Expert Systems With Applications 115 (2019) 673–683

Contents lists available at ScienceDirect

Expert Systems With Applications

journal homepage: www.elsevier.com/locate/eswa

An ensemble learning approach to lip-based biometric verification, with a dynamic selection of classifiers

Piotr Porwik, Rafal Doroz*, Krzysztof Wrobel

University of Silesia, Institute of Computer Science, ul. Bedzinska 39, Sosnowiec 41-200, Poland

ARTICLE INFO

Article history:
Received 6 February 2018
Revised 27 July 2018
Accepted 22 August 2018
Available online 23 August 2018

Keywords:
Lip-based biometrics
Dynamic classifiers selection
Pattern recognition
Ensemble classification
Person verification

ABSTRACT

Machine learning approaches are largely focused on pattern or object classification, where a combination of several classifier systems can be integrated to help generate an optimal or suboptimal classification decision. Multiple classification systems have been extensively developed because a committee of classifiers, also known as an ensemble, can outperform the ensemble's individual members. In this paper, a classification method based on an ensemble of binary classifiers is proposed. Our strategy consists of two phases: (1) the competence of the base heterogeneous classifiers in a pool is determined, and (2) an ensemble is formed by combining those base classifiers with the greatest competences for the given input data.

We have shown that the competence of the base classifiers can be successfully calculated even if the number of their learning examples was limited. Such a situation is particularly observed with biometric data. In this paper, we propose a new biometric data structure, the *Sim* coefficients, along with an efficient data processing technique involving a pool of competent classifiers chosen by dynamic selection.

1. Introduction

In the last decade we have observed a focus on research into the use of hybrid intelligent systems for pattern classification. For such systems, various computational-intelligence techniques have been proposed to solve realistically complex problems as seen in the fields of medical diagnosis and biomedical technology (Koprowski & Bocklitz, 2016), image analysis, biometrics (Cpałka, Zalasiński, & Rutkowski, 2016; Doroz, Porwik, & Orczyk, 2016; Ghoualmi, Draa, & Chikhi, 2016; Khan, Tahir, Khelifi, Bouridane, & Almotaeryi, 2017; Raman, Sa, Majhi, & Bakshi, 2017; Tadeusiewicz & Horzyk, 2014; Tirumala, Shahamiri, Garhwal, & Wang, 2017), banking, data analytics and many others (Cruz, Sabourin, & Cavalcanti, 2018; Kowalski, Naruniec, & Trzcinski, 2017; Porwik, Doroz, & Orczyk, 2016). Hybridization techniques have mostly been inspired by the human behavioral system because it integrates information from different parts of the body before finally coordinating the activity of all of the body's parts. In the technical sciences, this approach can be analogously treated as a hybrid composition of many diverse computational units which together help to form an ultimate decision. Multiple classifier systems (often known as an ensemble or as a committee of classifiers) are convenient tools that help to decompose complex problems into subproblems that are more easily solved. From a practical point of view, a committee of classifiers can use a combination of both homogeneous and heterogeneous classifier models in producing a final decision. A classifier ensemble will only work satisfactorily if the individual classifiers in the ensemble are each of a high quality, and if the preparation of the classification model depends on the structure of the ensemble. Heterogeneous ensemble classifiers are easier to model: we only need to train each of the ensemble's individual classifiers. So, this task is similar to training a single classifier. A homogeneous approach requires a more advanced training procedure, where the feature space is partitioned or the dataset is tailored to obtain an initial diversity of the ensemble's members.

The main task of pattern recognition is to assign a given feature to one of a set of predefined categories. Given such an assumption, the pattern recognition algorithm Υ maps the feature space X to the set of class labels C:

$$\Upsilon : X \mapsto C. \tag{1}$$

In practice, the algorithm Υ is obtained via computer programs called "classifiers". The mapping (1) is predominantly established on the basis of samples from a training set containing learning examples: observations of features together with their correct classifications as given by experts.

* Corresponding author.
E-mail addresses: piotr.porwik@us.edu.pl (P. Porwik), rafal.doroz@us.edu.pl (R. Doroz), krzysztof.wrobel@us.edu.pl (K. Wrobel).

https://doi.org/10.1016/j.eswa.2018.08.037
0957-4174/

Let the recognized subject $\mathbf{x} \in X$ be described by the vector of L features, then $\mathbf{x} \mapsto \mathbf{F} = [f^{(1)}, \ldots, f^{(L)}]$. For the case of two class labels $\{c_1, c_2\} \in C$ as often appear in biometrics – either legitimate or non-legitimate users – the formula (1) can be expressed as follows:

$$\Upsilon : \mathbf{x} \mapsto \mathbf{F} \mapsto \{c_1, c_2\}, \tag{2}$$

then the classifier Υ makes a decision using the conditional rule:

$$\Upsilon(\mathbf{x}) = \begin{cases} c_1 & \text{if } \lambda(\mathbf{x}, c_1) \geq \lambda(\mathbf{x}, c_2) \\ c_2 & \text{otherwise} \end{cases}, \tag{3}$$

where $\lambda(\mathbf{x}, c_1)$ and $\lambda(\mathbf{x}, c_2)$ are the support functions of the classifier for a given subject $\mathbf{x}$, characterized by a collection of the ordered features $\mathbf{F}$. The support functions can be different and can depend on the nature of the classifier used and on the type of its input data. Classifier's support function for a given class c_i, $i = 1, 2$ is computed as probability of correct classification in a stochastic manner: $\lambda(\mathbf{x}, c_i) = Pr(\Upsilon(\mathbf{x}) \mapsto c_i)$, $\mathbf{x} \in X$, $\lambda(\mathbf{x}, c_i) \in [0, 1]$ and $\lambda(\mathbf{x}, c_1) + \lambda(\mathbf{x}, c_2) = 1$.

In the last decade, significant development in committee of classifiers algorithms, also called multiple classifier systems, can be observed (Krysmann & Kurzynski, 2013; Kuncheva, 2004; Woloszynski & Kurzynski, 2009; Zhang, Luo, García, Tang, & Herrera, 2017). Of the different strategies, the most popular has been the construction of an ensemble, as widely proposed in biometrics, medicine, other biomedical approaches, and for time series analyses (Ahmed, Rasool, Afzal, & Siddiqi, 2017; Cruz et al., 2018; Doroz et al., 2016; Fierrez, Morales, Vera-Rodriguez, & Camacho, 2018; Koprowski & Bocklitz, 2016; Kowalski et al., 2017; Porwik, Doroz, & Orczyk, 2014; Zhang et al., 2017). Combining classifiers has been proposed as a way to improve the performance of individual classifiers. Such a strategy utilizes an ability to tackle complex tasks by decomposing various properties of each base classifier. It has also been proven that the accuracy of a pool of classifiers often exceeds the accuracy achieved by a single classifier (Britto, Sabourin, & Oliveira, 2014; Cruz et al., 2018; Kuncheva, 2004).

Assume that we manage the set of classifiers $\Upsilon_1, \ldots, \Upsilon_D$ in the ensemble E, and that D is odd. If we analyze a two-class problem, the classification process for the ensemble E can be expressed in a compact form as follows:

$$E(\mathbf{x}) = \begin{cases} c_1 & \text{if } \sum_{d=1}^{D} I(\lambda_d(\mathbf{x}, c_1) \geq \lambda_d(\mathbf{x}, c_2)) > \lceil \frac{D}{2} \rceil \\ c_2 & \text{otherwise} \end{cases}, \tag{4}$$

where $\lambda_d(\mathbf{x}, c)$ stands for the support function value of the dth classifier Υ_d in the ensemble E for a given subject $\mathbf{x}$ belonging to the class c_1, or to c_2, respectively. $I(con)$ is an indicator function that produces values from the set {0, 1}. The function $I(con)$ returns the value 1 if the condition con is fulfilled, and 0 otherwise. The advantage of the formula (4) is its ability to be used for any type of classifier.

In the strategy proposed in this paper, instead of a single classifier we employ a multiple classifier system. The multi-classifier technique supplies advantages which can not be achieved when only a single classifier is used in the classification process. Additionally, the use of an ensemble has gained wide acceptance in the machine learning and statistics communities due to the significant improvement in accuracy it delivers (Doroz et al., 2016; Kuncheva, 2004; Woźniak, Graña, & Corchado, 2014). The advantages of an ensemble approach can be briefly listed:

- An ensemble can work in parallel and on a distributed computer architecture. Such a solution results in partial solutions that are always independently solved with only the final decision being formed as a combination of the networked nature of the solution. This strategy reduces computation time, especially when dealing with big data or in a classifier's learning mode when the final classification model is not yet known.
- Machine learning strategies are often heuristic search algorithms with predefined but modeled evaluation functions (Woźniak et al., 2014). Thus, these algorithms could work with different initial points. For each initial point, these machine learning strategies can be taken to be a multi-local search which should increase the likelihood of achieving an optimal solution. A classifier trained starting from various initial conditions can compute a locally optimal solution, which can then be expanded to a global domain.
- Ensembles behave well across the range of data availability, from when we have only scarce data samples for learning, to situations in which we have a huge volume of samples at our disposal. In the latter such cases, we can utilize additional applied methods, train the classifier by means of partitioned datasets, or instead of the original data introduce other data substitutions. This last technique will be employed in our approach, as will be described later.
- Each single classifier has a specific, often dedicated, competence domain, and it is not possible to design a single classifier which outperforms all others in every classification task (Britto et al., 2014; Woźniak et al., 2014; Zhang et al., 2017). In complex recognition tasks, the optimization function can not be modeled by one single classifier. However, an increase in the number of representative functions may result in the dataset domain being properly covered.

2. The new type of training datasets for ensembles

In biometric verification mode there are mostly just two types of data: legitimated and non-legitimated. This is the reason why, in such domains, the most popular approach is two-class classifications with two class labels c_1 and c_2. It should be noted that in bioinformatics also multi-class classification is often used. In the most convenient case, all example classes are balanced or are roughly equal. This means that the same number of examples appear in each class. Unfortunately, in some cases it is difficult to collect such data. For example, in verification mode it is no problem to gather a given person's genuine signature, whereas in most cases we have a problem with forged signatures, or their collection is simply impossible. The second problem that appears in biometrics is unstable data which can change over some range. An example might be in the domain of signatures or voice analysis where, to generate a highly efficient classifier, several copies of a given object have to be collected. This is uncomfortable for users and very difficult for forged objects.

In our new method of input sample interpretation, instead of raw data in the form of a vector of features $\mathbf{F}$, a new kind of similarity coefficients, *Sim*, is introduced. Through utilizing the fact that the same feature $f \in \mathbf{F}$ occurs in two instances O_i and O_j, the similarity of f in these instances can be computed.

In our case, the similarity *Sim* of the feature f_l, $l = 1, \ldots, L$ in the instance O_i and O_j, will be expressed in the space $\mathbb{R}^L$. The similarities *Sim* will be measured by means of the formula:

$$Sim(O_i, O_j)^{f_l} = |f_l^i - f_l^j|, \tag{5}$$

where f_l^i and f_l^j represents the f_l feature in the instances O_i and O_j, respectively.

The training phase provides a machine learning algorithm to train the classifier (Kuncheva, 2004) using a subject-independent learning strategy (Zois, Alewijnse, & Economou, 2016), in which the classifier is developed on training datasets containing similarities between all pairs of legitimated instances of a given subject, and between non-legitimated instances of others subjects. As with the non-legitimated instances, randomly selected legitimated instances

retrieved from other subjects were selected. In verification mode, the user presents him or herself by means of an unknown mouth image and another ID (say his/her name). Classifier compares instances of a given subject from a database and on the basis of *Sim* coefficients confirms or not, whether such subject is authorized.

Suppose that we have N instances of samples from the legitimated subject O. These instances were captured at different times and from different environments and form the set $\pi_1 = \{O_1, \ldots, O_N\}$. Let the matrix $\mathbf{X}$ contain values of the similarity coefficients *Sim*, calculated between all pairs of instances from the set π_1 that belongs to the same subject O. Let $(O_i, O_j)^{f_l}$ denote a pair of instances, O_i and O_j, of a given legitimated subject. In future, the similarity of these instances will be compared by means of the feature f_l. If an instance of the subject O is described by L features, then a global matrix $\mathbf{X} = \{Sim(O_i, O_j)\}, i = 1, \ldots, N-1, j = 2, \ldots, N$ of similarities between the subject's instances can be constructed:

$$\mathbf{X} = \begin{bmatrix} Sim(O_1, O_2)^{f_1} & \cdots & Sim(O_1, O_2)^{f_L} \\ Sim(O_1, O_3)^{f_1} & \cdots & Sim(O_1, O_3)^{f_L} \\ \vdots & \ddots & \vdots \\ Sim(O_{N-1}, O_N)^{f_1} & \cdots & Sim(O_{N-1}, O_N)^{f_L} \end{bmatrix}_{\binom{N}{2} \times L}, \tag{6}$$

where $Sim(O_i, O_j)^{f_l}$ is the similarity coefficient of the feature f_l, measured for the two instances O_i, $O_j \in \pi_1$ of a given subject.

Let the set $\pi_2 = \{O_1^{\Delta}, \ldots, O_M^{\Delta}\}$ comprise M randomly selected non-legitimate instances. Then the matrix $\mathbf{Y} = \{Sim(O_i, O_j^{\Delta})\}, i = 1, \ldots, N, j = 1, \ldots, M$ will contain values of the similarity coefficients *Sim* calculated between some pairs of instances from the set π_1 and π_2, so that the matrix $\mathbf{Y}$ will include only the similarities *Sim* between the original instances of a given subject and instances generated by other subjects, as recorded in the same database:

$$\mathbf{Y} = \begin{bmatrix} Sim(O_1, {O_1}^{\Delta})^{f_1} & \cdots & Sim(O_1, {O_1}^{\Delta})^{f_L} \\ Sim(O_1, {O_2}^{\Delta})^{f_1} & \cdots & Sim(O_1, {O_2}^{\Delta})^{f_L} \\ \vdots & \ddots & \vdots \\ Sim(O_N, {O_M}^{\Delta})^{f_1} & \cdots & Sim(O_N, {O_M}^{\Delta})^{f_L} \end{bmatrix}_{M \cdot N \times L}, \tag{7}$$

where $O_i \in \pi_1$, $O_j^{\Delta} \in \pi_2$ are the ith legitimate instance from a given subject and the jth non-legitimate instance from another subject member of the set π_2. To avoid imbalanced data (Lee & Kim, 2018), the number of elements in the matrix $\mathbf{X}$ should be close to the number of elements in the matrix $\mathbf{Y}$. This assumption is satisfied for $M = \lceil (N-1)/2 \rceil$. The idea, as presented above, of the formation of *Sim* coefficients is shown in Fig. 1.

Suppose (see Fig. 1) that we have $N = 5$ legitimated instances of a given subject and that each instance is characterized by two features, say f_1 and f_2. In the first stage, the similarities (*Sim*) between all original (legitimated) instances of a given subject, as measured on the basis of these features, are calculated. Thus we obtain $L \cdot \binom{5}{2} = 2 \cdot 10 = 20$ training instances from a legitimated subject. These are marked by blue circles. Each circle designates two similarities of an subject's instances, $(O_i, O_j)^{f_1}$ and $(O_i, O_j)^{f_2}$, respectively. In the second stage are calculated the similarities between the $N = 5$ selected original instances supplied by a given user and the $M = 2$ randomly selected instances supplied by other (non-legitimated) users $(O_i, O_j^{\Delta})^{f_1}$ and $(O_i, O_j^{\Delta})^{f_2}$. These are marked by red triangles. This means that we also obtain 20 non-legitimated instances.

It should be noted that for the *Sim* coefficients, in the training phase we have a total of 40 such coefficients belonging to the classes c_1 (matrix $\mathbf{X}$) and c_2 (matrix $\mathbf{Y}$), whereas for the raw data in the form of the vectors $\mathbf{F} = \{f_1, f_2\}$ there are only 20 instances. The matrices $\mathbf{X}$ and $\mathbf{Y}$ form matrix $\mathbf{\Omega} = [\mathbf{X}; \mathbf{Y}]$, that will be used in the classification phase.

3. Committee of classifiers

The goal of the ensemble generation step is to create a pool $E = \{\Upsilon_1, \ldots, \Upsilon_D\}$ of base classifiers that are both accurate and diverse. We can observe in the literature many paradigms covering the generation of a diverse pool of classifiers (Britto et al., 2014; Cruz et al., 2018; Kuncheva, 2004). An interesting review of these strategies, with comments, can be found in (Cruz et al., 2018). In our approach, we prefer to select classifiers based on different classifier models. For example: k-Nearest Neighbor (kNN), J48, Bayes NET, Random Forest, and many others. Some, for example Probabilistic Neural Networks (PNNs) and Particle Swarm Optimization (PSO) techniques, have been successfully employed in biometrics (Porwik et al., 2016; Rodrigues, Silva, Papa, Marana, & Yang, 2016). For a dynamic selection of base classifiers in the ensemble, a pool of classifiers has to be determined. We propose using eight popular, algorithmically advanced classifiers: PNN+PSO, Bayes NET, Hoeffding Tree, k-NN, J48, Random Forest, Random Tree, RIDOR, Naive Bayes, and kStar. These mentioned base classifiers give stable results over a wide range of input data (Cruz et al., 2018; Kuncheva, 2004). This range assures an initial diversity and complementarity, which leads to a better performance of the ensemble. Additionally, a number of these individual classifiers can be easily applied in a distributed environment, leading to a decreased time taken for the classification. Nevertheless, it is difficult to find a universal tool for handling classification problems, and research into classification efficiency is ongoing (Fierrez et al., 2018; Krysmann & Kurzynski, 2013; Porwik et al., 2016).

4. Measuring classifier competence

Perhaps the most important aspect of classifier selection is the need to supply a diverse, accurate pool of classifiers. It follows that, at the same time, more than one model may have desirable properties for a given classification problem (Krysmann & Kurzynski, 2013; Kuncheva, 2004). It is for this reason that we introduce dynamic classifier selection from a pool of popular classifiers.

Nowadays, selection of dynamic classifiers or selection of their dynamic features is seen in addressing many practical problems, including in biometrics, banking, image recognition and many other domains (Cruz et al., 2018; Doroz et al., 2016; Woźniak et al., 2014). In practice, we can distinguish two strategies for forming ensembles of classifiers: homogeneous and heterogeneous. Heterogeneous classifiers can be trained by different algorithms, so they potentially offer a more flexible data description with class-independent base classifiers. To properly accommodate the collected observations, a selection of the appropriate classifiers seems to be a promising strategy.

It is well known that classification can be performed on the basis of a single classifier, a classifier ensemble, or by using an ensemble with additional competence parameters. In this paper, these factors will be independently investigated via practical experiments. For each base classifier, the classifier's competence will be separately calculated, following which each classifier's competence will be used to decide which classifier from the ensemble E is the best. Suppose that the available input data $\mathbf{\Omega}$ will be divided into a training set $T \in \mathbf{\Omega}$ and a validation set $V \in \mathbf{\Omega}$, with each datum having a known class label. First, using Algorithm 1 , each classifier will be trained on set T such that the source competences of each classifier at each point $\mathbf{x}_k \in V$ are determined,

There are many techniques which provide a measure of classifier source competence (Britto et al., 2014; Cruz et al., 2018; Woloszynski & Kurzynski, 2009). In our method, the source com-

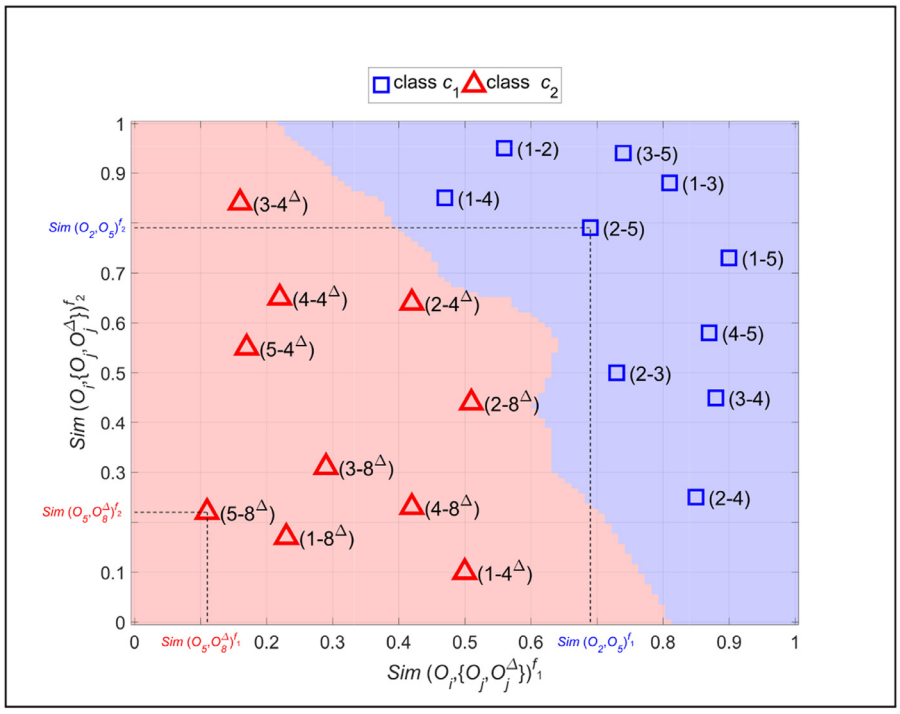


Fig. 1. Principles of the *Sim* similarities' locations on the transformed Cartesian features plane. The description $(5-8^\Delta)$ is equivalent to $Sim(O_5, O_8^\Delta)$, while $(2-5)$ is equivalent to $Sim(O_2, O_5)$, and similarly for other descriptions.

Algorithm 1: Scheme of the source competence calculation.

Data: The pool of classifiers $E = \{\Upsilon_1, \ldots, \Upsilon_D\}$;
the datasets T and V, both with instances described by the feature vector $(\mathbf{x}_k, c_j)$ with known class label c_j;
Result: The source competence $K^{\Upsilon_d}(\mathbf{x}_k, c_j)$;
1 Train each classifier $\Upsilon_d \in E$ on dataset T;
2 **foreach** *validated instance* $(\boldsymbol{x}_k, c_j) \in V$ **do**
3 send the instance $(\mathbf{x}_k, c_j)$ to all classifiers in ensemble E;
4 **foreach** $\Upsilon_d \in E$ **do**
5 calculate the source competence $K^{\Upsilon_d}(\mathbf{x}_k, c_j)$ of a given classifier Υ_d for the instance $\mathbf{x}_k$;

petence of a given classifier Υ_d is defined as follows:

$$K^{\Upsilon_d}(\mathbf{x}_k, c_j) = (2 \cdot I(\Upsilon_d(\mathbf{x}_k, c_j) = c_j) - 1) \\ (\lambda_d(\mathbf{x}_k, c_j) - \lambda_d(\mathbf{x}_k, c \neq c_j)), \tag{8}$$

where $I(con)$ is an indicator function.

The $\lambda_d(\mathbf{x}_k, c_j)$ represents the support function value for the dth classifier Υ_d for a given instance $\mathbf{x}_k$ that belongs either to the class c_j or to the unknown class c. Next determined is the global competence $C(\Upsilon_d|\mathbf{x}_n)$ of a given classifier Υ_d at the new unknown instance $\mathbf{x}_n$, where $\mathbf{x}_n \notin V$ and $\mathbf{x}_n \notin T$. This is calculated utilizing the following rule:

$$C(\Upsilon_d|\mathbf{x}_n) = \sum_{\mathbf{x}_k \in V} K^{\Upsilon_d}(\mathbf{x}_k, c_j) \cdot \exp\left(-dist(\mathbf{x}_n, \mathbf{x}_k)\right)^{\varepsilon}, \tag{9}$$

the *dist* is Euclidean distance between $\mathbf{x}_n$ and $\mathbf{x}_k$.

The Gaussian potential function in (9) strongly promotes the $\mathbf{x}_n$ elements lying close to $\mathbf{x}_k$, while penalizing other elements. In the research conducted into the coefficient ε, we assumed $\varepsilon = 2.5$. In practice, changes in this coefficient affect the biometric factors FAR and FRR. These dependences will be presented in the experimental part of this paper (see Experiment 4). The FAR (False Acceptance Rate) and FRR (False Rejection Rate) factors will be formally defined later. It should be noticed that, in practice, instead of the $\mathbf{x}_n$ and $\mathbf{x}_k$ variables, their equivalent in the form of the *Sim* coefficient will be used.

5. Subject classification by the ensemble

The last step of the proposed approach is to estimate the ensemble's accuracy. This will be done on the basis of *Sim* coefficients. In the classification phase, a subject O provides an instance O^* to be verified. This instance is compared with k reference instances, where $k \leq N$, stored in the database against the subject with the claimed identity. Each time, when the instance O^* is compared with another reference instance, O_i, $i = 1, \ldots, k$ from the database, the additional vectors $\mathbf{v}_i^*$ are created. The construction of these vectors is shown in detail here:

$$\begin{aligned} \mathbf{v}_1^* &= \left[Sim(O_1, O^*)^{f_1}, Sim(O_1, O^*)^{f_2}, \ldots, Sim(O_1, O^*)^{f_L}\right], \\ \mathbf{v}_2^* &= \left[Sim(O_2, O^*)^{f_1}, Sim(O_2, O^*)^{f_2}, \ldots, Sim(O_2, O^*)^{f_L}\right], \\ &\vdots \\ \mathbf{v}_k^* &= \left[Sim(O_k, O^*)^{f_1}, Sim(O_k, O^*)^{f_2}, \ldots, Sim(O_k, O^*)^{f_L}\right], \end{aligned} \tag{10}$$

where f_l, $l = 1, \ldots, L$ are features of the instances O_i, O^*.

After creating vectors $\mathbf{v}_i^*$, for each of them, the global competence of a given classifier is established in accordance with (9). This means that, for every vector $\mathbf{v}_i^*$, the value $C(\Upsilon_d|\mathbf{v}_i^*)$ is computed for $d = 1, \ldots, D$. Classifications using a dynamic selection of classifiers from ensemble E are realized using Algorithm 2 .

6. Experimental investigations

In this section, a series of experiments that confirm the advantages of the presented approach are reported on. In these investigations, ensembles with dynamic selection of base classifiers

Algorithm 2: Verification of instance O^* using dynamic classifier selection.

Data: The unknown instance O^* described by input vectors $\mathbf{v}_1^*, \ldots, \mathbf{v}_k^*$ (see Eq. (10)); the pool of classifiers in the ensemble $E = \{\Upsilon_1, \ldots, \Upsilon_D\}$;$c_1$ is a label of legitimated and c_2 non-legitimated person

Result: Verification of the unknown instance O^* of the subject O;

1 **for** $j = 1, \ldots, k$ **do**

2 **for** $d = 1, \ldots, D$ **do**

3 determine the global competence $C(\Upsilon_d|\mathbf{v}_j)$ (see Eq. (9));

4 On the basis of the $C(\Upsilon_d|\mathbf{v}_j)$ values, form the dynamic ensemble $E^p \in E$ of the p most competent classifiers at recognizing the instance $\mathbf{v}_j$;

5 Classify the instance $\mathbf{v}_j$ using the new ensemble E^p: $E^p(\mathbf{v}_j) = \begin{cases} +1 & \text{if } \sum_{i=1}^{p} \mathrm{I}(\lambda_i(\mathbf{v}_j, c_1) \geq \lambda_i(\mathbf{v}_j, c_2)) > \lceil \frac{p}{2} \rceil \\ -1 & \text{otherwise} \end{cases}$;

6 On the basis of the values $E^p(\mathbf{v}_j)$, determined separately for each vector $\mathbf{v}_1^*, \ldots, \mathbf{v}_k^*$, verify the unknown instance O^* by majority voting: $\Psi := sign\left[\sum_{j=1}^{k} E^p(\mathbf{v}_j)\right]$;

7 instance $O^* = \begin{cases} \text{legitimated} & \text{if } \Psi = '' + '' \\ \text{non} - \text{legitimated} & \text{if } \Psi = '' - '' \end{cases}$;

are compared to those without such selection. Additionally, various variants of the classifiers' work are evaluated.

6.1. Dataset

All the experiments conducted were carried out using a face-content database. This database includes images of various sizes and with different facial expressions. In experiments, the Multi-PIE face database was used (Gross, Matthews, Cohn, Kanade, & Baker, 2010). This dataset is comprised of 750 000 color images of 337 people recorded in up to four sessions over the span of five months. Subjects were imaged from 15 view points and under 19 illumination conditions. Since our study focuses on static lip features, only face images displaying neutral, smiling, squint or indifferent expressions were selected from the database. Images containing non-frontal poses and subjects with their mouthes open were omitted from our experiments. From the database various features were taken into consideration, and seven images per subject were selected. The available data were divided into training, validation and test set. Training and testing sets were extracted using the Leave-One-Out cross validation strategy. The original training set was split in proportion 50 to 50% for training and validation set respectively. It ensures that the risk of overoptimistic results is significantly reduced.

6.2. Data preparation

Our data preparation followed the same method extensively described in our previous work (Wrobel, Doroz, Porwik, Naruniec, & Kowalski, 2017). However, for the reader's convenience, it will be briefly again described here. In the proposed strategy, the data preparation only took the lips' contours and their geometrical measurements into consideration. This meant that in contrast to other methods the texture of the lips' surfaces were ignored. This simplified calculations without any deterioration in the recognition quality. In the first stage, each face's and mouth's region of interest (ROI) were designated. In the next stage, various landmarks were located on each face in the area of the mouth and lips (Kowalski et al., 2017; Wrobel et al., 2017) and specific Euclidean distance-based measurements were calculated between these landmark points.

A face-ground initial shape model $S^0 = [(x_1, y_1), \ldots, (x_n, y_n)]$ was now constructed, consisting of the n facial landmarks and represented by their Cartesian coordinates (Fig. 2). These landmarks can be independently moved which allows for the fitting of individual landmarks to the given face elements. The goal of face alignment is to transform an initial shape S^0 to the shape of a real face S^{Im} on the image Im. This can be done by minimizing alignment error. At the beginning the shape S^{Im} is unknown, so the alignment procedure has to be derived via a multistage learning process. The realization of this task requires a pool of U training images. The region to be searched is limited to the face's ROI (Fig. 3). Individual landmarks $(x, y) \in S^0$ are manually imposed on a given image Im (Fig. 2) and then moved and fitted to all face elements, so that a set $\{Im_i, S_i^0\}_{i=1}^{U}$ of training examples is formed. The learning process stops when the alignment error κ no longer decreases: the differences between the ground shape and the last estimation have stabilized. This means that we have found the landmarks' mapping function $f(Im, S^0) \rightarrow \delta S$ for the facial image Im that minimizes the Euclidean norm error of the landmarks' alignment:

$$\left\| S^0 + \delta S - S^{Im} \right\| = \kappa, \tag{11}$$

where S^{Im} is the true ground shape of the target image Im. This formula shows that the difference between the ground shape and the current estimated shape can be computed. The proposed technique of imposing landmarks is robust to changes in the skin as well as to the background color. The stages discussed above are presented in Figs. 2 and 3. In our approach, on the basis of the final landmarks' points, the Euclidean distances between the appropriate points are calculated. Each of the Euclidean distances form a specific feature f_l. These features are then grouped into a vector of features. The grouping principles and the Euclidean distances are clearly presented in Fig. 4. Distances are now grouped and form various vectors of features (Wrobel et al., 2017). Finally, on the basis of the lip's features, the *Sim* coefficients are established. The technique for the determination of the *Sim* coefficients was presented above. The same vector of features and the same Multi-PIE database were employed in (Wrobel et al., 2017), so the advantages of our new verification strategy can here be reliably presented.

6.3. Statistical evaluation of obtained results

The evaluation of each biometric system primarily depends on number of samples which were used in biometric parameters evaluation. In analysis of measurement results we can estimate the interval in which lies the real measured value. So, results obtained from samples allow to make help generalizations about population.

Let x_1, ... ,x_n be a sample of size of n with mean value $\bar{x}$ and variance s. Let $t = \frac{(\bar{x}-\mu)}{s}\sqrt{n}$, where μ is the estimated arithmetic mean of the population, then variable t has the t-Student distribution with $n-1$ degrees of freedom. Confidence interval (CI) of the μ at $100(1-\alpha)\%$, can be then computed from the formula (Irad, 2008):

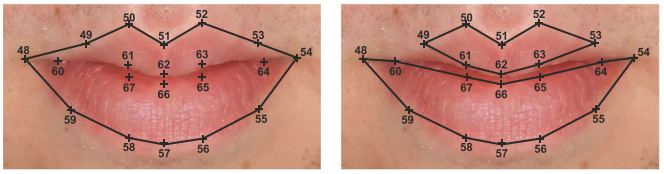
$$CI = \left[\bar{x} - t_{\alpha,n-1}\frac{s}{\sqrt{n}},\ \bar{x} + t_{\alpha,n-1}\frac{s}{\sqrt{n}}\right] = \bar{x} \pm, t_{\alpha,n-1}\frac{s}{\sqrt{n}}, \tag{12}$$

where α is a significance level.

All measurements presented in this paper were generated as average results obtained using the Leave-One-Out technique, where the n-element sample is divided into n subsets, containing one element.

(a) (b) (c)

Fig. 2. A few iterative steps of landmark localization: (a) the initial image, (b) the first alignment step, and (c) the final image with the best alignment of landmarks.



(a) (b)

Fig. 3. (a) Region of interest (ROI) of a given image: (a) contour of the mouth's ROI, (b) the upper and lower numbered landmarks for lip characterization.

We have assumed that $\alpha = 0.05$, so confidence level has been established at 95%. Presented in the following tables confidence intervals confirm the statistical soundness of the proposed method.

6.4. Experiment 1

The lip verification system was measured using three common factors (Bolle, 2004): FAR (False Acceptance Rate), FRR (False Rejection Rate), and ACC (Accuracy):

$$\mathrm{FAR} = \frac{\text{number of non}-\text{legitimate users accepted}}{\text{number of non}-\text{legitimate tested}} \cdot 100\%, \tag{13}$$

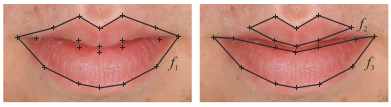
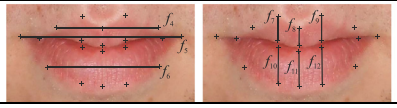
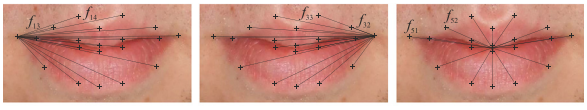
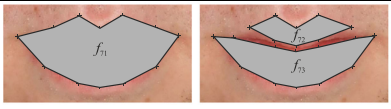
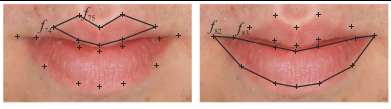
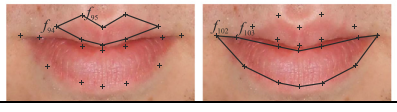
	The lengths of the mouth's contour around the ROI, and upper and lower lip contours: f_1, f_2, f_3.
	Widths and heights of the lips at selected points: $f_4, \ldots, f_{12}$.
	Distances to the remaining contour points from the left mouth corner, right mouth corner and from the center of the mouth: $f_{13}, \ldots, f_{70}$.
	Areas of the mouth's ROI and of the upper and lower lip: f_{71}, f_{72}, f_{73}.
	Coordinates (x, y) of the upper and lower lips' contour: $f_{74}, \ldots, f_{93}$.
	Curvatures α of the upper and lower lips' contour: $f_{94}, \ldots, f_{113}$.

Fig. 4. Measurement principles of the lip-based features.

Table 1
Influence of the number of instances (images) of a selected person O on a classifier's PNN+PSO accuracy and CI at the 95% around Accuracy (ACC), FAR and FRR factors.

	No. of instances in the set π_1					
	2	3	4	5	6	7
ACC [%]	20.31 [20.16–20.46]	65.78 [65.32–66.24]	84.45 [83.89–85.01]	86.95 [86.34–87.56]	86.96 [86.38–87.54]	86.95 [86.32–87.58]
FAR [%]	77.69 [77.10–78.28]	25.91 [25.71–26.11]	11.44 [11.26–11.62]	11.19 [11.10–11.28]	11.18 [11.12–11.24]	11.20 [11.10–11.30]
FRR [%]	82.54 [82.07–83.01]	44.32 [44.03 –44.61]	18.63 [18.30–18.96]	17.93 [17.80 18.06]	17.93 [17.80–18.06]	17.92 [17.70–18.14]

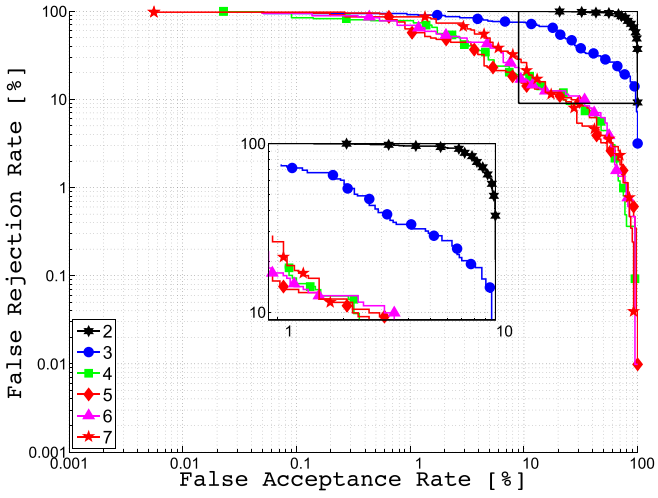


Fig. 5. The ROC curves for various instances used in PNN+PSO classifier. (ROC: Receiver Operating characteristic Curve).

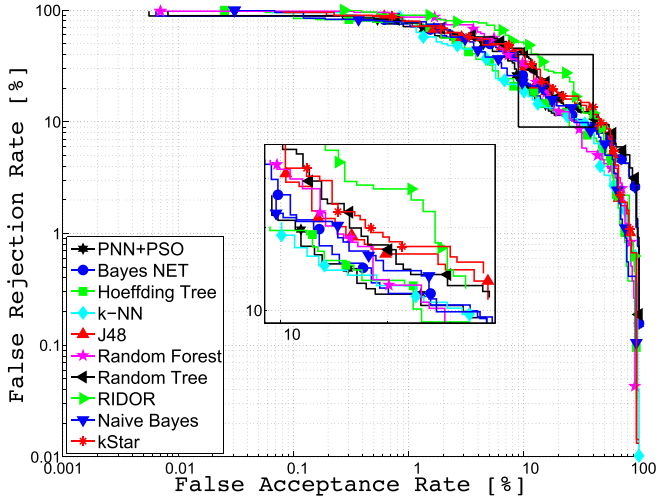


Fig. 6. Comparisons of ROC curves for ten single base classifiers that, subsequently, will form an ensemble E.

$$\mathrm{FRR} = \frac{\text{number of legitimate users rejected}}{\text{number of legitimated users tested}} \cdot 100\%, \qquad (14)$$

Let A be a number of legitimated users correctly recognized.
Let B be a number of non-legitimate users correctly recognized.
Let C be a number of users tested.

For such assumptions, Accuracy (ACC) of the biometric system can be defined as follows:

$$\mathrm{ACC} = \frac{A+B}{C} \cdot 100\%. \qquad (15)$$

As was explained in the previous sections of this paper, we proposed ten popular classification algorithms as base classifiers to form an ensemble E which together will be used to verify the subjects. Our main goal is to improve the accuracy of a biometric system based on extracted lip features (Fig. 4). In the first experiment we investigated how many instances of a given subject (here, how many lip images of a given person O) were needed to obtain a classifier's highest accuracy. In practice, what was previously explained, instead of lip images we employ *Sim* coefficients but from biometrics point of view the most interesting is number of images of a given subject for assessment of the method. From experiments performed follow that only five images of a given person is enough for high recognition accuracy. For more number of images, accuracy factor did not increased. So, proposed solution is better comparing to results in Table 5 All results presented in this paper were generated as average results obtained using the Leave-One-Out technique. The results are gathered in Table 1 and Fig. 5.

From Table 1 and Fig. 5 it follows that for the biometric Multi-PIE face dataset, the highest classification accuracy is reached at five lip-based image samples for each subject. For this number of instances the lowest FAR/FRR value was also observed. Thus, the proposed solution seems to be optimal. Computed confidence intervals are low, thus allowing us to assume with relative certainty that the ACC, FAR and FRR presented in this section are reliable. In all future experiments, five instances per class c_1 (legitimated) and c_2 (non-legitimated) will be treated as the reference, regardless of the type of classifier employed.

6.5. Experiment 2

Set up: The aim of the second experiment was to investigate the behavior of various classifiers in the lip recognition process, where the quality of classification was measured by the classifiers' accuracy levels. In comparative studies, literature-proven canonical classifiers were implemented in Matlab, R, or WEKA systems, and each was then evaluated against *Sim*-type data. The classifiers were evaluated separately as single algorithms. It should be noted that, in the future, classifiers will work as ensembles. The parameters used in this experiment are listed in Table 2.

Results: Investigations utilizing the biometric Multi-PIE dataset were performed separately for each classifier. The accuracy values for each classifier are presented in Table 3. It can be seen that the accuracy of the base classifiers is quite high and that, in the majority of cases, it exceeds 80%, and for a PNN classifier boosted by the PSO strategy is even over 86%.

Similarly to previously, as both Table 3 and Fig. 6 confirm, in the group of ten classifiers, the PNN classifier boosted by the PSO procedure achieves the highest accuracy factor. All investigations in this experiment were carried out on the basis of five instances per subject (person).

6.6. Experiment 3

Set up: As was stated previously, ten base classifiers operate to form the proposed ensemble E. In this experiment, the classifiers

Table 2
Parameter settings for each classifier.

Classifier	Parameter setting.
PNN	Kernel type: Gaussian; Nodes in hidden layer: 220; Number of particles: 30; Number of epochs: 20; Smoothing parameter range $\sigma \in [0.1, 5]$.
Bayes NET	
Hoeffding Tree	Minimum fraction of weight: 0.02.
J48	Pruned, Decision trees, Number of trees: 100; Minimum number of instances per leaf: 2.
Random Forest	Number of trees: 10.
Random Tree	The number of randomly chosen attributes: $log_2(numberofattributes)$
RIDOR	Number of folds: 3; Number of shuffles: 1; The minimal weights of instances within a split: 2.
k-NN	k=3, Distance weighting: no distance weighting; Distance type: Euclidean.
Naive Bayes	Maximum a posteriori probability estimation with normal Gaussian data distribution.
kStar	Global blend: 20.

Table 3
Accuracy for various individual base classifiers in the proposed ensemble E and CI at the 95% around Accuracy (ACC), FAR and FRR factors.

Classifier	ACC [%]	FAR [%]	FRR [%]
PNN+PSO	86.95 [85.88–88.02]	11.19 [11.00–11.38]	17.93 [17.61–18.25]
Bayes NET	84.67 [83.60–85.74]	12.59 [12.45–12.73]	18.01 [17.76–18.26]
Hoeffding Tree	84.11 [83.14–85.08]	12.49 [12.36–12.62]	19.13 [18.90–19.36]
k-NN	83.58 [82.68–84.48]	13.98 [13.85–14.11]	18.38 [18.01–18.71]
J48	81.79 [80.81–82.77]	14.97 [14.70–15.24]	22.55 [22.27–22.83]
Random Forest	84.23 [83.22–85.24]	13.69 [13.59–13.79]	19.21 [18.97–19.45]
Random Tree	83.00 [82.05–83.95]	15.32 [15.04–15.60]	20.85 [19.60–22.10]
RIDOR	78.46 [77.50–79.42]	19.82 [19.50–20.14]	25.47 [25.13–25.81]
Naive Bayes	83.44 [82.44–84.44]	14.71 [14.41–15.01]	20.63 [20.37–20.89]
kStar	80.43 [79.34–81.52]	16.99 [16.79–17.19]	23.87 [23.38–24.36]

comprising the pool have been dynamically selected on the basis of their individual competences, in accordance with Algorithm 2. In our biometric case, only legitimated (c_1) and non-legitimated (c_2) users were considered. The verification procedure was processed on the basis of three vectors $\mathbf{v}^*$ (see Eq. (10)), randomly selected from the set of instances for a verified person.

Next, competences for an unknown instance (represented by the *Sim* coefficients) were computed on the basis of the formulas (8) and 9).

The best classifiers in the pool were determined three times, each time for another $\mathbf{v}_i^*$, $i = 1, \ldots, 3$.

Results: Investigations were carried out into various numbers p of dynamically selected classifiers in the pool $E^p \in E$. These dependences are presented in Table 4 and Fig. 7.

If we compare Table 3 – in which only single classifiers were considered – with the new results of the multiple classification system, we can conclude that classification by dynamic ensemble represents a significantly improvement. Analysis of Table 4 confirms these observations: for the same dataset, verification accuracy is now considerably higher.

6.7. Experiment 4

Set up: In Experiment 1, the FAR and FRR factors were stated and all investigations then performed were carried out on the basis of these factors. In biometric systems, authentication is approved when the difference between the system's enrollment and test phases is less than some selected acceptance threshold. This

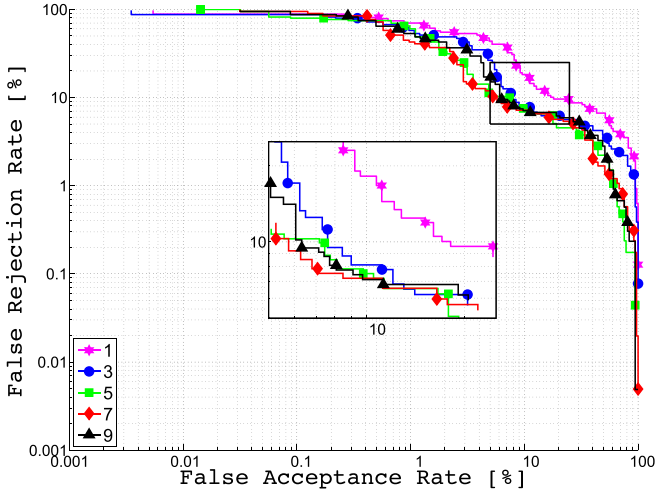


Fig. 7. Ensemble accuracy levels for different numbers $p = 1, 3, 5, 7, 9$ of base classifiers in the pool.

Table 4
Ensemble accuracy as dependent on the number p of classifiers in the pool E^p and CI at the 95% around Accuracy (ACC), FAR and FRR factors .

Number p	ACC [%]	FAR [%]	FRR [%]
1	87.02 [86.04–88.00]	10.93 [10.80–11.06]	15.22 [15.00–15.44]
3	91.34 [90.72–91.96]	7.45 [7.38–7.52]	10.05 [9.90–10.20]
5	91.86 [91.01–92.71]	7.59 [7.32–7.96]	10.11 [9.90–10.32]
7	91.90 [91.05–92.75]	7.49 [7.42 –7.56]	10.13 [9.84–10.42]
9	91.86 [91.00–92.71]	7.55 [7.08–8.02]	10.01 [9.42–10.50]

threshold should change depending on the system's requirements. In the earlier studies in this paper such a change was not possible. Thanks to the introduction of the constant ε into formula (9), the proposed solution is now more elastic and allows for the modified system's acceptance error level to depend on the system's requirements.

As stated in the theoretical part of this paper, the global competence $C(\Upsilon_d|\mathbf{x}_n)$ of a given classifier Υ_d depends, among other things, on a Gaussian potential function and a constant ε (see Eq. (9)). Values of this Gaussian function additionally promote unknown elements lying close to reference points, while penalizing others. Further, by means of a constant ε, we can change - within some limited range - the ensemble's important FAR/FRR factors. It can be observed that the source competence $K^{\Upsilon_d}(\mathbf{x}_k, c_j)$ of the base classifier Υ_d from (9) directly depends on the Gaussian func-

Table 5
Comparison of our results with other strategies by the mean and standard deviation of the classification accuracy (ACC), FAR, FRR and ERR factors of each run over all calculations.

Method	FAR [%]	FRR [%]	EER [%]	ACC [%]	Database	Comments
Verification process						
Our new approach	**7.55** ± 0.23	**10.01** ± 0.28	–	**91.90** ± 1.47	Multi-PIE: 337 persons, 5 images per person.	Color image ✓
						Grayscale images ✓
Wrobel et al. (2017)	15.32 ± 0.35	12.63 ± 0.56	12.71 ± 0.53	86.95 ± 1.21	Multi-PIE: 337 persons, 5 images per person.	Color image ✓
						Grayscale images ✓
	15.53 ± 0.43	11.67 ± 0.51	12.45 ± 0.52	87.14 ± 1.06	PUT: 100 persons, 5 images per person.	
	15.99 ± 0.57	11.63 ± 0.47	12.57 ± 0.41	87.26 ± 0.91	Own: 50 persons, 5 images per person.	
Wang and Liew (2012)	–	–	18.50	–	Own: 40 persons.	Color image ✓
						Grayscale images ✗
Travieso et al. (2011)	–	–	12.75	–	PIE: 68 persons, 11 images per person.	Color image ✓
						Grayscale images ✗
Identification process[1)]						
Kim, Lee, Hwang, Baik, and Chung (2004)	1.10	3.60	–	95.30	Own: 24 persons.	Color image ✓
						Grayscale images ✗
Choraś (2010)	–	–	–	82.00	Own: 38 persons, 3 images per person.	Color image ✓
						Grayscale images ✗
Travieso et al. (2014)	–	–	–	99.89 ± 0.16	Own: 50 persons, 10 images per person.	Color image ✓
						Grayscale images ✗
	–	–	–	97.13 ± 0.56	PIE: 68 persons, 11 images per person.	
	–	–	–	98.10 ± 0.40	RaFD: 60 persons, 9 images per person.	

[1)] Results were achieved for closed sets of data (closed-set), so without counterfeits. It should be stressed that our method is designated to work in verification mode, therefore direct comparison with an identification system is not possible.

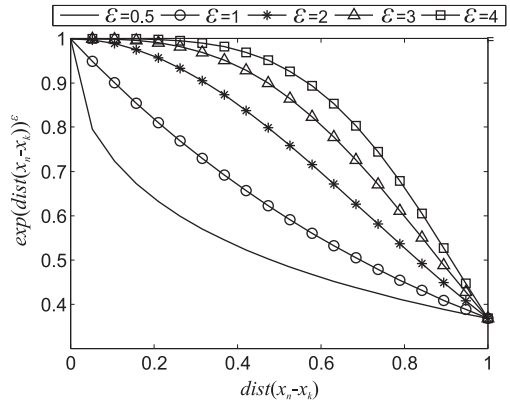


Fig. 8. Family of the Gaussian function curves and their dependence on the distance $dist(\mathbf{x}_n - \mathbf{x}_k)$ and the constant ε (see Eq. (9).

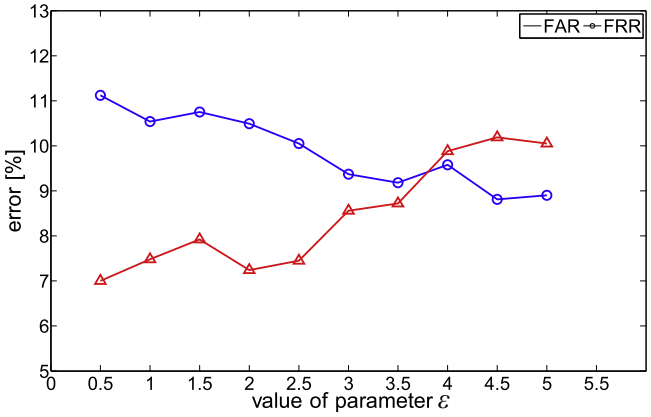


Fig. 9. Dependences between FAR/FRR and constant ε for ensemble E^p with $p = 3$ base classifiers in ensemble.

tion in this equation. If the normalized distance *dist* between an unknown element $\mathbf{x}_n$ and a known element $\mathbf{x}_k$ decreases, then the Gaussian function will additionally promote (boost) $\mathbf{x}_n$. What this means is that $\mathbf{x}_n$ will be recognized as an element belonging to the class affirmed by $\mathbf{x}_k$. Each of the Gaussian function graphs, for each of the various distances between $\mathbf{x}_n$ and $\mathbf{x}_k$, and their dependence on the parameter ε, are depicted in Fig. 8

Results: Taking into consideration Eqs. (8) and (9) and the curves from Fig. 8, we can designate the characteristic of the ensemble E^p in this particular biometric system in which lip-based features are analyzed. In such a case, an adaptively controlled acceptance threshold is formed by changes of the constant ε. The characteristics of the biometric system evaluated by FAR/FRR and with a constant ε are presented in Fig. 9.

Similar experiments can be carried out for various numbers of base classifiers in an ensemble. Fig. 9 can also be compared with Table 4 especially for the $p = 3$ base classifiers in ensemble E^p. To the best of our knowledge, this is the first time that such research has been undertaken.

6.8. Experiment 5

For the purposes of the current experimental analysis, we compared the dynamic classifier selection model with other models of classification. These comparison results are presented in Fig. 10. From these box plots, we can observe that the single best classifier (Table 3) and ensemble $E^{p=1}$ (Table 4), work with similarly low Accuracy (ACC) factors. For $p = 3$ up to $p = 9$ base classifiers working together as an ensemble E^p, the verification system achieves

Table 6
Average time execution of the most important stages of verification procedure.

Stage	time [ms]
Determination of the landmarks for a single face image (see Fig. 2)	177
Determination of the features $f_1 \dots f_{113}$ for a single image (see Fig. 4)	104
Selection of the most competent classifiers (see Section 4)	13
Classification time of a single person (see Section 5)	31
Total verification time of a single person[a]	325

[a] Time measurement was made in the Matlab. In our experiments, the measurement time was obtained on a PC class computer equipped with an Intel i5-2500 processor running at 3.3 GHz, with 16GB of RAM and a Windows 7 x64 operating system.

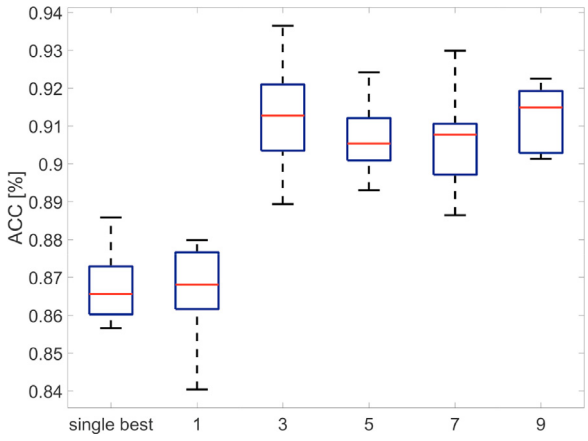


Fig. 10. Statistical distribution: accuracy comparison of single classifier and ensemble.

significantly better results. Each box indicates: the max and min accuracy value; the upper and lower quartile; the median. Ensembles with 3, 5, 7 and 9 base classifiers work similarly, although the ensemble $E^{p=3}$ seems to be the best because its statistical characteristics are strongly symmetrical.

6.9. Experiment 6

The most important state-of-the-art work in the field of lip-based processing were presented in the Introduction. From our point of view, these trends are not particularly interesting, although they use, in some of their steps, similar techniques to those in this paper. Our approach is focused on the biometric recognition of persons, so our strategy will be compared to relevant papers. The recent and most interesting investigations in this field are summarized in Table 5. From Table 5 we can evaluate how many images have to be used to obtain highest classifier's accuracy and how large is the dataset. In other words how many samples have been taken into account during biometric experiments.

In this table, all recognition factors as published by their authors have been shown. It should be noted that our verification method was tested in an "open set" mode using a large and publicly accessible face database. In practice, a verification mode is most frequently used. In an open-set, there is no guarantee that a record of any given individual is contained in the database. Thus, a biometric verification system can be deliberately cheated and, if so, this should be indicated. Analysis of Table 5 shows that the most important results were presented in the work (Travieso, Zhang, Miller, & Alonso, 2014), in which recognition accuracy was very high when the biometric system worked in its identification mode. Additionally, in the works (Travieso et al., 2014; Travieso, Zhang, Miller, Alonso, & Ferrer, 2011), the databases had from 9 to 11 images per person, whereas we utilized datasets with only 5 instances per person. This means that in our approach the classifiers' learning process was more restricted and thus more realistic from a biometric point of view. Results of the lip-based identifications were achieved for closed sets of data (closed-set), so without any counterfeits. Such a mode is less restricted as compared to our strategy in which counterfeits were acceptable. It should be stressed that our method is designated to work in verification mode, therefore in direct comparison with an identification system is not possible.

Usefulness of the proposed method can also be assessed in terms of time complexity. Average time complexity has been evaluated separately for the most important stages of the verification procedure. Time dependences have been gathered in Table 6. Time complexity also depends on the type of classifiers used. For another pull of classifiers presented in 6 estimates can be slightly different.

7. Final remarks

Nowadays, it is trivial to remark that high volumes of data can be stored. However, these data have to be efficiently analyzed, so the use of efficient computational techniques has become a problem of the utmost importance. In this paper we have summarized the main research on ensembles of biometric classifiers. The key issues in this area are the classifiers' diversity and the methods of classifier combination. Diversity is believed to provide improved accuracy and classifier performance. The results of the experiments we conducted demonstrate that our classification model uses a new kind of similarity score that brings benefits in terms of the verification of biometric data. The investigations we performed demonstrate the validity of our results. A public dataset retrieved from a biometric repository was used to test the quality of the proposed approach. All the results clearly showed the remarkable performance of ensembles generated by our proposed dynamic base-classifier selection, an approach which outperformed all other methods. Our studies made it clear that the extraction of face-based lip features might become an important tool for biometric verification. Our conclusions concerning this new lip-based biometric verification method in a multiple classifier environment can be summarized as follows:

- Our verification method is robust to changes in skin color and can be used for color as well as gray-scale images.
- The proposed ensemble-based strategy demonstrates a high recognition accuracy. The accuracy all of each classifier was checked using a Leave-One-Out validation procedure (Tables 1–4).
- The results obtained were achieved thanks to the use of a new composed lip feature and the dynamic selection of base classifiers. The *Sim* feature was computed on the basis of an accessible similarity coefficient and on the lip features extracted from the face-based image. Using the *Sim* coefficients allows for increased input data, which is important for learning procedures.

- The correctness and usefulness of the proposed algorithm has been statistically analyzed (Fig. 10) and confidence intervals analyzes.
- We have shown that a biometric verification system can be tuned by an appropriate ensemble given a global competence determination (see Eq. (9)). By means of an additional constant ε, the characteristics of the FAR/FRR charts can be modeled (Figs. 8 and 9)
- The main goal of building a classifier is high accuracy. The investigations undertaken confirm that this goal has been achieved via a successive, dynamic choice of base classifiers from a pool of classifiers.

The results indicate that, in real problems, a multiple classifiers approach with a dynamic selection of classifiers for lip-based verification is a new, successful, strategy that has never before been used in biometrics.

What is clear is that any biometric system will make mistakes, and that the true value of the various error rates cannot be computed or theoretically established. It can be done by statistical estimation of the errors using of biometric samples and it was presented in this paper. Mistakes occurring in the our biometric system result from the various reasons:

- FAR mistakes can occur because for each person, only mouth ROI is analyzed and no other biometric features are taken into consideration. Many persons have similar mouth shape, so these errors are registered. However, such strategy allows to speed up processing procedures.
- FRR mistakes result from the inability to stabilize the lip muscles. So, slight changes of mouth shapes (with slight face wince for example) will be always observed.

Is should be also mentioned that good results will be obtained when images with mouth areas will be taken with a high resolution.

References

Ahmed, M., Rasool, A. G., Afzal, H., & Siddiqi, I. (2017). Improving handwriting based gender classification using ensemble classifiers. *Expert Systems with Applications, 85*, 158–168.

Bolle, R. (2004). Guide to biometrics. *Springer professional computing*. New York: Springer.

Britto, A. S., Sabourin, R., & Oliveira, L. E. (2014). Dynamic selection of classifiersa comprehensive review. *Pattern Recognition, 47*(11), 3665–3680.

Choraś, M. (2010). The lip as a biometric. *Pattern Analysis and Applications, 13*(1), 105–112.

Cpałka, K., Zalasiński, M., & Rutkowski, L. (2016). A new algorithm for identity verification based on the analysis of a handwritten dynamic signature. *Applied Soft Computing, 43*(Supplement C), 47–56.

Cruz, R. M., Sabourin, R., & Cavalcanti, G. D. (2018). Dynamic classifier selection: Recent advances and perspectives. *Information Fusion, 41*(Supplement C), 195–216.

Doroz, R., Porwik, P., & Orczyk, T. (2016). Dynamic signature verification method based on association of features with similarity measures. *Neurocomputing, 171*, 921–931.

Fierrez, J., Morales, A., Vera-Rodriguez, R., & Camacho, D. (2018). Multiple classifiers in biometrics. part 2: Trends and challenges. *Information Fusion, 44*, 103–112.

Ghoualmi, L., Draa, A., & Chikhi, S. (2016). An ear biometric system based on artificial bees and the scale invariant feature transform. *Expert Systems with Applications, 57*, 49–61.

Gross, R., Matthews, I., Cohn, J., Kanade, T., & Baker, S. (2010). Multi-pie. *Image Vision Comput., 28*(5), 807–813.

Irad, B.-G. (2008). *Encyclopedia of statistics in quality and reliability*. American Cancer Society.

Khan, F. A., Tahir, M. A., Khelifi, F., Bouridane, A., & Almotaeryi, R. (2017). Robust off-line text independent writer identification using bagged discrete cosine transform features. *Expert Systems with Applications, 71*, 404–415.

Kim, J. O., Lee, W., Hwang, J., Baik, K. S., & Chung, C. H. (2004). Lip print recognition for security systems by multi-resolution architecture. *Future Generation Computer Systems, 20*(2), 295–301. Modeling and simulation in supercomputing and telecommunications

Koprowski, R., & Bocklitz, T. (2016). Image processing in biomedical diagnosis. *Journal of Biophotonics, 9*(5), 434–435.

Kowalski, M., Naruniec, J., & Trzcinski, T. (2017). Deep alignment network: A convolutional neural network for robust face alignment. *The IEEE conference on computer vision and pattern recognition (CVPR) workshops*.

Krysmann, M., & Kurzynski, M. (2013). Methods of learning classifier competence applied to the dynamic ensemble selection. In R. Burduk, K. Jackowski, M. Kurzynski, M. Wozniak, & A. Zolnierek (Eds.), *Proceedings of the 8th international conference on computer recognition systems CORES 2013* (pp. 151–160)). Heidelberg: Springer International Publishing.

Kuncheva, L. I. (2004). *Combining pattern classifiers: Methods and algorithms*. Wiley-Interscience.

Lee, H. K., & Kim, S. B. (2018). An overlap-sensitive margin classifier for imbalanced and overlapping data. *Expert Systems with Applications, 98*, 72–83.

Porwik, P., Doroz, R., & Orczyk, T. (2014). The k-NN classifier and self-adaptive hotelling data reduction technique in handwritten signatures recognition. *Pattern Analysis and Applications, 18*(4), 983–1001.

Porwik, P., Doroz, R., & Orczyk, T. (2016). Signatures verification based on PNN classifier optimised by PSO algorithm. *Pattern Recognition, 60*, 998–1014.

Raman, R., Sa, P. K., Majhi, B., & Bakshi, S. (2017). Fusion of shape and texture features for lip biometry in mobile devices. *Book: Mobile Biometrics, Chap. 6*, 155–176.

Rodrigues, D., Silva, G. F., Papa, J. P., Marana, A. N., & Yang, X.-S. (2016). Eeg-based person identification through binary flower pollination algorithm. *Expert Systems with Applications, 62*, 81–90.

Tadeusiewicz, R., & Horzyk, A. (2014). Man-Machine interaction improvement by means of automatic human personality identification. In *Lecture notes in computer science (including subseries lecture notes in artificial intelligence and lecture notes in bioinformatics): 8838* (pp. 278–289).

Tirumala, S. S., Shahamiri, S. R., Garhwal, A. S., & Wang, R. (2017). Speaker identification features extraction methods: A systematic review. *Expert Systems with Applications, 90*, 250–271.

Travieso, C. M., Zhang, J., Miller, P., & Alonso, J. B. (2014). Using a discrete hidden markov model kernel for lip-based biometric identification. *Image and Vision Computing, 32*(12), 1080–1089.

Travieso, C. M., Zhang, J., Miller, P., Alonso, J. B., & Ferrer, M. A. (2011). Bimodal biometric verification based on face and lips. *Neurocomputing, 74*(1415), 2407–2410.

Wang, S.-L., & Liew, A. W.-C. (2012). Physiological and behavioral lip biometrics: A comprehensive study of their discriminative power. *Pattern Recognition, 45*(9), 3328–3335.

Woloszynski, T., & Kurzynski, M. (2009). On a new measure of classifier competence in the feature space. In M. Kurzynski, & M. Wozniak (Eds.), *Computer recognition systems 3* (pp. 285–292)). Berlin, Heidelberg: Springer Berlin Heidelberg.

Woźniak, M., Graña, M., & Corchado, E. (2014). A survey of multiple classifier systems as hybrid systems. *Information Fusion, 16*, 3–17. Special Issue on Information Fusion in Hybrid Intelligent Fusion Systems

Wrobel, K., Doroz, R., Porwik, P., Naruniec, J., & Kowalski, M. (2017). Using a probabilistic neural network for lip-based biometric verification. *Engineering Applications of Artificial Intelligence, 64*(Supplement C), 112–127.

Zhang, Z.-L., Luo, X.-G., García, S., Tang, J.-F., & Herrera, F. (2017). Exploring the effectiveness of dynamic ensemble selection in the one-versus-one scheme. *Knowledge-Based Systems, 125*, 53–63.

Zois, E. N., Alewijnse, L., & Economou, G. (2016). Offline signature verification and quality characterization using poset-oriented grid features. *Pattern Recognition, 54*(Supplement C), 162–177.