presence of the buried oxide (BOX), the injection of signal to the silicon substrate is well reduced especially at low frequencies. As it can be seen in Fig. 3b, this can be modeled in the lumped equivalent circuit by a capacitance $C_{BOX}$ between the pad and the silicon substrate, defined as:

$$C_{BOX} = K \frac{\varepsilon_0 \varepsilon_{ox} S}{t_{ox}}, \tag{9}$$

where $\varepsilon_{ox}$ and $t_{ox}$ are, respectively, the permittivity and thickness of the oxide layer.

As explained before, the propagation through the silicon substrate is modeled by the $R_{Lateral}$ and $C_{Lateral}$, which depend only on the Si properties. By reducing the conductivity of the silicon substrate in Eq. (7) we reduce this conductive coupling path through the substrate. This can be achieved by the use of a high resistivity (HR) SOI substrate.

However, because the buried oxide layer presents a low but not null density of oxide fixed charges and interface charges at the BOX/HR-Si interface region, a parasitic surface conduction (PSC) will greatly affect the equivalent resistance $R_{Lateral}$ in Eq. (7) by increasing the effective conductivity of silicon. In fact, an electron inversion layer is created producing a highly conductive surface layer. If a highly trap-rich layer (such as polysilicon) is deposited below the BOX and the Si substrate, these free carriers will be trapped and the parasitic conduction effect will be reduced. The nominal value of silicon resistivity is then recovered.

## 3. Impact of PSC Effect in HR-Si with and without a Trap-Rich Passivation Layer

To analyze the impact of the PSC effect in oxidized HR-Si we measure the attenuation of a coplanar waveguide (CPW) transmission line. CPW lines are used to characterize the interface properties due to their high sensitivity to the properties of the substrate surface they are printed on.

As explained before, the presence of fixed oxide charges in the silicon dioxide creates an inversion/accumulation layer at the $Si/SiO_2$ interface, hence increasing the conductivity at that zone. It has been demonstrated [9] that, for a CPW line, when a highly conductive layer is located between the top insulator and a high-resistivity substrate, the electric field is mainly located in the conductive layer, instead of going deeper into the substrate. The effective resistivity of the substrate is no longer the nominal value but a much lower value, typically one order of magnitude lower [7].

It has been demonstrated [7], [8], [10] that the impact of the PSC effect can be reduced by introducing traps at the $Si/SiO_2$ interface. These traps will freeze the free carriers, and by reducing their mobility, the nominal resistivity of the substrate will be recovered. The introduction of such a trap-rich passivation layer can be easily done by deposition of a polysilicon layer between the top oxide and the HR-Si substrate. This solution has the advantage that it is thermally stable and compatible with CMOS processes [9].

The extracted attenuation constants of a CPW line lying on different substrates are shown in Fig. 4. The CPW dimensions are 26 þm for the width ($W$) of the central conductor, 12 $\mu$m for the slot width ($S$), and 208 $\mu$m for the width of the planar ground conductors ($Wg$). Three different substrates are measured: a bulk Si with standard resistivity ($\rho = 20$ Ωcm), one HR-Si substrate with $\rho = 5$ kΩcm and the same HR-Si substrate but including a sandwiched

polySi layer of 287 nm-thick. In all cases front and back metal layers are 1 $\mu$m-thick aluminum and the thermal $SiO_2$ top layer is of 50 nm. As it can be observed, the CPW with the passivation layer is the line which presents the lowest attenuation level, at least 0.2 dB/mm lower than that of the same HR-Si substrate without passivation.

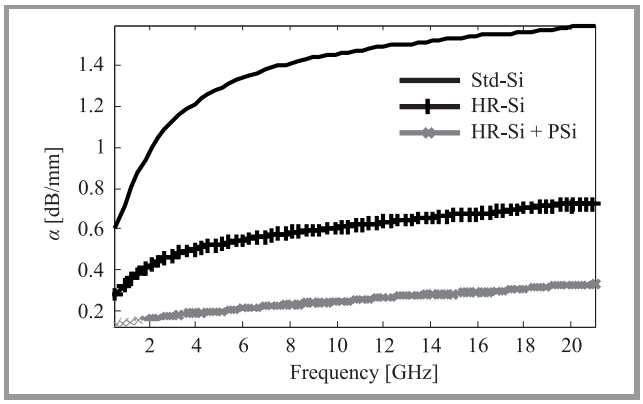



***Fig. 4.*** Measured losses as a function of frequency for CPW line lying on three different Si substrates.

To have a more clear representation of the impact of the resistivity variation at the interface we can compare the effective resistivity of each substrate. It is calculated as described in [11], by comparing the extracted conductance of the RLCG equivalent circuit of the CPW transmission line with the one of a theoretical line with the same dimensions lying on a lossless Si substrate. The effective resistivity for the three different substrates is presented in Fig. 5. It can be clearly seen that the effective resistivity of the standard Si substrate remains closer to its nominal

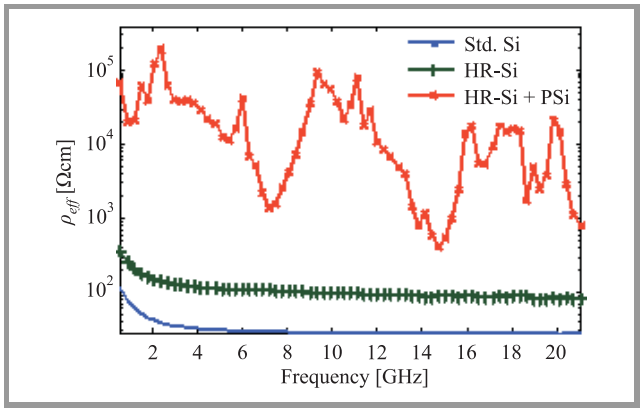



***Fig. 5.*** Effective resistivity as a function of frequency for CPW line lying on three different Si substrates.

value, as expected since the PSC effect has a low impact on such type of substrate that already presents intrinsically a certain concentration of free carriers. However, for the HR-Si substrate the effective resistivity ($p_{eff}$) is approximately equal to 100 Ωcm. This value is 50 times lower than its nominal value. On the other hand, the same HR-Si substrate with a trap-rich passivation layer exhibits an effective resistivity which is kept higher than 1 kΩcm. The efficiency of the passivation layer to reduce the PSC effect is demonstrated as the high-resistivity characteristic of the oxidized HR-Si substrate is recovered. Similar conclusions can be extrapolated to substrate crosstalk, where thanks to the introduction of a trap-rich passivation layer the impact of the conductive layer below the oxide will effectively vanish, hence reducing the conductive coupling through the substrate.

## 4. Device and Measurement Setup Description

Two partially depleted PD SOI MOSFETS with gate lengths of 0.13 and 0.24 $\mu$m are used to analyze the impact of substrate crosstalk in their RF performance. Both PD MOSFETS are fabricated using a commercial 0.13 and 0.24 $\mu$m, respectively, SOI CMOS process by ST-Microelectronics. This is a single poly CMOS process using SOI UNIBOND wafers with high resistivity Si ($\rho_{Si} > 3$ kΩcm) as starting substrate material, and 450 nm-thick BOX. This process features lines with six copper metal layers (M1 to M6) and one top aluminium layer to reduce conductor losses [9].

To compare their performance when a passivated wafer is used, a second set of identical PD SOI MOSFETS is obtained by transferring the processed layers onto a passivated HR Si wafer, with a 300 nm-thick polysilicon layer, with the method presented in [9] and patented by TraciT Technologies. It must be mentioned that an additional 525 nm-thick PECVD oxide, needed for the transfer process, increases the final BOX thickness of the passivated wafer to 975 nm.

The aim of our experimental setup is to assess the impact on the RF performance of the PD SOI MOSFETs when a noise signal is injected in the vicinity of the transistor, for passivated and unpassivated substrates. This can be achieved by comparing the spectral response (output signal) of the device with and without the injection of a known signal noise. On-wafer measurements are performed using a probe station, a WILTRON 68147A (10 MHz – 20 GHz) sweep generator for the generation of the 1 GHz RF input

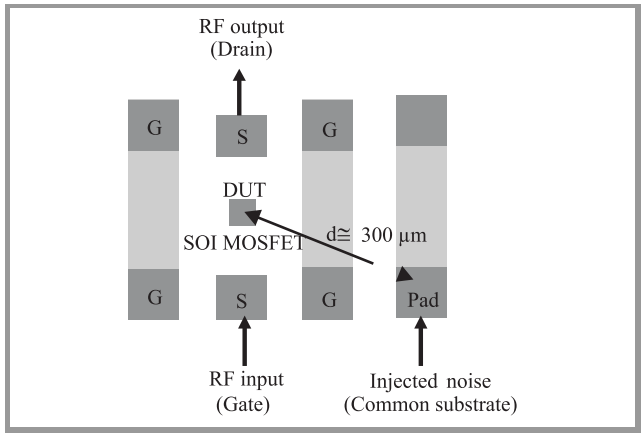



***Fig. 6.*** SOI MOSFET crosstalk structure, S and G stand for signal and ground, respectively.

signal, and a HP33120 low frequency generator for the injection of a noise signal through a metallic pad near the device under test and lying on the same substrate. The RF output of the transistor is visualized using an Agilent E4440 wideband spectrum analyzer. A top view of the measured device, shown in Fig. 6, demonstrates that the distance between the device and the noise source is 300 $\mu$m.

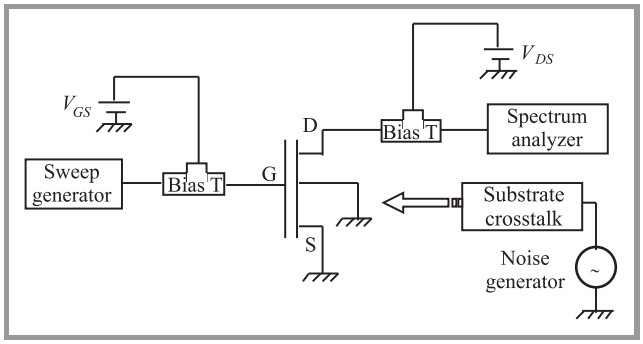



***Fig. 7.*** Measurement setup of a transistor spectral response with injected noise signal.

A diagram of the measurement setup is presented in Fig. 7. It shows the biasing of the transistor, the RF input signal at the gate node, and the output spectral response at the drain node of the transistor.

## 5. Results and Discussion

Without injection of a noise signal, Fig. 8 shows the output signal spectrum, (a) without and (b) with a passivation layer, for a 0.24 $\mu$m SOI MOSFET biased at $V_{GS}$=1 V and $V_{DS} = 1$ V for an input power of +10 dBm applied at the gate of the transistor. We can see that the generated intrinsic noise levels of the transistor are similar for both wafers. The amplitude of the output single tone for the transistor on the passivated substrate is slightly better (2 dB).

When a noise signal of a 10 V peak-to-peak at a frequency of 3 MHz ($f_{noise}$) is injected near the transistor we observe harmonics at 1 GHz $-f_{noise}$ and 1 GHz $+f_{noise}$ as shown in Fig. 9a. The amplitude of these sideband tones is $-56$ dBm.

The generation of these mixing products of the RF input and the noise signal at the RF output can be explained by the substrate coupling of the noise signal into the MOSFET. Although it was expected that at such low frequencies the BOX layer of the SOI provides a high enough level of isolation, the presence of a highly conductive surface at the BOX/HR-Si interface allows an efficient coupling of the large noise signal. As it is shown in Fig. 10, two possible coupling paths can be identified:

– one, that goes from the metallic pad of the noise injection to the RF input pad of the MOSFET and to the pad-to-gate interconnection,

– and a second one, that goes from the metallic pad of the noise injection to the back gate of the transistor.

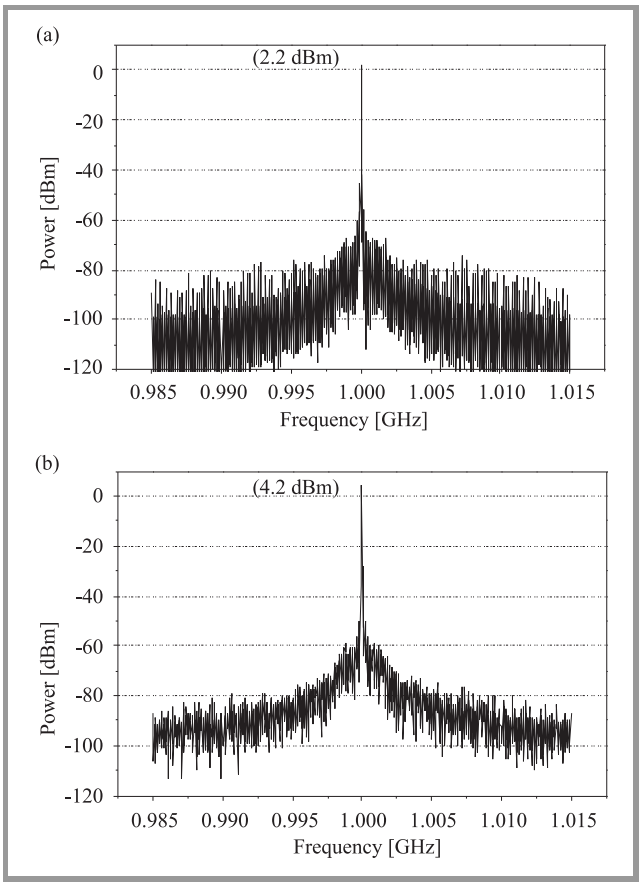



***Fig. 8.*** Spectral response at 1 GHz without injected noise signal for a SOI PD MOSFET ($L_g = 0.24$ $\mu$m) on a (a) unpassivated and (b) passivated HR SOI substrate.

These paths are labelled as "front gate coupling" and "back gate modulation", respectively. As it is known the transconductance of the transistor from the back gate is lower than that from the front gate, we can assume that the "front gate coupling" is more important than the "back gate modulation", and thus that the sum of the noise and the RF input signals reach the gate of the MOSFET, and are mixed inside the transistor.

These assumptions are well confirmed when a passivation layer underneath the BOX is introduced. In fact, as it can be seen in Fig. 9b, the mixing products due to the presence of the noise signal fall below the noise floor of our measurement. A reduction of at least 15 dB is then measured for the sideband tones induced by the low frequency noise signal. The decrease of the noise level can be explained by the reduction of free carriers at the $Si/SiO_2$ interface. Indeed, the introduction of a high density of traps into the silicon surface will pin the surface potential and minimize the creation of a conductive inversion layer. Thus, the passivated SOI HR-Si substrate proved to have better crosstalk immunity.

Table 1 shows the detected power levels of the mixing products due to the noise signal for the two SOI MOSFETs (0.13 $\mu$m and 0.24 $\mu$m) when the frequency of the noise signal varies from 100 kHz to 3 MHz. These levels are

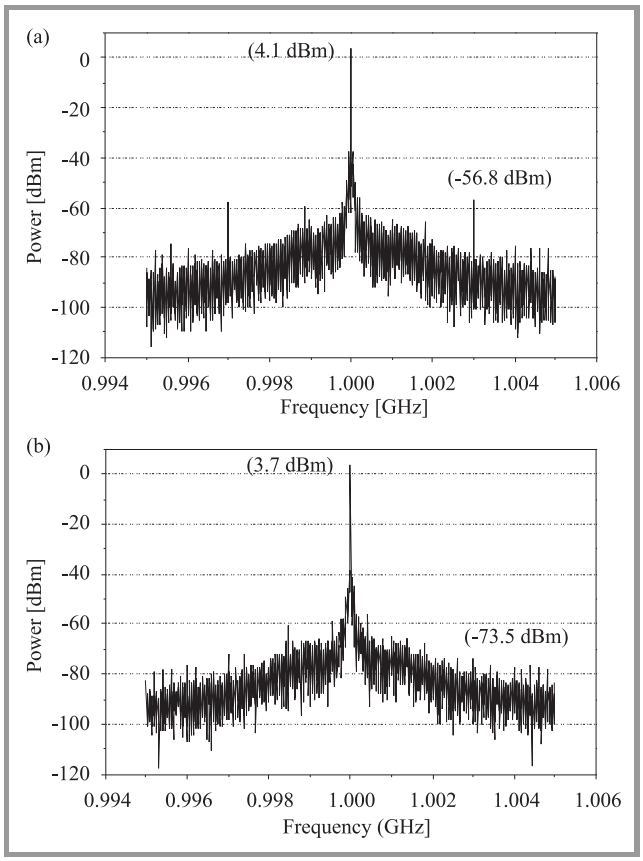



***Fig. 9.*** Ouput spectrum around 1 GHz for a SOI PD MOSFET (0.24 $\mu$m), $f_{noise} = 3$ MHz: (a) unpassivated and (b) passivated HR SOI substrates.

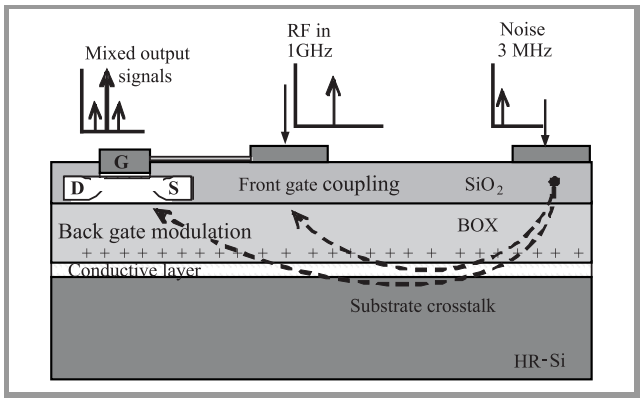



***Fig. 10.*** Substrate crosstalk schematic between the SOI MOSFET and the metallic pad at which the noise signal is injected.

Table 1
Harmonics levels in dB at 1 GHz $\pm f_{noise}$
for two PD SOI MOSFETs

| Freq. | $L_g = 0.13$ $\mu$m | | $L_g = 0.24$ $\mu$m | |
|---|---|---|---|---|
| [MHz] | Unpassivated | Passivated | Unpassivated | Passivated |
| 0.1 | –63 | –62 | –56 | –64.4 |
| 0.5 | –53.7 | –71 | –56.7 | –69.6 |
| 1 | –52.8 | –61.4 | –52.8 | –64.3 |
| 3 | –58.1 | –71.7 | –56.8 | –73.5 |

similar for both technologies, and in the case of passivated wafer they are below the noise floor of the measurement.

This improvement confirms that the use of the trap-rich passivation layer under the BOX act as a very effective stabilizing layer for SOI HR-Si wafers.

In order to model the coupling effect through the high resistivity SOI substrate a simple equivalent circuit that is presented in Fig.3b is proposed. The substrate is modeled as R-C networks where the elements $R_{Lateral}$ and $C_{Lateral}$ model the substrate crosstalk. As explained before, the use of the polysilicon layer can improve the effective resistivity of the substrate and hence reduce the coupling effect. This translates to the increase of the values of $R_{Lateral}$ as seen in Table 2.

Table 2
$R_{Lateral}$ with varying substrate effective resistivities

| $\rho_{eff}$ [$\Omega$cm] | $R_{Lateral}$ [k$\Omega$] |
|---|---|
| 20 | 5.05 |
| 200 | 50.5 |
| 5,000 | 1,250 |

These values derive from Eq. (7) where the distance between the two devices is $d = 300$ $\mu$m and the width of the metallic pad is $W = 100$ $\mu$m.

To assess the attenuation of the noise levels with and without the passivation layer, this equivalent circuit is then connected to the small-signal equivalent circuit of the SOI MOSFET transistor and simulated with SPICE Eldo software.

In Fig. 11, dual-tone input signals are introduced at the gate node of the transistor. The output signal shows a mixing product of the low frequency noise and the RF 1 GHz signals. Mitigation of R-C networks modeling the transfer function of the substrate is calculated in the case of unpassivated SOI substrate, and then introduced in terms of noise signal amplitude.

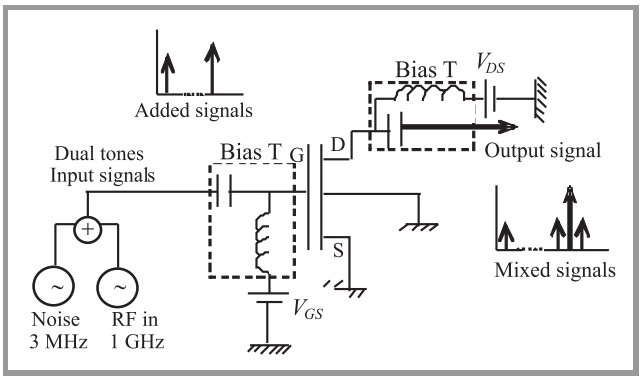



***Fig. 11.*** Simulated equivalent circuit.

Using a noise signal level of 0.01 V peak-to-peak at 3 MHz we simulate sideband tones with an amplitude of –56 dBm which was previously observed experimentally in Fig. 9a.

When a passivated layer is introduced the effective resistivity of the wafer is improved. This is translated into the increase of $R_{Lateral}$ and hence an attenuation of the sideband tones levels due to the noise signal. This reduction is assumed to be more than 15 dB as presented in the measurement results.

## 6. Conclusions

The significance of the noise crosstalk through the Si substrate and the related effects on RF circuit are highlighted. Our measurements demonstrate the efficiency of a trap-rich layer, e.g., the polycrystalline silicon, to reach a very high recombination rate at the Si substrate-BOX interface and thus minimize the impact of crosstalk and surface conduction. A model of the substrate losses and propagated noise signal at the circuit level is proposed. Measurement and simulation results depict a reduction of the noise level by more than 15 dB when the substrate is passivated. In fact, a thin polycrystalline silicon surface layer acts as a very effective stabilizing layer for HR SOI substrate. Theses results are of importance for mixed-mode ICs in which coupling phenomena between digital and sensitive analog parts must be minimized. Thus high-resistivity SOI substrate with a polysilicon surface layer underneath the BOX is a viable substrate for low loss RF and mixed-mode applications.

## Acknowledgments

The authors would like to thank P. Simon for RF device measurement, ST microelectronics, Crolles, France, for processing the wafer, TraciT Technologies, Grenoble, France, for transferring the processed SOI layers onto the passivated substrate.

## References

[1] *International Technology Roadmap for Semiconductor (ITRS)*. 2007 Edition.

[2] O. Rozeau, J. Jomaah, S. Haendler, J. Boussey, and F. Balestra, "SOI technologies overview for low power low voltage radio frequency applications", *Analog Integr. Circ. Sign. Proc.*, vol. 25, pp. 93–114, 2000.

[3] H. Xiao, R. Huang, J. Liang, H. Liu, Y. Tian, R. Wang, and Y. Wang, "The localized-SOI MOSFET as a candidate for analog/RF applications", *IEEE Trans. Electron Dev.*, vol. 54, no. 8, pp. 1978–1984, 2007.

[4] H. S. Gamble, B. M. Armstrong, S. J. N. Mitchell, Y. Wu, V. F. Fusco, and J. A. C. Stewart, "Low-loss CPW on surface stabilized high-resistivity silicon", *IEEE Microw. Guid. Wave Lett.*, vol. 9, no. 10, pp. 395–297, 1999.

[5] D. Lederer and J.-P. Raskin, "Substrate loss mechnisms for microstrip and CPW transmission lines on lossy silicon wafers", *Solid-State Electron.*, vol. 47, pp. 1927–1936, 2003.

[6] J.-P. Raskin, A. Viviani, D. Flandre, and J.-P.Colinge, "Substrate crosstalk reduction using SOI technology", *IEEE Trans. Electron Dev.*, vol. 44, no. 12, pp. 2252–2261, 1997.

[7] D. Lederer, J.-P. Raskin, "RF performance of a commercial SOI technology transferred onto a passivated HR silicon substrate", *IEEE Trans. Electron Dev.*, vol. 55, no. 7, pp. 1664–1671, 2008.

[8] Y. H. Wu, A. Chin, K. H. Shih, C. C. Wu, C. P. Liao, S. C. Pai, and C. C. Chi, "RF loss and crosstalk on extremely high resistivity (10 k – 1 MΩcm) Si fabricated by ion implantation", *IEEE MTT-S Digest*, 2000.

[9] B. Rong, J. N. Burghartz, L. K. Nanver, B. Rejaei, and M. Van der Zwan, "Surface-passivated high resistivity silicon substrates for RFICs", *IEEE Electron Dev. Lett.*, vol. 25, no. 4, pp. 176–178, 2004.

[10] D. Lederer, "WideBand characterization of advanced SOI material and MOS devices for high frequency applications", Ph.D. thesis, UCL, Louvain-la-Neuve, Belgium, October 2006.

[11] D. Lederer and J.-P. Raskin, "Effective resistivity of fully-processed SOI substrates", *Solid-State Electron.*, vol. 49, no. 3, pp. 491–496, 2005.



**Khaled Ben Ali** was born in Sfax, Tunisia, in 1978. He received the Electronics Master degree from the Faculty of Sciences of Tunis, Tunisia, in 2001 and a Master degree in instrumentation and measurement from the National Institute of Applied Sciences and Technologies in 2004. From 2004 to 2008, he was assistant lecturer at the higher institute of dramatic arts and the higher institute of medical technologies of Tunis. He has been working towards his Ph.D. degree jointly at the Microwave Laboratory at the Université Catholique de Louvain (UCL), Louvain-la-Neuve, Belgium, and the Faculty of Sciences of Tunis, since 2008. His research interests are the characterization of high-resistivity Si substrates, losses and crosstalk of passive and active devices for RF applications and RF MEMS devices.
e-mail: khaled.benali@uclouvain.be
Université catholique de Louvain
Place du Levant, 3, Maxwell Building
B-1348 Louvain-la-Neuve, Belgium



**César Roda Neve** was born in Madrid, Spain, in 1975. He received the Industrial Engineer degree from the ICAI Universidad Pontificia de Comillas, Madrid, Spain, in 2000. From 2004 to 2006, he was teaching assistant and researcher engineer at the Electronics Department of the University Carlos III of Madrid, Madrid, Spain, where he worked on pulsed mode-locked lasers and laterally coupled diode lasers for ROF links and optoelec-

tronic up/down conversion. In 2006, he joined the Microwave Laboratory at the Université Catholique de Louvain (UCL), Louvain-la-Neuve, Belgium, where he is working towards his Ph.D. degree. His research interests are high-resistivity Si substrates, wide-band on-wafer characterization of advanced passive and active devices in silicon MOS technology, substrate crosstalk, on wafer temperature characterization, transmission lines losses, and interconnects.

e-mail: cesar.rodaneve@uclouvain.be
Université catholique de Louvain
Place du Levant, 3, Maxwell Building
B-1348 Louvain-la-Neuve, Belgium



**Ali Gharsallah** received the degree in radio frequency Engineering from the Higher School of Telecommunication of Tunis in 1986 and the Ph.D. degree in 1994 from the Engineering School of Tunis. Since 1991, he was with the Department of Physics at the Faculty of Sciences of Tunis. Actually, he is a Full Professor in Electrical Engineering and Director of the Engineering studied in the Higher Ministry Education of Tunisia. His current research interests include smart antennas, array signal processing, multilayered structures and microwave integrated circuits. He has about thirty five papers published in scientific journals and fifty papers conferences. Professor Gharsallah supervises more then ten thesis and thirty Masters.

e-mail: ali.gharsallah@fst.rnu.tn
Research Unit of Microwaves Electronic Circuits
and Systems
Faculty of Science of Tunis
Elmanar 2092, Tunisia



**Jean-Pierre Raskin** (M'97, SM'06) was born in Aye, Belgium, in 1971. He received the Industrial Engineer degree from the Institut Supérieur Industriel d'Arlon, Belgium, in 1993, and the M.Sc. and Ph.D. degrees in applied sciences from the Université catholique de Louvain (UCL), Louvain-la-Neuve, Belgium, in 1994 and 1997, respectively. From 1994 to 1997, he was a Research Engineer at the Microwave Laboratory, UCL, Belgium. He worked on the modeling, characterization and realization of MMIC's in Silicon-on-Insulator (SOI) technology for low-power, low-voltage applications. In 1998, he joined the EECS Department of The University of Michigan, Ann Arbor, USA. He has been involved in the development and characterization of micromachining fabrication techniques for microwave and millimeter-wave circuits and microelectromechanical transducers/amplifiers working in harsh environments. In 2000, he joined the Microwave Laboratory of UCL, Louvain-la-Neuve, Belgium, as Associate Professor. Since 2007, he has been a Full Professor and Head of the Microwave Laboratory of UCL. He was visiting professor at Newcastle University, Newcastle Upon Tyne, UK, from September 2009 until September 2010. His research interests are the modeling, wideband characterization and fabrication of advanced SOI MOSFETs as well as micro and nanofabrication of MEMS/NEMS sensors and actuators. He is an IEEE Senior Member, EuMA Associate Member and Member of the Research Center in Micro and Nanoscopic Materials and Electronic Devices of the Université catholique de Louvain. He is author or co-author of more than 350 scientific articles.

e-mail: jean-pierre.raskin@uclouvain.be
Université catholique de Louvain
Place du Levant, 3, Maxwell Building
B-1348 Louvain-la-Neuve, Belgium

Paper

# Capacity and Quality Optimization of CDMA Networks

Ayodeji James Bamisaye and Michael O. Kolawole

**Abstract—Coverage and capacity are important issues in the planning process for cellular third generation (3G) mobile networks. The planning process aims to allow the maximum number of users sending and receiving adequate signal strength in a cell. This paper describes the conceptual expressions require for network coverage and capacity optimization analysis, examines service quality issues, and presents practical solutions to problems common to sub-optimality of CDMA networks.**

***Keywords—CDMA, link pilot, network planning, quality of service.***

## 1. Introduction

Development of the third generation (3G) systems, such as code division multiple access (CDMA) – including 2000 series that utilises CDMA as an underlying channel access method-provided connectivity to packet data networks via cellular systems while increasing voice capacity. As one would expect, many of the rapidly growing internet applications and services are finding their way into the mobile wireless domain and taking advantage of the 3G system. Services such as real time streaming video and music, and on-line interactive gaming are just a few examples of services whose popularity is growing beyond expectations. Hence, services of this nature have challenged 3G networks standardization capable of providing increase data throughput. The 3G networks may also be referred to as universal mobile telecommunication system (UMTS) [1].

Spectrum is a valuable and limited resource. Therefore for an operator, cost effective improvement in capacity is always an important goal. Capacity gain, both for voice and new data services, is crucial for an operator's competitiveness. It is possible to achieve significant capacity improvements in existing networks without deploying additional carrier and base stations or drafting new standards. By following proper *radio frequency* (RF) network planning and optimization techniques; CDMA operators would see immediate benefits on their network capacity. CDMA is a digital cellular technology that uses spread-spectrum techniques [2]. It does not assign a specific frequency to each user. Rather, every channel uses the full available spectrum. Individual conversations are encoded with a pseudo-random digital sequence.

The objective of network planning is to maximize the coverage and capacity, and the quality of service. However, in the 3G planning, since all carriers in the network use the same frequency range, frequency planning is not required. Furthermore, coverage and capacity planning should be performed in tandem since capacity requirement and traffic distribution influence the coverage [3], [4].

## 2. Network Planning

There are three distinct standards operating worldwide for 3G networks namely WCDMA, CDMA2000 and TD-SCDMA, but all have similar planning process as well as overall operational objectives. This planning process can be divided into three parts:

1. Initial phase (also called system dimensioning) involves estimating traffic throughput, coverage area, and coverage threshold.
2. Digital planning phase includes meeting traffic, configuration, coverage threshold and capacity requirements.
3. Radio frequency optimization and monitoring; involving coverage verification and capacity availability.

We see that the whole process of 3G planning needs to take into account coverage and capacity planning. In a cellular system where all the air interface connections operate on the same carrier, the number of simultaneous users directly influences the receivers' noise floors [5]. Consequently, the performance of any digital modulation technique can be described in terms of the normalized ratio of energy per bit ($E_b$) to interference density ($I_o$) required to achieve the minimum desired bit error ratio (BER). The relationship between normalized value and the signal-to-noise ratio over the entire occupied bandwidth ($S/N$) needs to be established ensuring the required minimum or threshold value ($E_b/I_o$) for an acceptable network performance.

### *2.1. Derivation of CDMA Network Coverage and Capacity*

CDMA, being a spread spectrum technique, requires that the occupied bandwidth ($W$) to be much greater than the information bit rate, $R$. The ratio, ($W/R$) (also known as *processing gain*) then becomes a factor by which ($E_b/I_o$) is improved over the entire occupied bandwidth. The relationship between normalized value and the signal-to-noise ratio ($S/N$) can be expressed as

$$\frac{S}{N} = \frac{E_b}{N_o}\frac{R}{W}. \tag{1}$$

Rearranging to have

$$\frac{E_b}{N_o} = \frac{S}{N}\frac{W}{R}. \tag{2}$$

We assume the system's capacity requirements are inherently the same for uplink and downlink. Strictly, this assumption holds in voice systems when defined by the number of users, and may not be true for data systems; capacity requirements may be asymmetric and would require some subtle modification. The full bandwidth $(S/N)$ ratio is the ratio of received power of the desired signal $P_r$, to the sum of the noise over the full bandwidth $((N_oW)$ and the total interference $(I_o)$ including from other users, which can be formalized as:

$$\frac{S}{N} = \frac{P_r}{I_o + N_oW}, \tag{3}$$

where $N_o$ is the total floor noise or thermal noise.
For $k$ users per channel occupying the same spectrum, the "same cell" interference power $I_{usc}$ is

$$I_{usc} = (k-1)P_r,$$

while "adjacent cell" interference $I_{uad}$ is equivalent to the "same cell" interference but modified by *loading factor*, $\eta$. That is

$$I_{uad} = \eta I_{usc} = \eta(k-1)P_r. \tag{4a}$$

The loading factor is proportional to frequency reuse, $f_r$, which is the ratio of total interference to own cell interference; that is,

$$f_r = \frac{1}{1+\eta}. \tag{4b}$$

Adding interferences from "same cell" and "adjacent cell" and taking into consideration interference reduction achieved through transmission – muting during voice pauses and sectionalizing – we can write expression for total interference:

$$I_o = \frac{G_vG_s(k-1)}{f_r}P_r, \tag{5}$$

where $G_v$, and $G_s$ are scaling factors, i.e., voice gain and section gain respectively. In view of Eq. (5) in Eq. (3), we have

$$\frac{S}{N} = \frac{P_r}{\left(\dfrac{G_vG_s(k-1)P_r}{f_r} + N_oW\right)}. \tag{6}$$

Given a digital modulation technique, we can define $\gamma_t$ as the minimum or threshold value of $(E_b/I_o)$ required for an acceptable network performance. As a result, Eq. (2) must be greater than $\gamma_t$. So, Eq. (6) in Eq. (2) and applying threshold, the desired signal

$$P_r > \frac{\gamma_t N_oW}{\dfrac{W}{R} - \dfrac{\gamma_t G_sG_v(k-1)}{f_r}}. \tag{7}$$

This equation can be used to demonstrate the dependency of received signal on the user data rates, $R$, and the total number of active users in a cell. The terrain under which a user operates and the user's proximity to the base station affect the effective power received by user's mobile devices (i.e., *mobile station*, MS). Suppose an MS is at distance $d$ from the *base transceiver station*, BTS. The received power at the BTS, $P_{r(BTS)}$), from MS, can be expressed as

$$P_{r(BTS)} = P_{t(MS)} - PL(d) - z. \tag{8}$$

where:

- $P_{r(BTS)} = P_r$ is the power received by BTS (in dBm),
- $P_{t(MS)}$ is the transmission power of the MS (in dBm),
- the propagation loss at distance $d$ from the MS to BTS (in dB).
- $z$ is error due to shadow fading (in dB), where motion of the mobile results in variation of the received signal strength.

This presupposes that network coverage is limited by the maximum transmission power at the mobile and no blocking nor outage takes place in the cellular system (since CDMA technology can provide the enough codes for the new call to the cell in the ideal situation).
We ignore here fast shadow fading. A statistical model [6] can be used to generate correlated shadow-fading patterns in the absence of detailed propagation and landscape information, or cross-correlation function model [7], or a simple power lognormal distribution.
In the cellular situation, it is impossible to apply the free-space loss rule because of the proximity of the earth and the effects of trees, building, and hills in, or close to, the transmission path. The propagation loss $PL(d)$ is modeled on the Okumura-Hata [8] path prediction model. The basic equation for path loss (in dB) is

$$\begin{aligned} PL(d) \;=\; & 69.55 + 26.16\log f - 13.82\log h_{BTS} \\ & + \left[44.9 - 6.55\log h_{BTS}\right]\log d - a(h_{MS}) + C, \end{aligned} \tag{9}$$

where: $f$ is the transmission frequency (in MHz); $h_{BTS}$ is base transceiver station antenna height (in m); $h_{MS}$ is mobile station antenna height (in m); $C$ and $a(h_{MS})$ are correction factors whose values depend on the propagation terrain's type and height.
For medium to large cities, $C = 0$. Like all empirical models, the Okumura-Hata model can be adjusted for specific propagation terrains and then use experimentally acquired correction factors to reduce their uncertainty. We can now build a relationship among the received power, number of users, the coverage area or terrain, and other essential parameters drawing from Eqs. (7), (8) and (9):

$$P_{r(BTS)} = P_{r(MS)} - PL(d) - z = 10\log\left[\frac{\gamma_t N_oW}{\dfrac{W}{R} - \dfrac{\gamma_t G_vG_s(k-1)}{f_r}}\right]. \tag{10}$$

The equations can be used in cellular system planning to set hard limits on the maximum number of users that can

be admitted into the cell. Identifying and analyzing these relationships can achieve effective planning in designing cell capacity and coverage to matching specified data services.

### *2.2. Capacity Improvements with Optimization*

Theoretically, we can obtain the maximum capacity ($k_{\max}$, also called *pole capacity*) by equating the denomination of the logarithm argument of Eq. (10) to zero. That is,

$$\frac{W}{R} - \frac{\gamma_t G_v G_s (k-1)}{f_r} = 0 \tag{11a}$$

giving the pole capacity

$$k_{\max} = 1 + \frac{f_r \left(\frac{W}{R}\right)}{\gamma_t G_v G_s}. \tag{11b}$$

Naturally, this maximum capacity implies that receive power goes to infinity and/or coverage shrinking to zero. In reality, this is unrealistic due to signal-quality constraints in uplink and downlink directions, applicable power control mechanism, and pilot signaling.

Network optimization is an integral part of the operation and maintenance of mobile networks. It is often performed whenever there is a change in the network. Proper optimization techniques enable the network operator to fine-tune the network for maximum attainable capacities. The next subsection covers the important steps required in network optimization process to achieving improvements in network capacity and guarantee superior quality of service.

### *2.3. RF Network Planning and Optimization*

Most CDMA networks that suffer from RF capacity degradation are the result of poor RF network planning and optimization. An optimized RF environment is vital for operators seeking to maximize capacity. Figure 1 illustrates the relationship between base transceiver station (BTS Tx) power reductions and combined forward link pilot – i.e., the ratio of the average power of a channel, $E_c/I_o$ – under different forward link loading scenarios. The figure shows that for higher combined $E_c/I_o$, lower traffic channel $(E_c/I_{or})$ is required and more BTS *high power amplifier* (HPA) power is conserved to accommodate more users in the same sector. Typically, as shown in Fig. 1, a 1 dB BTS Tx power reduction for each voice channel increases RF capacity by approximately 15% for any given BTS sector. In essence, to achieve a uniform service quality is provided to all channels, transmission power of the forward traffic channel would need to be controlled by base stations according to the pilot powers measured at the respective mobile stations.

We see that in CDMA systems, coverage and capacity are heavily coupled and cannot be planned separately such as they could in the second-generation (2G) systems. The expression given by Eq. (10) describing the relationship between coverage, capacity and data rates is very useful in planning the networks. However, there are shortcomings

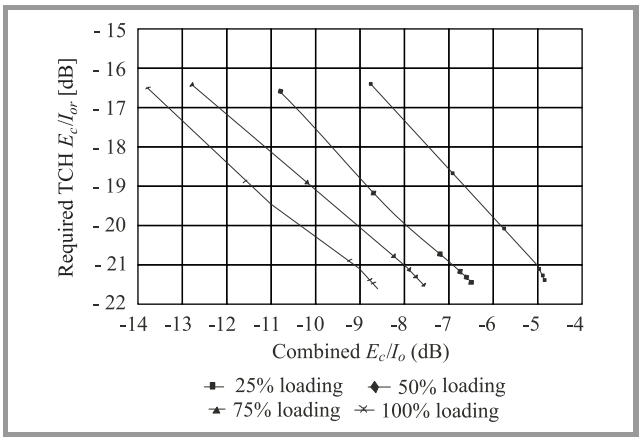



***Fig. 1.*** BTS Tx power reduction versus combined $E_c/I_o$.

in the analysis and modeling provided herein. The modeling is largely based upon an ideal situation, wherein no call will be blocked. While CDMA technology can itself provide enough codes for the mobile terminals to assure the ideal situation (enough for every user, therefore there is no blocked call) these are not used in practice. To overcome the shortcomings of the modeling presented herein, in future work we would need to consider the outage probability and dynamic loading control in the planning process and minimizing costs.

#### 2.3.1. Capacity Improvement Areas

Problems common to sub-optimality of CDMA network having RF capacity issues and solutions include:

- **Forward/reverse link imbalance** – this problem is normally caused by boomer sites with elevated antenna radiation centres. The BTS forward link covers distant areas or deep inside buildings where the reverse link of a mobile cannot reach back to the base station. In this case, excessive forward link traffic channel power is always allocated to compensate for the path loss. Link imbalance areas typically can be identified if the forward link coverage is sufficient (good pilot $E_c/I_o$) but call setup failure is high due to exhausted mobile transmit power on the reverse link. Link imbalance can be identified by drive testing into problematic areas or analyzing the network performance data for problematic clusters / sectors. RF coverage on the forward link is much larger than that on reverse link; excessive BTS Tx power is allocated for the remote user. Primarily involves making decisions on where to install new base stations and how to select their configuration, including antenna height and tilt, sector orientations, maximum emission power, pilot signal [9].
- **Excessive soft handoff area** – caused by improper cell site layout, misuse of base station antennas. This may require reduction of excessive handoff areas. In a well optimized, lightly loaded CDMA network, the typical soft handoff reduction factor is be-

tween 1.6 and 1.8; this range is reduced to 1.4 to 1.6 for a loaded network. If a higher value is seen in some areas of the network, this could indicate that those areas have softer handoff than necessary. Soft handoff increases the reliability of the radio link, but the base station requires more power to maintain the soft handoff, which reduces the forward link capacity.

- **Improper RF parameter settings** – RF parameter settings should be fine-tuned according to the traffic loading distribution in order to improve the overall network performance. RF parameter adjustments could be considered after the RF environment has been optimized and the network has reached a stable stage. There are no fixed rules on parameter changes. When voice capacity enhancement is the objective, some channel power management and power control parameter can be considered for the tuning. Nevertheless, base station's transmitter HPA power limits forward link capacity. Forward overhead channels and traffic channels share the HPA power: depends on the individual base station products.

## 3. Conclusion

Efficient planning and optimization of mobile networks is key to guaranteeing superior quality of service and user experience. This paper has developed expressions that can be used for detailed analysis of the criterion of optimization, as well as for network planning. CDMA network operators have various solutions, both short term and long term, to enhance their system capacity. Anomalies such as forward/reverse link imbalance, excessive soft handoff areas, and improper RF parameter settings could lead to under utilization of system capacity. This paper has also explained that with proper network planning and network optimization of the installed CDMA network, operators can quickly and efficiently utilize their network resources to achieve optimum system capacity.

## References

[1] H. Harri and A. Toskala, *WCDMA for UMTS radio access for third generation mobile communication.* Wiley, 2000.

[2] A. J. Viterbi, *CDMA: Principles of Spread Spectrum Communication*. Reading: Addison-Wesley, 1995.

[3] R. Hoppe, H. Buddendick, G. Wolfle, and F. M. Landstorfer, "Dynamic simulator for studying WCDMA radio network performance", in Proc. *53rd Veh. Technol. Conf. VTC 2001 Spring*, Rhodes, Greece, 2001, vol. 4, pp. 2771–2775.

[4] L. Jaana, W. Achim, and N. Tomas, *Radio Network Planning and Optimisation for UMTS*. Wiley, 2002.

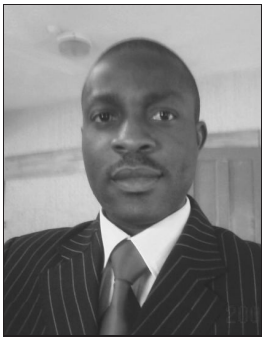
[5] A. J. Bamisaye, "Effect of multiuser interference on subscriber location in CDMA network", M. Eng. thesis, the Federal University of Technology, Akure. 2009.

[6] K. Kumaran, S. E. Golowich, and S. Borst, "Correlated shadow-fading in wireless networks and its effect on call dropping", *Wirel. Netw.*, vol. 8, no. 1, pp. 61–71, 2002.

[7] F. Graziosi and F. Santucci, "A general correlation model for shadow fading in mobile radio systems", *Commun. Lett. IEEE*, vol. 6, no. 3, pp. 102–104, 2002.

[8] M. Hata, "Empirical formula for propagation loss in land mobile radio service", *IEEE Trans. Veh. Technol.*, vol. VT-29, pp. 317–325, 1981.

[9] E. Amaldi, A. Capone, and F. Malucelli, "Radio planning and coverage optimization of 3G cellular networks", *Wirel. Netw.*, vol. 14, no.4, pp. 435–447, 2008.



**Ayodeji James Bamisaye** obtained degree in electrical engineering and Master degree (M.Eng.) in electrical and electronics engineering (communication option) from Bayero University, Kano, Nigeria and The Federal University of Technology, Akure, Ondo State, Nigeria, respectively. He is a member of several professional societies including Nigerian Society of Engineers (MNSE), and Council for the Regulation of Engineering in Nigeria (COREN). He has over 8 years experience in mobile communications Engineering, he is presently a Core Network/Switch Engineer with Starcomms Telecommunication PLC, the largest CDMA 3G network in Nigeria, and a researcher in switching optimization.
e-mail: ayobamisaye@yahoo.com
Department of Electrical and Electronics Engineering
Federal University of Technology
Akure, Nigeria



**Michael O. Kolawole** received B.Eng. and Ph.D. in electrical engineering, respectively, from the Victoria University Melbourne and University of New South Wales, Sydney. He also earned a Master of environmental studies from the University of Adelaide. He is an experienced project and research leader with knowledge across a broad range of business and technology environments. He has over 15 years of experience in satellite and radar systems design and applications, and has overseen a number of operational innovations. He has lectured in satellite communications and communications engineering at the Victoria University Melbourne Australia. He is concurrently Professor of Electrical Engineering (Communication) of the Federal University of Technology, Akure Nigeria. He is a member of several professional societies including the New York Academy of Science, and the Institution of Engineers (Australia) .
e-mail: kolawolm@yahoo.com
Department of Electrical and Electronics Engineering
Federal University of Technology
Akure, Nigeria

Paper

# Routing Misbehavior Detection in MANETs Using 2ACK

Sunilkumar S. Manvi, Lokesh B. Bhajantri, and Vittalkumar K. Vagga

**Abstract—This paper proposes routing misbehavior detection in MANETs using 2ACK scheme. Routing protocols for MANETs are designed based on the assumption that all participating nodes are fully cooperative. However, due to the open structure and scarcely available battery-based energy, node misbehavior may exist. In the existing system, there is a possibility that when a sender chooses an intermediate link to send some message to a destination, the intermediate link may pose problems such as, the intermediate node may not forward the packets to destination, it may take very long time to send packets or it may modify the contents of the packet. In MANETs, as there is no retransmission of packets once it is sent, care must be taken not to loose packets. We have analyzed and evaluated a technique, termed 2ACK scheme to detect and mitigate the effect of such routing misbehavior in MANETs environment. It is based on a simple 2-hop acknowledgment packet that is sent back by the receiver of the next-hop link. 2ACK transmission takes place for only a fraction of data packets, but not for all. Such a selective acknowledgment is intended to reduce the additional routing overhead caused by the 2ACK scheme. Our contribution in this paper is that, we have embedded some security aspects with 2ACK to check confidentiality of the message by verifying the original hash code with the hash code generated at the destination. If 2ACK is not received within the wait time or the hash code of the message is changed then the node to next hop link of sender is declared as the misbehaving link. We simulated the routing misbehavior detection using 2ACK scheme to test the operation scheme in terms of performance parameters.**

***Keywords—2ACK, MANETs, routing misbehavior, selfish node.***

## 1. Introduction

A mobile ad hoc network (MANET) is a collection of mobile nodes (hosts) which communicate with each other via wireless links either directly or relying on other nodes as routers. The operation of MANETs does not depend on pre-existing infrastructure or base stations. Network nodes in MANETs are free to move randomly. Therefore, the network topology of a MANETs may change rapidly and unpredictably. All network activities such as discovering the topology and delivering data packets have to be executed by the nodes themselves either individually or collectively. Depending on its application, the structure of a MANET may vary from a small, static network that is highly power-constrained to a large-scale, mobile, highly dynamic network.

There are two types of MANETs: closed and open [1]. In a closed MANET, all mobile nodes cooperate with each other towards a common goal, such as emergency search/rescue or military and law enforcement operations. In an open MANET, different mobile nodes with different goals share their resources in order to ensure global connectivity. However, some resources are consumed quickly as the nodes participate in the network functions. For instance, battery power is considered to be most important in a mobile environment. An individual mobile node may attempt to benefit from other nodes, but refuse to share its own resources. Such nodes are called selfish nodes or misbehaving nodes and their behavior is termed as selfishness or misbehavior. One of the major sources of energy consumption in the mobile nodes of MANETs is wireless transmission. A selfish node may refuse to forward data packets for other nodes in order to conserve its own energy [2], [3].

In MANETs, routing misbehavior can severely degrade the performance at the routing layer. Specifically, nodes may participate in the route discovery and maintenance processes but refuse to forward data packets. How do we detect such misbehavior? How to make such detection process more efficient (i.e., with less control overhead) and accurate (i.e., with low false alarm rate and missed detection rate). We analyzed the 2ACK technique [4] to detect such misbehaving nodes or links. Routes containing such nodes will be eliminated from consideration. The source node will be able to choose an appropriate route to send its data. The 2ACK scheme is a network-layer technique to detect misbehaving links and to mitigate their effects. The 2ACK scheme detects misbehavior through the use of a new type of acknowledgment packet, termed 2ACK. A 2ACK packet is assigned a fixed route of two hops (three nodes) in the opposite direction of the data traffic route. In this work, we provide security features to 2ACK, where confidentiality of the message is checked by verifying the original hash code with the hash code generated at the destination.

The rest of the paper is organized as follows. Section 2 discusses related work in this area. Section 3 describes the proposed work. Section 4 presents the simulation procedure, performance parameters and the results of the proposed work. Finally, we conclude in Section 5.

## 2. Related Work

The security problem and the misbehavior problem of wireless networks including MANET's have been studied by many researchers. Various techniques have been proposed to prevent selfishness in MANETs. Some of the related works are as follows.

The work given in [5] explains detection of malicious nodes by the destination node, isolation of malicious nodes by discarding the path and prevention data packets by using dispersion techniques.
The work given in [4] describes the performance degradation caused by selfish (misbehaving) nodes in MANETs. They have proposed and evaluated a technique, termed 2ACK, to detect and mitigate the effect of such routing misbehavior.
The work given in [6] presents cooperative, distributed intrusion detection architecture for MANETs that is intended to address some challenges. The architecture is organized as a dynamic hierarchy in which data acquisition occurs at the leaves, with intrusion detection data being incrementally aggregated, reduced, analyzed, and correlated as it flows upward towards the root.
The work given in [7] explains the problem of identification of misbehaving nodes and refusing to forward packets to a destination. They have proposed a reactive identification mechanism that does not rely on continuous overhearing or intensive acknowledgment techniques, but is only activated in the event of performance degradation.
The work given in [8] proposes a general solution to packet dropping misbehavior in mobile ad hoc networks. The solution allows monitoring, detecting, and isolating the droppers.
The work given in [9] proposes signal strength based routing for wireless ad hoc networks. It uses signal strengths on the multi hop to identify stable route from source to destination in an ad hoc networks. A stable route helps to reduce control packets overhead during route maintenance and avoids route interruptions. Some of the related work is given [10], [11], [12].

# 3. Proposed Work

The proposed system is used to detect the misbehavior routing using 2ACK and also check the confidentiality of the data message in MANETs environment. Here, we used a scheme called 2ACK scheme, where the destination node of the next hop link will send back a 2 hop acknowledgement called 2ACK to indicate that the data packet has been received successfully. The proposed work (2ACK with confidentiality) is as follows.

- If the 2ACK time is less than the wait time and the original message contents are not altered at the intermediate node then, a message is given to sender that the link is working properly.
- If the 2ACK time is more than the wait time and the original message contents are not altered at the intermediate node, then a message is given to sender that the link is misbehaving.
- If the 2ACK time is more than the wait time and the original message contents are altered at the intermediate node, then message is given to sender that the link is misbehaving and confidentiality is lost.
- If the 2ACK time is less than the wait time and the original message contents are altered at the intermediate node then, a message is given to sender that the link is working properly and confidentiality is lost.

At destination, a hash code will be generated and compared with the sender's hash code to check the confidentiality of message. Hence, if the link is misbehaving, sender to transmit messages will not use it in future and loss of packets can be avoided.
This section presents system model, and functioning scheme.

### *3.1. System Model*

In the existing system, there is a possibility that when a sender chooses an intermediate link to send some message to destination, the intermediate link may give problems such as the intermediate node may not forward the packets to destination, it may take very long time to send packets or it may modify the contents of the packet. In MANETs, as there is no retransmission of packets once it is sent, hence care is to be taken that packets are not lost.
Noting that a misbehaving node can either be the sender or the receiver of the next-hop link, we have focused on the problem of detecting misbehaving links instead of misbehaving nodes using 2ACK scheme. In the next-hop link, a misbehaving sender or a misbehaving receiver has a similar adverse effect on the data packet. It will not be forwarded further. The result is that this link will be tagged. Our approach is used to discuss the significantly simplification of the routing detection mechanism and also checking the confidentiality of the message in MANETs environment.
Figure1 shows the system model of the proposed work. The various modules in the system model are as follows.

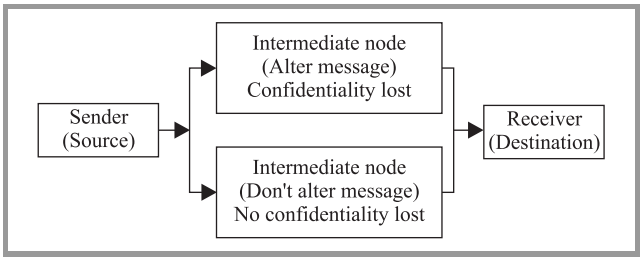



***Fig. 1.*** System model.

**Module 1: Sender module (Source node)**. The task of this module is to read the message and then divide the message into packets of 48 bytes in length, send the packet to receiver through the intermediate node and receive acknowledgement from the receiver node through the intermediate node. After sending every packet the "Cpkts" counter is incremented by 1. 2ACK time is compared with the wait time. If 2ACK is less than wait time, "Cmiss" counter is incremented by 1. The ratio of "Cmiss" to "Cpkts" is compared with the "Rmiss" (a threshold ratio). If it is less than "Rmiss", link is working properly otherwise misbehaving.

**Module 2: Intermediate module (Intermediate node)**. The task of this module is to receive packet from sender, alter/don't alter the message and send it to destination. Get 2ACK packet from the receiver and send 2ACK packet to sender.

**Module 3: Receiver module (Destination node)**. The task of this module is to receive message from the intermediate node, take out destination name and hash code and decode it. Compare the hash code of source node and destination node for security purpose. Send 2ACK to source through the intermediate node.

### *3.2. Functioning of Scheme*

#### 3.2.1. Algorithm of 2ACK Scheme

We have used the triplet of N1 → N2 → N3 as an example to illustrate 2ACK's pseudo code. Where N1 is assumed as the source node, N2 is the intermediate node and N3 is the destination node. Note that such codes run on each of the sender/receiver of the 2ACK packets.

**Nomenclature**: {**Cpkts** = the number of the message packets sent, **Cmiss** = the number of the 2ACK packets missed, **d** = the acknowledgement ratio. **WT** = waiting time, i.e., the maximum time allotted to receive 2ACK packet}

**A. At node N1**

**while** (true) **do**

- Read the destination address;
- Read the message;
- Find the length of the message.
  Cmiss=0, Cpkts=0, WT=20 ms, d=0.2,
  2ACK Time=Current Time (Acknowledgement accepted time) – Start Time.
  **while** (length > 48 bytes) **do**
  Take out 48 message packet;
  Length = length – 48;
  Encode message using hash function;
  Send message along with the hash key;
  Cpkts++ ;
  Receive 2ACK packet;
  **if** (2ACK time > WT) **then**
  Cmiss++ ;
  **end**
  **end**
  **if** (length < 48 bytes) **then**
  Encode message using hash function;
  Send message along with the hash key;
  Cpkts++;
  Receive 2ACK packet;
  **if** (2ACK time > WT) **then**
  Cmiss++;
  **end**
  **end**

**end**

**B. At node N2**

**while** (true) **do**
Read message from source N1
**if** (Alter) **then**
Add dummy bytes of characters;
Process it and forward to destination N3;
Receive 2ACK from N3 and send it to N1;
**else if** (Do not Alter) **then**
Process it and forward to destination N3;
Receive 2ACK from N3 and send it to N1;
**end**
**end**

**C. At node N3**

**while** (true) **do**
Read message from N2;
Take out destination name and hash code;
Decode the message;
Send 2ACK packet to N2;
**end**

**D. At N1 and N3 parallel**

**while** (true) **do**
**if** ((Cmiss/Cpkts)>d and (hash code of source msg) ! = (hash code of destination msg)) **then**
Link is misbehaving and the confidentiality is lost;
**end**
**if** ((Cmiss/Cpkts)<d and (hash code of source msg) ! = (hash code of destination msg)) **then**
Link is working properly and the confidentiality is lost;
**end**
**if** ((Cmiss/Cpkts)>d and (hash code of source msg) = (hash code of destination msg)) **then**
Link is misbehaving;
**end**
**if** ((Cmiss/Cpkts)<d and (hash code of source msg) = (hash code of destination msg)) **then**
Link is working properly;
**end**
**end**

## 4. Simulation

We conducted simulation of the proposed scheme by using C programming language. The proposed scheme has been simulated in various network scenarios. Simulations are carried out extensively with random number for 100 iterations. This section presents the simulation model, simulation procedure and results and discussions.

### *4.1. Simulation Model*

Our simulation model consists of $N$ number of nodes. The nodes are selected randomly in MANETs environment. The first node is always assumed as the source node and the last node is assumed as the destination node. Remaining nodes are assumed as the intermediate nodes (e.g., $N = 70$ nodes, in that first, i.e., N1 is assumed as source node and last, i.e., N70 is assumed as the destination node and N2 to N69 are assumed as the intermediate nodes). We have used some of the functions in our simulation model.

- **Pm** – the fraction of nodes that are misbehaving. The misbehaving nodes are selected among all network nodes randomly;
- **Rmiss** – the threshold to determine the allowable ratio of the total number of 2ACK packets missed to the total number of data packets sent;
- **R2ack** – the acknowledgement ratio, the fraction of data packets that are acknowledged with 2ACK packets (maintained at the 2ACK sender).

### *4.2. Simulation Procedure*

To illustrate some of the results of simulation, we have considered the following environment variables as follows: N = 10 to 90 for different cases, Pm = 0, 0.1, 0.2, 0.3, 0.4, WT = 20 ms and R2ack = 0.05, 0.2, 0.5, and 1.

**Begin**

1) Randomly generate number of nodes N.
2) Compute the acknowledgement time in the absence of misbehaving nodes.
3) Compute for the selected parameter for different values of Pm ranging from 0 to 0.4 and find the number of misbehaving nodes.
4) Wait for some delay and the compute the same parameter for different R2ack values ranging from 0.05 to 1.
5) Apply the proposed scheme.
6) Compute the performance parameters.
7) Generate the graphs.

**End**

### *4.3. Performance Parameters*

We have used the following parameters to measure the performance of the 2ACK scheme in MANET's.

- **Packet delivery ratio (PDR)** – the ratio of the number of packets received at the destination and the number of packets sent by the source.
- **Routing overhead (RO)** – the ratio of the amount of routing related transmissions (such as misbehavior report, 2ACK etc) to the amount of data transmissions. The amount is in bytes. Both forwarded and transmitted packets are counted.
- **2ACK time** – it measures the time required to receive the 2ACK packet from destination node to source node during the absence of misbehaving nodes.
- **2ACK time1** – it measures the time required to receive the 2ACK packet from destination node to source node during the presence of some misbehaving nodes.
- **Throughput** – it measures the overall performance of the 2ACK scheme with respect to the misbehaviour ratio.

### *4.4. Results and Discussion*

Figure 2 shows the packet delivery ratio versus misbehavior ratio. The packet delivery ratio (PDR) of the 2ACK scheme with different acknowledgment ratios (R2ack). The varied Pm from 0 (all of the nodes are well behaved) to 0.4 (40% of the nodes are misbehave). We have observed

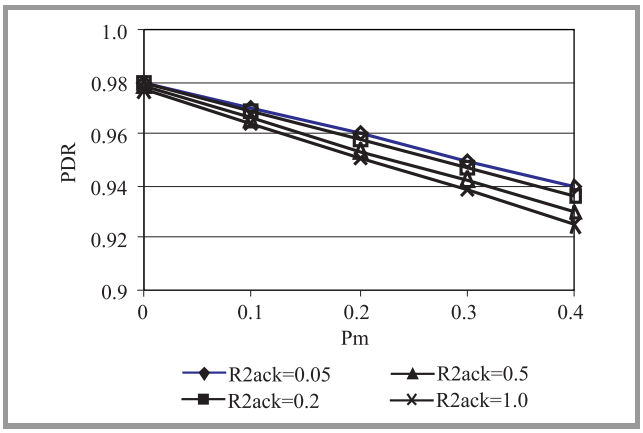



***Fig. 2.*** Packet delivery ratio (PDR) versus misbehavior ratio (Pm).

that most packets were delivered when Pm = 0 (no misbehaving nodes). The packet delivery ratio decreases as Pm increases. The 2ACK scheme delivered over 90% of the data packets even when Pm = 0.4. The acknowledgment ratio R2ack was set to 0.05, 0.2, 0.5 and 1 respectively. We can see R2ack does not appreciably affect the PDR performance of the 2ACK scheme.

Figure 3 shows the routing overhead (RO) of the 2ACK scheme with different acknowledgment ratios, R2ack. We varied Pm from 0 (all of the nodes are well behaved) to 0.4 (40% of the nodes are misbehave). Here, we compare routing overhead of the 2ACK scheme with different R2ack values. Overhead of the 2ACK scheme is highest when R2ack = 1. This is due to the large number of the 2ACK packets transmitted in the network. As the value

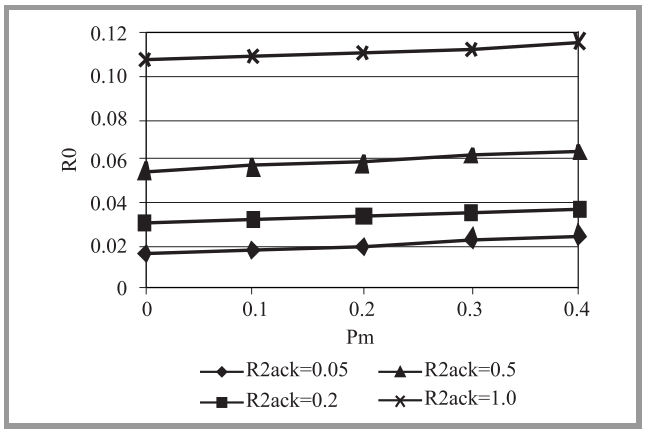



***Fig. 3.*** Routing overhead (RO) versus misbehavior ratio (Pm).

of R2ack decreases, the routing overhead reduces dramatically. Therefore, R2ack in the 2ACK scheme provides an effective "knob" to tune the routing overhead.

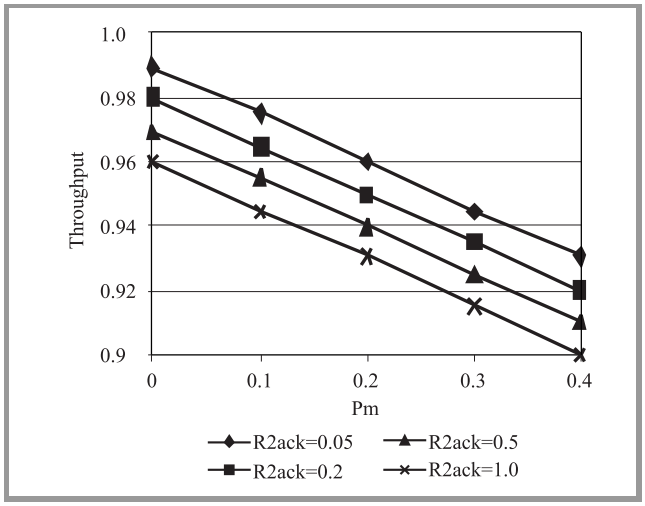



***Fig. 4.*** Throughput versus misbehavior ratio (Pm).

Figure 4 shows the relative throughput of the 2ACK scheme with different acknowledgment ratios, R2ack. We varied Pm from 0 (all of the nodes are well behaved) to 0.4 (40% of the nodes are misbehave). Here, we compare throughput of the 2ACK scheme with different R2ack values as well as with the different misbehavior ratios values. Throughput will be high when the misbehavior ratio is 0 (no misbehaving nodes) and R2ack is 0.05 (5 2ACK has to be sent for every 100 packets). The throughput decreases as Pm increases or R2ack increases. For instance, when Pm = 0.4 and R2ack = 1, the 2ACK scheme is able to support a relative throughput of 90%.

Figure 5 shows the number of the nodes increases, the 2ACK time will also increases in MANET environment. The number of nodes are randomly selected and wait time is set for 20 ms. The time is calculated for the expected 2ACK packet. If received within 20 ms, it is called a successful 2ACK. If not it called as lost 2ACK.

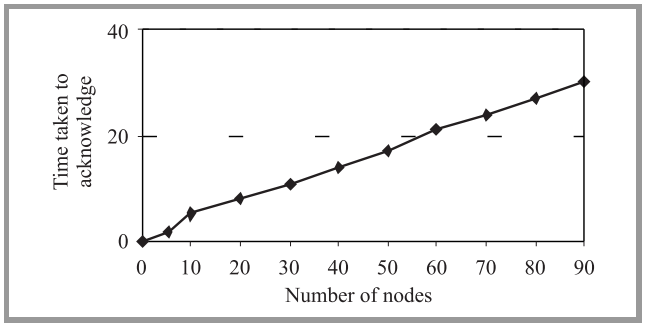



***Fig. 5.*** Number of nodes versus time taken to acknowledge.

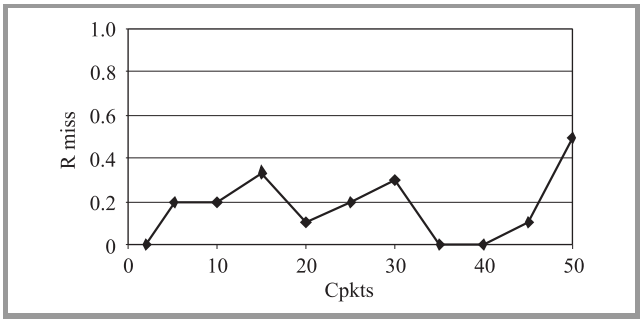



***Fig. 6.*** 2ACK miss ratio (Rmiss) versus number of packets sent.

Figure 6 shows the graph of 2ACK miss ratio (Rmiss) versus number of packets sent (Cpkts). Cmiss depends upon the 2ACK time which varies on the number of misbehaving nodes. Hence, the graph varies drastically.

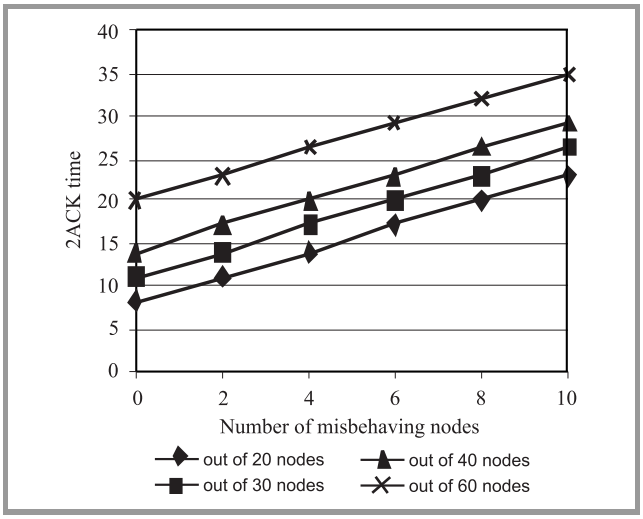



***Fig. 7.*** Number of misbehaving nodes versus 2ACK time.

Figure 7 shows the graph of 2ACK time with respect to the number of misbehaving nodes. As the number of misbehaving nodes increases, the time taken to receive the 2ACK packet will also increases gradually.

## 5. Conclusion

Mobile ad hoc networks have been an area for active research over the past few years, due to their potentially

widespread application in military and civilian communications. Such a network is highly dependent on the cooperation of all its members to perform networking functions. This makes it highly vulnerable to selfish nodes or misbehavior nodes. When such misbehaving nodes participate in the route discovery phase but refuse to forward the data packets, routing performance may be degraded severely.

In this paper, we have investigated the performance degradation caused by such selfish (misbehaving) nodes in MANETs. We have analyzed and evaluated a technique, termed 2ACK, to detect and mitigate the effect of such routing misbehavior. Extensive analysis of the 2ACK scheme has been performed to evaluate its performance. We have embedded some security aspects with 2ACK to check confidentiality of the message by verifying the original hash code with the hash code generated at the destination. Our simulation results show that the 2ACK scheme maintains up to 91% packet delivery ratio even when there are 40% misbehaving nodes in the MANETs that we have studied. The regular DSR scheme can only offer a packet delivery ratio of 40%. The false alarm rate and routing overhead of the 2ACK scheme are investigated as well. One advantage of the 2ACK scheme is its flexibility to control overhead with the use of the R2ack parameter.

## References


[1] E. Lorenzini, "Cooperation", in *Proc. Sust. Coop. Multi-Hop Wirel. Netw.*, 2007 [Online]. Available: http://www.research.microsoft.com/enus/um/people/ratul/.../nsdi2005-catch.pdf

[2] G. F. Mariasy, P. Georgiadis, D. Flitzanis, and K. Mandalas, "Cooperation enforcement schemes for MANETs: a survey: research articles", in *Wirel. Commun. Mobile Comput.*, vol. 6, iss. 3, pp. 319–332, 2006.

[3] L. Tamilselvan and V. Sankaranarayanan, "Prevention of cooperative black hole attack in MANET", *J. Netw.*, vol. 3, no. 5, pp. 13–20, 2008.

[4] K. Liu, J. Deng, P. K. Varshney, and K. Balakrishnan, "An acknowledgment-based approach for the detection of routing misbehavior in MANETs", *IEEE Trans. Mob. Comput.*, vol. 6, no. 5, pp. 536–550, 2007.

[5] S. Dhanalakshmi and, M. Rajaram, "A reliable and secure framework for detection and isolation of malicious nodes in MANET", *Int. J. Comp. Sci. Netw. Secur.*, vol. 8, no. 10, pp. 184–190, 2008.

[6] D. Sterne, P. Balasubramanyam, D. Carman, B. Wilson, R. Talpade, C. Ko, R. Balupari, C.-Y. Tseng, T. Bowen, K. Levitt, and J. Rowe, "A general cooperative intrusion detection architecture for MANETs", in *Proc. 3rd IEEE Int. Inform. Assur. Worksh.*, College Park, USA, 2005, pp. 57–70.

[7] W. Kozma Jr. and L. Lazos, "Reactive identification of misbehaviour in ad hoc networks based on random audits", 2008 [Online]. Available: http://www.ieeexplore.ieee.org/iel5/4557722/4557723/04557810.pdf

[8] D. Djenouri, O. Mahmoudi, M. Bouamama, D. Llewellyn-Jones, and M. Merabti, "On securing MANET routing protocol against control packet dropping" [Online]. Available: http://www.cscjournals.org/Journals/IJCSS/Volume2/Issue1/IJCSS-24.pdf

[9] S. R. Biradar, K. Sharma, S. K. Sarkar, and Puttamadappa C., "Signal strengths based stable route for wireless ad-hoc networks" in *Proc. Int. Worksh. Conf. Stat. Phys. Appr. Multi-Discip. Probl.*, Guwahati, India, 2008 [Online]. Available: http://www.iitg.ac.in/statphys/files/abs_cn_05.pdf

[10] K.-W. Chin, J. Judge, A. Williams, and R. Kermode, "Implementation experience with MANET routing protocols", *ACM SIGCOMM Comp. Commun. Rev.*, vol. 32, no. 5, pp. 49–59, 2002.

[11] S. Meguerdichian, F. Koushanfar, M. Potkonjak, and M. B. Srivastava, "Coverage problems in wireless ad-hoc sensor networks" [Online]. Available: http://www.ece.rice.edu/∼fk1/papers/Infocom_coverage_01.pdf

[12] K. Singh, A. Nedos, and S. Clarke, "Distributed computing and networking", in *Proc. 8th Int. Conf. ICDCN 2006*, Guwahati, India, 2006, Berlin Heidelberg: Springer [Online]. Available: http://www.springerlink.com/index/82r23149410vr703.pdf





**Sunilkumar S. Manvi** received M.E., and Ph.D. from the University of Visweshwariah College of Engineering (UVCE), Bangalore, India, and Indian Institute of Science (IISc.), Bangalore, India, respectively. He is currently working as a Professor and Head of Department of Electronics and Communication Engineering, REVA Institute of Technology and Management, Bangalore, India. He has experience of around 22 years in teaching and research. He is involved in research of agent based applications in multimedia communications, grid computing, mobile ad-hoc networks, sensor networks, e-commerce, vehicle networks and mobile computing. He has published 4 books, 4 book chapters, 40 refereed journal papers, and about 110 refereed conference papers. He has given many invited lectures and has conducted several workshops/seminars/conferences. He is reviewer for many journals/conferences including IEEE and Elsevier Publications. He is a member of IEEE, Fellow of IETE (India) and Fellow of IE (India).
e-mail: agentsun2002@yahoo.com
Department of Electronics and Communication Engineering
Wireless Information System Research Lab
REVA Institute of Technology and Management
Bangalore, Karnataka, India





**Lokesh B. Bhajantri** received M.Tech. degree in computer science and Engg. from Basaveshwar Engineering College, Bagalkot, India, in 2005. He is presently working as a lecturer in the Department of Information Science and Engineering, Basaveshwar Engineering College, Bagalkot, India. He has experience of around 6 years in

teaching and research. His areas of interest include e-commerce, u-commerce, mobile computing and communication, networking protocols, distributed sensor networks, genetic algorithms, applications of mobile agents and real time systems. He has given invited lectures in AICTE sponsored workshops/seminars. He has published chapter in *Handbook Research on Telecommunications Planning and Management for Business*, 5 referred international/national conferences papers. He is member of Board of Studies (BoS) in Department of Information Science and Engineering, BEC (Autonomous), Bagalkot, Karnataka, India.
e-mail: lokeshcse@yahoo.co.in
Department of Information Science
and Engineering
Basveshwar Engineering College
Bagalkot, Karnataka, India



**Vittalkumar K. Vagga** received B.E. degree in Electronics & Communication, 2005 and M.Tech. degree in computer science and Engg., 2009, from Vishveswariah Technological University (VTU), Belguam, Karnataka, India. He is currently working as a lecturer in the Department of Electronics and Communication, Government Polytechnic. His areas of interest include embedded systems, networking protocols, ad-hoc networks and real time systems. He is involved in conducting CHSSC classes in Infosys Campus Connect Program as a faculty.
e-mail: vkvagga@gmail.com
Department of Electronics and Communication
Government Polytechnic
Gadag-Betgeri, Karnataka, India

# *Information for Authors*

***Journal of Telecommunications and Information Technology*** **(JTIT)** is published quarterly. It comprises original contributions, dealing with a wide range of topics related to telecommunications and information technology. **All papers are subject to peer review**. Topics presented in the JTIT report primary and/or experimental research results, which advance the base of scientific and technological knowledge about telecommunications and information technology.

JTIT is dedicated to publishing research results which advance the level of current research or add to the understanding of problems related to modulation and signal design, wireless communications, optical communications and photonic systems, voice communications devices, image and signal processing, transmission systems, network architecture, coding and communication theory, as well as information technology.

Suitable research-related papers should hold the potential to advance the technological base of telecommunications and information technology. Tutorial and review papers are published only by invitation.

**Manuscript.** TEX and LATEX are preferable, standard Microsoft Word format (.doc) is acceptable. The author's JTIT LATEX style file is available:
http://www.itl.waw.pl/publ/jtit/doc/jtitaut.zip

Papers published should contain up to 10 printed pages in LATEX author's style (Word processor one printed page corresponds approximately to 6000 characters).

The manuscript should include an abstract about 150–200 words long and the relevant keywords. The abstract should contain statement of the problem, assumptions and methodology, results and conclusion or discussion on the importance of the results. Abstracts must not include mathematical expressions or bibliographic references.

Keywords should not repeat the title of the manuscript. About four keywords or phrases in alphabetical order should be used, separated by commas.

The original files accompanied with pdf file should be submitted by e-mail: redakcja@itl.waw.pl

**Figures, tables and photographs.** Original figures should be submitted. Drawings in Corel Draw and PostScript formats are preferred. Figure captions should be placed below the figures and can not be included as a part of the figure. Each figure should be submitted as a separated graphic file, in .cdr, .eps, .ps, .png or .tif format. Tables and figures should be numbered consecutively with Arabic numerals.

Each photograph with minimum 300 dpi resolution should be delivered in electronic formats (TIFF, JPG or PNG) as a separated file.

**References.** All references should be marked in the text by Arabic numerals in square brackets and listed at the end of the paper in order of their appearance in the text, including exclusively publications cited inside. Samples of correct formats for various types of references are presented below:


[1] Y. Namihira, "Relationship between nonlinear effective area and mode field diameter for dispersion shifted fibres", *Electron. Lett*., vol. 30, no. 3, pp. 262–264, 1994.

[2] C. Kittel, *Introduction to Solid State Physics*. New York: Wiley, 1986.

[3] S. Demri and E. Orłowska, "Informational representability: Abstract models versus concrete models", in *Fuzzy Sets, Logics and Knowledge-Based Reasoning*, D. Dubois and H. Prade, Eds. Dordrecht: Kluwer, 1999, pp. 301–314.


**Biographies and photographs of authors.** A brief professional author's biography of up to 200 words and a photo of each author should be included with the manuscript.

**Galley proofs.** Authors should return proofs as a list of corrections as soon as possible. In other cases, the article will be proof-read against manuscript by the editor and printed without the author's corrections. Remarks to the errata should be provided within one week after receiving the offprint.

**Copyright.** Manuscript submitted to JTIT should not be published or simultaneously submitted for publication elsewhere. By submitting a manuscript, the author(s) agree to automatically transfer the copyright for their article to the publisher, if and when the article is accepted for publication. The copyright comprises the exclusive rights to reproduce and distribute the article, including reprints and all translation rights. No part of the present JTIT should not be reproduced in any form nor transmitted or translated into a machine language without prior written consent of the publisher.
For copyright form see:
http://www.itl.waw.pl/publ/jtit/wskazowki.html

A copy of the JTIT is provided to each author of paper published.

---


*Journal of Telecommunications and Information Technology* has entered into an electronic licencing relationship with EBSCO Publishing, the world's most prolific aggregator of full text journals, magazines and other sources. The text of *Journal of Telecommunications and Information Technology* can be found on EBSCO Publishing's databases. For more information on EBSCO Publishing, please visit www.epnet.com.

(*Contents Continued from Front Cover*)

**Probabilistic Issues in Biometric Template Design**

*A. Pacut* — *Paper* — *75*

**The Template Selection in Biometric Systems Based on Binary Iris Codes**

*M. Chochowski* — *Paper* — *86*

**Impact of Crosstalk into High Resistivity Silicon Substrate on the RF Performance of SOI MOSFET**

*K. B. Ali, C. R. Neve, A. Gharsallah, and J.-P. Raskin* — *Paper* — *93*

**Capacity and Quality Optimization of CDMA Networks**

*A. J. Bamisaye and M. O. Kolawole* — *Paper* — *101*

**Routing Misbehavior Detection in MANETs Using 2ACK**

*S. S. Manvi, L. B. Bhajantri, and V. K. Vagga* — *Paper* — *105*

INSTYTUT ŁĄCZNOŚCI
PAŃSTWOWY INSTYTUT BADAWCZY

Editorial Office

National Institute
of Telecommunications
Szachowa st 1
04-894 Warsaw, Poland

tel. +48 22 512 81 83
fax: +48 22 512 84 00
e-mail: redakcja@itl.waw.pl
http://www.nit.eu

JOURNAL OF TELECOMMUNICATIONS AND INFORMATION TECHNOLOGY 4/2010