Int. J. Appl. Math. Comput. Sci., 2015, Vol. 25, No. 4, 915–925
DOI: 10.1515/amcs-2015-0065

STATISTICAL TESTING OF SEGMENT HOMOGENEITY IN CLASSIFICATION OF PIECEWISE–REGULAR OBJECTS

ANDREY V. SAVCHENKO [a,*], NATALYA S. BELOVA [b]

[a] Laboratory of Algorithms and Technologies for Network Analysis
National Research University Higher School of Economics, 136 Rodionova St., Nizhny Novgorod 603093, Russia
e-mail: avsavchenko@hse.ru

[b] Faculty of Computer Science
National Research University Higher School of Economics, 20 Myasnitskaya St., Moscow 101000, Russia
e-mail: nbelova@hse.ru

The paper is focused on the problem of multi-class classification of composite (piecewise-regular) objects (e.g., speech signals, complex images, etc.). We propose a mathematical model of composite object representation as a sequence of independent segments. Each segment is represented as a random sample of independent identically distributed feature vectors. Based on this model and a statistical approach, we reduce the task to a problem of composite hypothesis testing of segment homogeneity. Several nearest-neighbor criteria are implemented, and for some of them the well-known special cases (e.g., the Kullback–Leibler minimum information discrimination principle, the probabilistic neural network) are highlighted. It is experimentally shown that the proposed approach improves the accuracy when compared with contemporary classifiers.

Keywords: statistical pattern recognition, classification, testing of segment homogeneity, probabilistic neural network.

1. Introduction

In a classification task it is required to assign the query object X (facial photo, speech signal, image of natural scenes, text) to one of $C > 1$ classes (Theodoridis and Koutroumbas, 2008). It is usually assumed that all classes are specified with a given database $\{X_r\}$, $r \in \{1, \ldots, R\}$ of $R \geq C$ cases (instances or models). For each model object X_r, a class label $c(r) \in \{1, \ldots, C\}$ is given. If the analyzed objects are represented as *feature vectors* with a fixed dimension, traditional classification methods can be used, e.g., linear/quadratic discriminant analysis (LDA/QDA), feed-forward multi-layer perceptrons (MLPs) and support vector machines (SVMs) (Haykin, 2008). Recently the research has shifted the focus to the objects containing several independent homogeneous (regular, "stationary") parts (Świercz, 2010). Each part can be considered a sample of independent identically distributed (i.i.d.) feature vectors. We will call such objects *composite* or *piecewise-regular*.

Recognition of piecewise-regular objects includes image and speech processing tasks. For instance, the image of the whole object or the keypoint neighborhood (Dalal and Triggs, 2005; Lowe, 2004) is divided into a grid of blocks; each block is processed independently (as in JPEG/MPEG compression algorithms). Speech can be considered a sequence of independent phonemes (Benesty *et al.*, 2008): features of different segments inside one word may have nothing in common as they correspond to distinct phonemes. Thus, modern recognition methods determine the structure of the classifier based on a piecewise-regular model of analyzed objects. The query object X and all models X_r are considered sequences of, respectively, K and K_r relatively independent homogeneous (regular) segments. Segmentation of practically important objects (images, speech) has been well studied (Theodoridis and Koutroumbas, 2008) and is not discussed in this article.

The main contribution of our paper is as follows. We introduce a novel approach to design classifiers of audiovisual objects (images, speech utterances) by testing segment homogeneity based on the probabilistic

*Corresponding author

model of composite object representation as a sequence of i.i.d. segments. Based on this approach, we present asymptotically minimax classifiers for parametric and non-parametric estimates of unknown probability densities. In the former case, an exponential family of distributions of each segment is assumed. In the latter case, nonparametric Parzen kernels are used. Our approach allows strictly proving the known insufficient quality of Bayesian classifiers, e.g., the probabilistic neural network (PNN), the Kullback–Leibler (KL) minimum information discrimination principle, usually explained by the naive assumption of the statistical independence of features. We presented another (probabilistic) explanation when the classification task is referred to as a composite hypothesis testing of segment homogeneity.

The rest of the paper is organized as follows. Section 2 presents the literature review of recognition methods of composite objects. In Section 3, we introduce our probabilistic mathematical model of composite objects and present classification criteria for parametric and non-parametric estimates of probability densities. In Section 4, experimental results in classification of images and Russian speech are presented. Finally, concluding comments are given in Section 5.

2. Literature review

Methods of piecewise-regular object recognition are primarily determined by the characteristics of the available database of models. The most well-studied are tasks with a *large number of available models for each class* ($R \gg C$), e.g., optical character recognition or classification of traffic signs and phonemes. Another practically important case is the *small sample size* ($C \approx R$). Let us describe both instances in detail.

Following the traditional approach to pattern recognition (Theodoridis and Koutroumbas, 2008), feature extraction is the crucial step to achieve high accuracy. Experimental studies clearly show that popular classifiers (LDA/QDA, MLP, SVM, etc.) are characterized by the best quality for *uncorrelated features* (Haykin, 2008). Hence, classical recognition procedures include normalization and decorrelation with, e.g., principal/independent component analysis (PCA/ICA) for primitive descriptions of analyzed objects (color matrix for images, signal of acoustic pressure amplitude of its fast Fourier transform (FFT) for speech). For instance, weighed histograms of gradient orientation are calculated in the neighborhood of image keypoints (Dalal and Triggs, 2005). For speech fragments the spectrum is estimated in several acoustic frequency bands and further decorrelated by the modified cosine transform. As a result, mel-frequency cepstral coefficients (MFCCs) are obtained (Benesty *et al.*, 2008). Thus, the most evident way to recognize a composite object is to divide it into a fixed number of homogeneous segments, estimate features for all segments, unite them into a single feature vector and classify them with the traditional MLP or SVM.

Unfortunately, homogeneous segments are extracted inaccurately—several important segments can be duplicated (e.g., vowels in speech signals) or missed (consonant phonemes). Hence, the described approach is ineffective for many tasks, such as automatic speech recognition (ASR) (Sas and Żołnierek, 2013). To overcome this drawback, preliminary *segmentation* of the query object and all models is performed. Next, the segments are dynamically aligned with dynamic programming techniques (dynamic time warping, DTW) (Benesty *et al.*, 2008): each segment of the input object is compared with several segments of each model in some neighborhood of the segment. It is obvious that such alignment causes a further increase in the average recognition time. Unfortunately, such an approach is known to be characterized with low accuracy if the number of classes becomes high. In consequence, since 1980 the most popular methods have been based on hidden Markov models (HMMs), specially developed for classification of piecewise-stationary objects (Benesty *et al.*, 2008). The HMM is a standard de-facto in modern ASR libraries (CMU Sphinx, HTK, Kaldi, etc.).

The major restriction of this approach is the requirement of features to be uncorrelated or independent. It is not surprising that the recent research has shifted the focus to the usage of *primitive correlated features and more complex classifiers*, e.g., the deep neural network (DNN) (LeCun *et al.*, 2015), which showed higher accuracy when compared with the state-of-the-art SVM for several model tasks. At first, restricted Boltzmann machines were used as unsupervised stacked auto encoders to extract features with final layers trained by the back propagation on the modern GPUs (Hinton *et al.*, 2006). However, the best results are achieved with other neural network methods with a deep architecture, such as the convolution neural network (CNN) (LeCun *et al.*, 1998), which do not use unsupervised learning. The CNN consists of sequentially mapped layers of convolution and sub-sampling. Recently, the subsampling layer has been replaced with the max pooling layer and several CNNs are united in a committee in the multi-column GPU max-pooling CNN (MC-GPU-MPCNN) (Ciresan *et al.*, 2012), which allowed reaching a better-than-human recognition rate in a traffic sign image recognition task ($C = 43$ classes, $R = 39209$ models). For handwritten digits recognition from the MNIST dataset ($C = 10$ classes, $R = 60000$), this DNN showed 99.65% accuracy.

It is important to note that the CNN can be applied not only in image recognition tasks. For instance, its special case, the time-delay neural network, showed good

accuracy of phoneme recognition (Bottou *et al*., 1990). However, most widely-used in the ASR task are other methods with prior phoneme segmentation. They allow using the phonetic approach widely developed in HMM studies. The DNN is used instead of the GMM (Gaussian mixture model) to decrease the word error rate to 10–15% (Hinton *et al*., 2012). It was shown by the Microsoft research team (Huang *et al*., 2013) that the usage of simple FFT features instead of the MFCC allows the DNN to increase the accuracy.

Thus, classifiers with a deep architecture and a large number of parameters for several model tasks allow achieving an equal-to-human accuracy. However, the situation becomes quite more complicated if the training set contains a small number of models per each class (in the worst case, one model per class, $C = R$) (Tan *et al*., 2006). In this case a proper distance metric between analyzed objects should be chosen. To classify the query object, nearest neighbor (NN) based rules, e.g., the k-NN or radial-basis networks (Haykin, 2008), are widely applied. A remarkable method is the binary classification of the distances between segments (Liao *et al*., 2007). It is assumed that from 2 to 5 models are available for each class.

Despite the impossibility to train complex classifiers, e.g., the CNN, for such a training set, it is possible to assign the vector of distances between corresponding segments to one of two classes when these distances are calculated between objects of the same or distinct classes. The query object is segmented, and the histograms of local binary patterns (LBPs) are estimated for each segment. The distance vector between the corresponding segments of the query image and each model from the database is calculated and recognized by the AdaBoost classifier. The decision is made in favor of the class of the model with the highest confidence. In fact, it is the same NN rule, but the distance function is the AdaBoost confidence. A similar approach to the face verification task allowed achieving the best-known accuracy for the Labeled Faces in the Wild (LFW) dataset (Zhou *et al*., 2015). In this paper, the CNN was trained based on an external independent dataset. This CNN was used as a feature extractor. The outputs of the last layer of the CNN for the query image and every model from the database were matched with the Euclidean distance. Thus, practically all recognition methods for the case $C \approx R$ are implemented as a special case of the k-NN classifier.

3. Materials and methods

Let the query object X be represented as a sequence of K regular (homogeneous) parts by any segmentation procedure (Theodoridis and Koutroumbas, 2008): $X = \{X(k)|k \in \{1,\dots,K\}\}$. Every k-th segment $X(k) = \{\mathbf{x}_j(k)|j \in \{1,\dots,n(k)\}\}$ is put in correspondence with a sequence of (primitive) feature vectors $\mathbf{x}_j(k) = \{x_{j;1}(k),\dots,x_{j;M}(k)\}$ with fixed dimension M, where $n(k)$ is the number of features in the k-th segment. Similarly, every model X_r is represented as a sequence $X_r = \{X_r(k)|k \in \{1,\dots,K_r\}\}$ of K_r segments and the k-th segment is defined as $X_r(k) = \{\mathbf{x}_j^{(r)}(k)|j \in \{1,\dots,n_r(k)\}\}$ of feature vectors $\mathbf{x}_j^{(r)}(k)$ with the same dimension M. Here $n_r(k)$ is the number of features in the k-th segment of the r-th model. As the procedure of automatic segmentation may be inaccurate, the segment $X(k)$ should be compared with a set $N_r(k)$ of segments of the r-th model. This neighborhood is determined for a specific task individually.

To apply statistical approach, we assume the following:

1. Vectors $\mathbf{x}_j(k)$, $\mathbf{x}_j^{(r)}(k)$ are *multivariate random variables*.

2. Segments $X(k)$ and $X_r(k)$ are random samples of i.i.d. feature vectors $\mathbf{x}_j(k)$ and $\mathbf{x}_j^{(r)}(k)$, respectively.

There are two possible approaches to estimate unknown class densities, namely, *parametric* and *nonparametric* (Theodoridis and Koutroumbas, 2008; Rutkowski, 2008). First, the parametric approach is discovered in detail. It is assumed that distributions of vectors $\mathbf{x}_j(k)$ and $\mathbf{x}_j^{(r)}(k)$ are of the multivariate exponential type $f_{\boldsymbol{\theta};n}(\widetilde{X})$ generated by a fixed (for all classes) function $f_0(\widetilde{X})$ with a p-dimensional parameter vector $\boldsymbol{\theta}$ (Kullback, 1997):

$$f_{\boldsymbol{\theta};n}(\widetilde{X}) = \exp(\boldsymbol{\tau}(\boldsymbol{\theta}) \cdot \widehat{\boldsymbol{\theta}}(\widetilde{X}))\frac{f_0(\widetilde{X})}{M(\boldsymbol{\tau})}, \tag{1}$$

where $\widehat{\boldsymbol{\theta}}(\widetilde{X})$ is an estimate of parameter $\boldsymbol{\theta}$ using available data (random sample) $\widetilde{X}$ of size n,

$$M(\boldsymbol{\tau}) = \int \exp(\boldsymbol{\tau}(\boldsymbol{\theta}) \cdot \widehat{\boldsymbol{\theta}}(\widetilde{X}))f_0(\widetilde{X})\,\mathrm{d}\widetilde{X}, \tag{2}$$

and $\boldsymbol{\tau}(\boldsymbol{\theta})$ is a normalizing function (p-dimensional parameter vector) defined by the following equation if the parameter estimation $\widehat{\boldsymbol{\theta}}(\widetilde{X})$ is unbiased (for details, see Kullback, 1997):

$$\int \widehat{\boldsymbol{\theta}}(\widetilde{X})f_{\boldsymbol{\theta};n}(\widetilde{X})\,\mathrm{d}\widetilde{X} \equiv \frac{\mathrm{d}}{\mathrm{d}\tau(\boldsymbol{\theta})}\ln M(\boldsymbol{\tau}) = \boldsymbol{\theta}. \tag{3}$$

Each r-th class of each k-th segment is determined by a parameter vector $\boldsymbol{\theta}_r(k)$. This assumption about the exponential family $f_{\widehat{\boldsymbol{\theta}}(X_r(k));n(k)}(X(k))$ in which parameter $\boldsymbol{\theta}_r(k)$ is estimated by using the observed (given) sample $X_r(k)$ covers many known distributions, e.g., polynomial, normal, etc.

In this paper we focus on the case of full prior uncertainty and assume that the prior probabilities of each

class are equal. In this case, if the recognition task is reduced to a problem of statistically testing a simple hypothesis, the Bayesian approach will be equivalent to the maximum likelihood criterion (Borovkov, 1998):

$$\max_{r\in\{1,\dots,R\}}\max_{k_r\in N_r(k)} f_{\widehat{\boldsymbol{\theta}}(X_r(k_r));n(k)}(X(k)). \tag{4}$$

It can be shown that (4) is equivalent to the minimum information discrimination rule (Kullback, 1997),

$$\min_{r\in\{1,\dots,R\}}\sum_{k=1}^{K} n(k)\min_{k_r\in N_r(k)}\widehat{I}(f_{\widehat{\boldsymbol{\theta}}(X(k))}:f_{\widehat{\boldsymbol{\theta}}(X_r(k_r))}), \tag{5}$$

where

$$\widehat{I}(f_{\widehat{\boldsymbol{\theta}}(X(k))}:f_{\widehat{\boldsymbol{\theta}}(X_r(k_r))}) = \int f_{\widehat{\boldsymbol{\theta}}(X(k))}(x)\ln\frac{f_{\widehat{\boldsymbol{\theta}}(X(k))}(\mathbf{x})}{f_{\widehat{\boldsymbol{\theta}}(X_r(k_r))}(\mathbf{x})}\,\mathrm{d}\mathbf{x} \tag{6}$$

is the KL divergence between the densities $f_{\widehat{\boldsymbol{\theta}}(X(k))}\equiv f_{\widehat{\boldsymbol{\theta}}(X(k));1}$ and $f_{\widehat{\boldsymbol{\theta}}(X_r(k_r))}\equiv f_{\widehat{\boldsymbol{\theta}}(X_r(k_r));1}$.

However, the criterion (5) is not correct as the true densities of segments of each class are not known and unbiased estimates of parameters $\boldsymbol{\theta}_r(k)$ should be used. In fact, the pattern recognition problem should be reduced to the statistical testing of *complex* hypothesis of samples homogeneity (Borovkov, 1998):

$$W_r(k;k_r): X(k) \text{ and } X_r(k_r) \text{ are homogeneous.} \tag{7}$$

The maximum likelihood decision of this problem,

$$\max_{r\in\{1,\dots,R\}}\max_{k_r\in N_r(k)}\sup_{\substack{\overline{\boldsymbol{\theta}}(k),\\ \overline{\boldsymbol{\theta}}_j(k_r),j\in\{1,\dots,R\}}} f(\{X(k),X_1(k_r),\dots,X_R(k_r)\}|W_r(k;k_r)), \tag{8}$$

is known to be asymptotically (if the size of the segment, i.e., the image resolution or the phoneme duration, is large) equivalent to the minimax criterion (Borovkov, 1998). Here $\overline{\boldsymbol{\theta}}(k)$ are the possible parameters of $X(k)$, $\overline{\boldsymbol{\theta}}_j(k_r)$ are the possible parameters of model $X_j(k_r)$, $f(\{X(k),X_1(k_r),\dots,X_R(k_r)\}|W_r(k;k_r))$ is the joint probability distribution of united sample $\{X(k),X_1(k_r),\dots,X_R(k_r)\}$ if the hypothesis $W_r(k;k_r)$ is true. Then we have the following result.

Theorem 1. *If $\widehat{\boldsymbol{\theta}}(\widetilde{X})$ is an unbiased maximum likelihood estimate of the parameter vector $\boldsymbol{\theta}$ in the distribution of the exponential type (1)–(3), then*

$$\min_{r\in\{1,\dots,R\}}\sum_{k=1}^{K}\min_{k_r\in N_r(k)}\Big(n(k)\widehat{I}(f_{\widehat{\boldsymbol{\theta}}(X(k))}:f_{\widehat{\boldsymbol{\theta}}(X_r(k;k_r))}) + n_r(k_r)\widehat{I}(f_{\widehat{\boldsymbol{\theta}}(X_r(k_r))}:f_{\widehat{\boldsymbol{\theta}}(X_r(k;k_r))})\Big) \tag{9}$$

is the asymptotically minimax criterion of testing hypothesis of samples homogeneity (7), where $X_r(k;k_r)=\{X(k),X_r(k_r)\}$ is the united sample of the query segment $X(k)$ and the model segment $X_r(k_r)$ while the KL divergences $\widehat{I}(f_{\widehat{\boldsymbol{\theta}}(X(k))}:f_{\widehat{\boldsymbol{\theta}}(X_r(k;k_r))}),\widehat{I}(f_{\widehat{\boldsymbol{\theta}}(X_r(k_r))}:f_{\widehat{\boldsymbol{\theta}}(X_r(k;k_r))})$ are defined in much the same way as in (6).

Proof. First, note that all vectors in the set $\{X(k),X_1(k_r),\dots,X_R(k_r)\}$ are independent. Hence, the likelihood function in (8) can be written as

$$\begin{aligned}&\sup_{\substack{\overline{\boldsymbol{\theta}}(k),\\ \overline{\boldsymbol{\theta}}_j(k_r),j\in\{1,\dots,R\}}} f(\{X(k),\dots,X_R(k_r)\}|W_r(k;k_r))\\ &\quad=\sup_{\overline{\boldsymbol{\theta}}(k)} f(X(k)|W_r(k;k_r))\\ &\qquad\times\prod_{j=1}^{R}\sup_{\overline{\boldsymbol{\theta}}_j(k_r)} f(X_j(k_r)|W_r(k;k_r)).\end{aligned} \tag{10}$$

If the hypothesis $W_r(k;k_r)$ is true, i.e., segments $X(k)$ and $X_r(k_r)$ are homogeneous, then the conditional density of $X_j(k_r)$ does not depend on the r-th class when $j\neq r$. In such a case, Eqn. (10) can be represented as

$$\begin{aligned}&\sup_{\substack{\overline{\boldsymbol{\theta}}(k),\\ \overline{\boldsymbol{\theta}}_j(k_r),j\in\{1,\dots,R\}}} f(\{X(k),\dots,X_R(k_r)\}|W_r(k;k_r))\\ &=\sup_{\overline{\boldsymbol{\theta}}(k)} f(X(k)|W_r(k;k_r))\sup_{\overline{\boldsymbol{\theta}}_r(k_r)} f(X_r(k_r)|W_r(k;k_r))\\ &\quad\times\prod_{\substack{j=1\\ j\neq r}}^{R}\sup_{\overline{\boldsymbol{\theta}}_j(k_r)} f_{\overline{\boldsymbol{\theta}}_j(k_r);n_j(k_r)}(X_j(k_r))\\ &=\frac{\sup_{\overline{\boldsymbol{\theta}}(k)} f(X(k)|W_r(k;k_r))\sup_{\overline{\boldsymbol{\theta}}_r(k_r)} f(X_r(k_r)|W_r(k;k_r))}{\sup_{\overline{\boldsymbol{\theta}}_r(k_r)} f_{\overline{\boldsymbol{\theta}}_r(k_r);n_r(k_r)}(X_r(k_r))}\\ &\quad\times\prod_{j=1}^{R}\sup_{\overline{\boldsymbol{\theta}}_j(k_r)} f_{\overline{\boldsymbol{\theta}}_j(k_r);n_j(k_r)}(X_j(k_r)).\end{aligned} \tag{11}$$

We divide (11) by

$$\sup_{\overline{\boldsymbol{\theta}}(k)} f_{\overline{\boldsymbol{\theta}}(k);n(k)}(X(k))\prod_{j=1}^{R}\sup_{\overline{\boldsymbol{\theta}}_j(k_r)} f_{\overline{\boldsymbol{\theta}}_j(k_r);n_j(k_r)}(X_j(k_r)),$$

which does not depend on r. The united sample $X_r(k;k_r)$ should be used to estimate $f(X(k)|W_r(k;k_r)), f(X_r(k_r)|W_r(k;k_r))$. The supremum in (11) is reached for a maximal likelihood estimate of $\boldsymbol{\theta}$. If $\widehat{\boldsymbol{\theta}}(\widetilde{X})$ is a unbiased maximum likelihood

estimate, the criterion (8) can be simplified:

$$\max_{r\in\{1,\dots,R\}}\max_{k_r\in N_r(k)}\frac{f_{\widehat{\boldsymbol{\theta}}(X_r(k;k_r));n(k)}(X(k))}{f_{\widehat{\boldsymbol{\theta}}(X(k));n(k)}(X(k))}\times\frac{f_{\widehat{\boldsymbol{\theta}}(X_r(k;k_r));n_r(k_r)}(X_r(k_r))}{f_{\widehat{\boldsymbol{\theta}}(X_r(k_r));n_r(k_r)}(X_r(k_r))}. \quad (12)$$

This expression is converted to the final form (9) by using the transformation of the criterion (4) to (5). ■

Thus, the criteria (5) and (9) are an implementation of the parametric approach for our probabilistic model of a piecewise-regular object. They can be implemented very efficiently since computation of the KL divergence usually requires $O(p^m)$ operations. For instance, $m = 1$ for the polynomial distribution and $m = 3$ for the p-variate normal distribution. Hence, the runtime complexity of the criteria (5) and (9) is equal to $O(p^m \cdot \sum_{r=1}^{R}\sum_{k=1}^{K}|N_r(k)|)$, where $|N_r(k)|$ is the size of the set $N_r(k)$.

The major assumption here was about the exponential family of distributions (1)–(3). Unfortunately, this assumption is not valid for arbitrary objects. Thus, a nonparametric approach is more popular nowadays (Rutkowski, 2008). We use the well-known Parzen estimates of probabilistic densities with Gaussian kernel $K(\cdot)$ (Specht, 1990). In such a case the maximum likelihood rule for the statistical testing of a *simple* hypothesis about the distributions of segments can be written as

$$\max_{r\in\{1,\dots,R\}}\sum_{k=1}^{K}\max_{k_r\in N_r(k)}\frac{1}{(n_r(k_r))^{n(k)}}\times\prod_{j=1}^{n(k)}\sum_{j_r=1}^{n_r(k_r)}K(\mathbf{x}_j(k),\mathbf{x}_{j_r}^{(r)}(k_r)). \quad (13)$$

The expression (13) is a generalization of the conventional PNN (Specht, 1990) for piecewise-regular objects. However, a proper way to perform classification of segments is to test their homogeneity.

Theorem 2. *If prior probabilities of all classes are equal, vectors $\mathbf{x}_j(k)$ and $\mathbf{x}_j^{(r)}(k)$ are i.i.d. random vectors with unknown densities which can be estimated with the Gaussian–Parzen kernel with fixed (for all classes) smoothing parameter σ, then the criterion*

$$\max_{r\in\{1,\dots,R\}}\sum_{k=1}^{K}\max_{k_r\in N_r(k)}\frac{n(k)^{n(k)}\cdot(n_r(k_r))^{n_r(k_r)}}{(n(k)+n_r(k_r))^{n(k)+n_r(k_r)}}\times\prod_{j=1}^{n(k)}\left(1+\frac{\sum_{j_r=1}^{n_r(k_r)}K(\mathbf{x}_j(k),\mathbf{x}_{j_r}^{(r)}(k))}{\sum_{j_1=1}^{n(k)}K(\mathbf{x}_j(k),\mathbf{x}_{j_1}(k))}\right)\times\prod_{j_r=1}^{n_r(k_r)}\left(1+\frac{\sum_{j_1=1}^{n(k)}K(\mathbf{x}_{j_r}^{(r)}(k_r),\mathbf{x}_{j_1}(k))}{\sum_{j_{r;1}=1}^{n_r(k_r)}K(\mathbf{x}_{j_r}^{(r)}(k_r),\mathbf{x}_{j_{r;1}}^{(r)}(k_r))}\right) \quad (14)$$

is the asymptotically minimax criterion of testing the complex hypothesis of samples homogeneity.

The proof is very similar to that of Theorem 1.

Unfortunately, the expressions (13) and (14) require a comparison of *all* features of *all* segments of *all* models. Their runtime complexity can be written as $O(M\sum_{r=1}^{R}\sum_{k=1}^{K}\sum_{k_r\in N_r(k)}n(k)n_r(k_r))$, i.e., they are much less computationally efficient than the parametric case (5), (9). Thus, the practical implementation of these rules may be unfeasible.

It is known (Savchenko, 2013b) that these criteria can be simplified if the feature vectors are discrete and certain, i.e., their domain of definition is a set $\{\mathbf{x}_1,\dots,\mathbf{x}_N\}$, where N is the number of different vectors. In such a case, the segment of the query object $X(k)$ can be described with the histogram $H(k) = \{h_1(k),\dots,h_N(k)\}$. Similarly, the model segment $X_r(k)$ can be described with the histogram $H_r(k) = \{h_1^{(r)}(k),\dots,h_N^{(r)}(k)\}$. This definition allows using the polynomial distribution, which is known to be of the exponential type (Kullback, 1997). Hence, it can be shown that the criterion (5) is equivalent to the minimum information discrimination principle,

$$\min_{r\in\{1,\dots,R\}}\sum_{k=1}^{K}\min_{k_r\in N_r(k)}\sum_{i=1}^{N}h_i(k)\ln\frac{h_i(k)}{h_i^{(r)}(k_r)}. \quad (15)$$

Similarly, the parametric criterion (9) based on homogeneity testing is equivalent to

$$\min_{r\in\{1,\dots,R\}}\sum_{k=1}^{K}\min_{k_r\in N_r(k)}\sum_{i=1}^{N}(n(k)h_i(k)\ln\frac{h_i(k)}{\widetilde{h}_{\Sigma;i}^{(r)}(k;k_r)}+n_r(k_r)h_i^{(r)}(k)\ln\frac{h_i^{(r)}(k)}{\widetilde{h}_{\Sigma;i}^{(r)}(k;k_r)}), \quad (16)$$

where

$$\widetilde{h}_{\Sigma;i}^{(r)}(k;k_r)=\frac{(n(k)h_i(k)+n_r(k_r)h_i^{(r)}(k_r))}{(n(k)+n_r(k_r))}.$$

If $n(k) = n_r(k_r)$, this criterion is equivalent to the NN rule with the Jensen–Shannon (JS) divergence widely used in various pattern recognition tasks (Martins *et al.*, 2008).

At the same time, if the nonparametric approach is used, an obvious generalization of the PNN (13) can be represented in the following form:

$$\min_{r\in\{1,\ldots,R\}}\sum_{k=1}^{K}\min_{k_r\in N_r(k)}\sum_{i=1}^{N}h_i(k)\ln\frac{h_{K;i}(k)}{h^{(r)}_{K;i}(k_r)}, \quad (17)$$

where

$$h_{K;i}(k)=\sum_{j=1}^{N}K_{ij}h_i(k)$$

and

$$h^{(r)}_{K;i}(k)=\sum_{j=1}^{N}K_{ij}h^{(r)}_i(k)$$

are the convolutions of the histograms with kernel $K_{ij}=K(\mathbf{x}_i,\mathbf{x}_j)$. Similarly, the proposed homogeneity testing for discrete patterns is implemented in the following way:

$$\min_{r\in\{1,\ldots,R\}}\sum_{k=1}^{K}\min_{k_r\in N_r(k)}\sum_{i=1}^{N}(n(k)h_i(k)\ln\frac{h_{K;i}(k)}{\widetilde{h}^{(r)}_{\Sigma;K;i}(k;k_r)} + n_r(k_r)h^{(r)}_i(k)\ln\frac{h^{(r)}_{K;i}(k)}{\widetilde{h}^{(r)}_{\Sigma;K;i}(k;k_r)}), \quad (18)$$

where

$$\widetilde{h}^{(r)}_{\Sigma;K;i}(k;k_r)=\frac{(n(k)h_{K;i}(k)+n_r(k_r)h^{(r)}_{K;i}(k_r))}{(n(k)+n_r(k_r))}.$$

In fact, the expressions (15), (16) are a special case of (17), (18) if a discrete delta function is used as a kernel, i.e., $\sigma\to\infty$. The runtime complexity of (17), (18) is $O(N\sum_{r=1}^{R}\sum_{k=1}^{K}|N_r(k)|)$, i.e., the computing efficiency is on average n^2M/N-times higher than the efficiency of (13) and (14), where $n=\sum_{k=1}^{K}n(k)/K$. It is obvious that our homogeneity-testing approach (9), (14), (16) is twice slower in comparison with the conventional statistical criteria (5), (13) and (15), respectively.

4. Experimental results

4.1. Image recognition. Let a set of $R>1$ grayscale images $\{X_r\}, r\in\{1,\ldots,R\}$ with width U_r and height V_r be given. In image recognition it is required to assign a query image X with width U and height V to one of R classes specified by these reference images. First, every image is put in correspondence with a set of feature descriptors (Theodoridis and Koutroumbas, 2008). The common part of most of modern algorithms is to divide the whole neighborhood into a regular grid of $S_1\times S_2$ blocks, S_1 rows and S_2 columns (in our previous notation, $K=K_1=K_2=\cdots=K_R=S_1S_2$), and separately evaluate the N bins-histogram $H^{(r)}(s_1,s_2)=\{h^{(r)}_1(s_1,s_2),\ldots,h^{(r)}_N(s_1,s_2)\}$ of widely used gradient orientation features for each block $(s_1,s_2), s_1\in\{1,\ldots,S_1\}, s_2\in\{1,\ldots,S_2\}$ of the reference image X_r (Dalal and Triggs, 2005; Lowe, 2004). The same procedure is repeated to evaluate the histograms of oriented gradients (HOG) $H(s_1,s_2)=\{h_1(s_1,s_2),\ldots,h_N(s_1,s_2)\}$ based on the query image X.

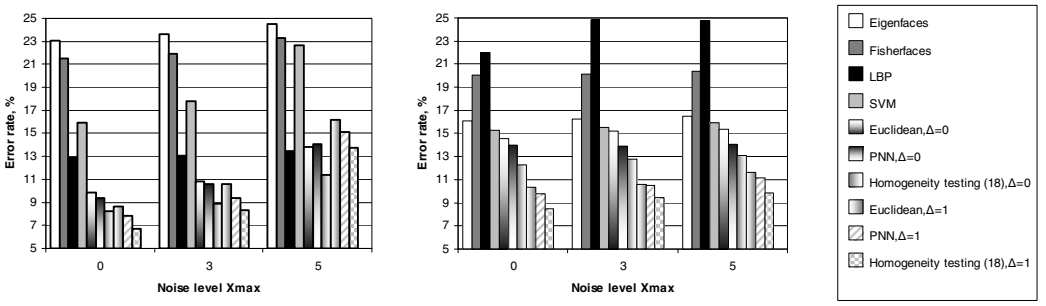
The second part is classifier design. If $C\approx R$ and the number of classes and the feature vector size are large, state-of-the-art classifiers (MLP, SVM, etc.) do not outperform the NN (Tan *et al.*, 2006). In view of the small spatial deviations due to misalignment after object detection, the following similarity measure with a mutual alignment and the matching of HOGs in the Δ-neighborhood of each block is used:

$$\min_{r\in\{1,\ldots,R\}}\sum_{s_1=1}^{S_1}\sum_{s_2=1}^{S_2}\min_{|\Delta_1|\le\Delta,|\Delta_2|\le\Delta}\rho_H(H(s_1+\Delta_1,s_2+\Delta_2),H^{(r)}(s_1,s_2)). \quad (19)$$

Here $\rho_H(H(s_1+\Delta_1,s_2+\Delta_2),H^{(r)}(s_1,s_2))$ is an arbitrary distance between HOGs $H(s_1+\Delta_1,s_2+\Delta_2)$ and $H^{(r)}(s_1,s_2)$, and neighborhood $N_r(k)$ in all criteria from the previous section for the cell (s_1,s_2) is described with the set $\{(\widetilde{s}_1,\widetilde{s}_2)\,|\,|\widetilde{s}_1-s_1|\le\Delta,|\widetilde{s}_2-s_2|\le\Delta\}$. In this paper, we explore the square of the Euclidean distance and describe distances based on the statistical approach, namely, KL (15), JS (16), the PNN (17) and the proposed criterion (18) on the basis of segment homogeneity testing. Additionally, we use the state-of-the-art SVM classifiers of HOGs and face recognition methods from the OpenCV library, namely, eigenfaces, fisherfaces and LBP histograms (Theodoridis and Koutroumbas, 2008).

In the first experiment we deal with face recognition with the FERET and AT&T datasets. From FERET 2720 frontal facial images of $C=994$ persons were selected. The training set contains $R=1370$ images (from 1 to 6 photos per person). The test set consists of other 1350 photos, i.e., from 1 to 4 faces per individual. The AT&T database contains 400 faces of $C=40$ persons; $R=80$ of them (2 photos per each person) formed the training set. The test set contains other 320 photos (8 images per individual).

These datasets contain a small number of models per person. Hence, we decided to avoid the standard methodology of tuning the parameters by splitting the whole dataset into training, validation and testing sets. Instead of this procedure, we used the large Essex face database (7900 images, 395 persons). In fact, a similar idea is popular in training DNN based face recognizers (Zhou *et al.*, 2015). The conventional 10-fold cross-validation was applied to obtain the following values of parameters. The median filter with a window



(a) (b)

Fig. 1. Error rate [%]: FERET dataset (a), AT&T dataset (b).

size (3×3) was applied to remove noise in detected faces. The following neighborhood sizes were tested: $\Delta = 0$ and $\Delta = 1$. All facial images are divided into regular segments (blocks) by a 12×12 grid ($S_1 = S_2 = 12$) if $\Delta = 1$ and by 6×6 grid ($S_1 = S_2 = 6$) if $\Delta = 0$. The number of bins in the gradient orientation histogram $N = 8$. The Gaussian kernel smoothing parameter $\sigma = 0.71$.

To measure the influence of the noise presence, we artificially added a random noise from the range $[-x_{\max}; x_{\max}]$ to *each* pixel of the image from the test set, where $x_{\max} \in \{0, 3, 5\}$. The error rate was estimated by repeating random sub-sampling cross-validation 100 times. The dependence of the estimated error rates on $x_{\max}$ for the FERET and AT&T datasets is shown in Fig. 1. The macro-averaging of recall and precision estimated for each class by one-vs-all procedure for $x_{\max} = 0$ and $\Delta = 1$ is presented in Tables 1 and 2. As the number of images per person in the test sample for the AT&T dataset is equal to 8 for all classes, macro-averaging recall here is equal to micro-averaging accuracy. However, accuracy and macro-averaging recall are slightly different for the FERET dataset, in which for most classes only one sample per person is available in the test set.

Here the quality of the conventional eigenfaces and SVM is appropriate only for the simpler AT&T dataset. For instance, the accuracy of eigenfaces is 15% higher when compared with the criterion (19) for the FERET dataset. However, if we repeat our experiment with the AT&T dataset but put into the training set $R = 160$ models, the situation will be changed. SVM achieves a 5.5% error rate, which is 2.1% lower than the NN rule with the Euclidean distance and even 1.2% lower than the error rate of the proposed approach without alignment ($\Delta = 0$). At the same time, even in this case the accuracy of the proposed approach with alignment of HOGs ($\Delta = 1$) is equal to 97.7%, which is 2.2% better than for the SVM. Secondly, the error rate of the traditional NN rule ($\Delta = 0$) with the Euclidean distance is too high.

Moreover, we confirmed that that of the PNN is less than the accuracy of the criterion based on homogeneity testing (18). According to McNemar's test with the confidence level of 0.05, this improvement of the proposed approach (18) is statistically significant. In fact, JS divergence is a special case of our similarity measure (18) if $\sigma \to 0$. Hence, our approach with segment homogeneity testing is much more robust to deviation of the smoothing parameter than the conventional PNN. Thirdly, the alignment of HOGs ($\Delta = 1$) is characterized by statistically significant higher accuracy than the conventional approach ($\Delta = 0$) in the case of small noise ($x_{\max} \leq 3$). Unfortunately, this alignment leads to worse performance: the traditional distance computation ($\Delta = 0$) is 9 ($(2 \cdot 1 + 1)^2$) times faster than HOGs alignment ($\Delta = 1$). For the more complex FERET

Table 1. Face recognition quality: FERET dataset.

Algorithm	Recall [%]	Precision [%]	F1 score [%]
Eigenfaces	77.8±2.8	71.7±2.8	74.6±2.8
Fisherfaces	77.3±2.8	71.9±2.9	74.5±2.9
LBP	88.0±1.7	84.2±1.9	86.0±1.8
SVM	83.6±1.9	80.8±2.0	82.2±1.9
Euclidean	93.4±1.2	91.3±1.2	92.4±1.2
KL	93.6±1.3	91.4±1.2	92.5±1.3
PNN	94.1±1.3	92.5±1.2	93.3±1.3
JH	94.3±1.1	92.8±1.2	93.5±1.1
Homogeneity testing (18)	95.3±0.9	93.6±1.0	94.4±0.9

Table 2. Face recognition quality: AT&T dataset.

Algorithm	Recall [%]	Precision [%]	F1 score [%]
Eigenfaces	83.9±1.6	86.7±1.5	85.3±1.6
Fisherfaces	79.9±2.1	85.9±1.6	82.8±1.9
LBP	75.2±2.3	80.9±2.1	78.0±2.2
SVM	84.7±1.5	87.7±1.3	86.1±1.4
Euclidean	89.7±1.3	91.5±1.1	90.6±1.2
KL	89.8±1.2	91.2±1.2	90.5±1.2
PNN	90.2±1.2	91.6±1.2	90.9±1.2
JH	91.1±1.1	92.7±1.0	91.9±1.1
Homogeneity testing (18)	91.5±1.0	93.1±1.0	92.3±1.0

dataset and high noise level $x_{\max} = 5$, the application of the alignment significantly *decreases* the recognition rate. Addition of large noise makes the estimated distribution of the gradient orientation (i.e., HOG) similar to HOGs of many other blocks. It is necessary to use simple classifiers (e.g., the criterion (19) with $\Delta = 0$) if the available training set is not representative.

In the last experiment we discover the viseme classification problem to show the potential of our approach in another application of the image recognition criterion (18), (19). This task usually appears in audio-visual ASR systems (Asadpour *et al.*, 2011). A viseme is a visual representation of a phoneme pronounced by a speaker. We collected 500 photos of 7 most important Russian visemes (pause and stressed vowels /AA/, /EE/, /II/, /OO/, /UU/, /Y/) taken by two Kinect cameras, namely, a normal camera and a depth sensor. The mouth region was detected with the OpenCV library.

Along with the conventional, in audio-visual recognition PCA features classified with the SVM and the NN rule, we used SIFT descriptors (Lowe, 2004) and HOGs. Further details of this experiment can be found in our previous paper (Savchenko and Khokhlova, 2014).

The dependence of the error rate obtained by random subsampling cross-validation on the number of models per one viseme class is shown in Fig. 2. Classification of PCA features calculated for depth-maps is characterized with poor accuracy so we do not show it in this figure. A detail comparison of two best classifiers, namely, PCA+SVM (normal camera) and our approach (depth sensor) in the case of 50 images per class in the training set ($R = 350$), is presented in Table 3.

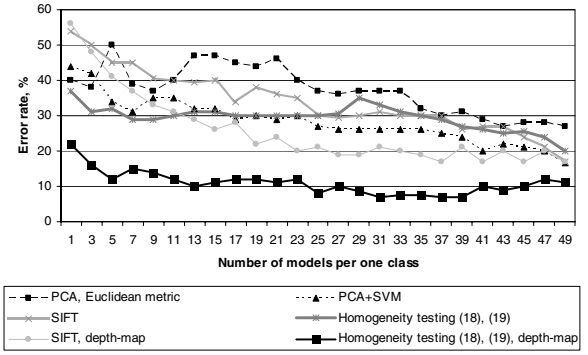


Fig. 2. Viseme recognition error rate [%].

Table 3. Quality of viseme recognition.

Viseme	PCA+SVM		Homogeneity testing (18), (19)	
	Recall [%]	Precision [%]	Recall [%]	Precision [%]
Pause	99±1.5	99±1.5	98±1.5	99±1.5
/AA/	61±4.8	99±1.5	99±1.5	99±1.5
/EE/	59±5.1	65±5.0	90±2.2	85±2.3
/II/	98±1.5	99±1.5	85±2.4	83±2.4
/OO/	85±2.9	79±3.2	98±1.5	97±1.5
/UU/	80±3.5	85±2.9	99±1.5	99±1.5
/Y/	98±1.5	55±5.2	67±4.5	75±4.4
Average	82.9±2.9	82.9±2.9	90.9±2.2	90.9±2.2

As expected, in this experiment the traditional, for audio-visual recognition PCA features are characterized with the higher accuracy if a normal camera is used and the number of models per class is large. However, our approach (19) based on hypothesis testing of segment homogeneity (18) shows the lowest error rate if the training set contains $R \leq 70$ models (Fig. 2). Moreover, the usage of a depth sensor allowed increasing the accuracy by 10–20%. In this case, the average viseme recognition quality is 8% higher than the best results achieved with conventional PCA+SVM (Table 3). At the same time, several visemes (e.g., /Y/ and /II/) are classified even worse when compared with recognition of normal images. Hence, a fusion of classifier outputs for normal images and depth-maps is a promising technique.

4.2. Voice command recognition. In this section we investigate the statistical (Bayesian) approach to ASR in a voice control application (Savchenko, 2013a). In this task, it is required to assign an utterance X to one of R commands $\{X_r\}$. Every r-th model is divided into a sequence of K_r syllables. Each syllable is put in correspondence with a code c of the vowel. Vowels should be specified by the model signals $\{\mathbf{x}_c\}, c \in \{1, \ldots, C\}$ pronounced by the speaker in isolated mode. Hence, the model phrase X_r is represented as a sequence of codes $\{c_{r,1}, \ldots, c_{r,K_r}\}$. Here, $c_{r,j} \in \{1, \ldots, C\}$ are the numbers of vowels from a given alphabet. To solve the ASR task, query utterance X is automatically divided into K syllables (Janakiraman *et al.*, 2010) and the vowel segment $X(k)$ is extracted from the k-th syllable (Pfau and Ruske, 1998). We assume that syllables are extracted without mistakes, e.g., the voice commands are produced in isolated syllable mode (Merialdo, 1988).

In such a case it is required to assign vowel $X(k)$ to one of C model phonemes. The parametric approach is much more popular in this particular task. It is assumed that phoneme distribution is either a Gaussian or a mixture of Gaussians (Benesty *et al.*, 2008). It is known that KL divergence is equivalent to the Itakura–Saito (IS) distance (Gray *et al.*, 1980) and the maximal likelihood estimate of the covariance matrix for a signal with a zero mean is unbiased. Thus, the parametric criterion (5) is written as follows:

$$\min_{r \in \{1,\ldots,R\}} \sum_{k=1}^{K} \rho_{IS}(X(k), \mathbf{x}_{c_{r,k}}). \tag{20}$$

If we assume that

$$\widehat{I}(f_{\widehat{\boldsymbol{\theta}}(X(k))} : f_{\widehat{\boldsymbol{\theta}}(X_r(k;k_r))}) \approx \widehat{I}(f_{\widehat{\boldsymbol{\theta}}(X(k))} : f_{\widehat{\boldsymbol{\theta}}(X_r(k_r))}), \widehat{I}(f_{\widehat{\boldsymbol{\theta}}(X_r(k_r))} : f_{\widehat{\boldsymbol{\theta}}(X_r(k;k_r))}) \approx \widehat{I}(f_{\widehat{\boldsymbol{\theta}}(X_r(k_r))} : f_{\widehat{\boldsymbol{\theta}}(X(k))}),$$

the consequence of Theorem 1 can be written as follows. If signals $X(k)$ and all model phonemes $\{\mathbf{x}_c\}, c \in \{1,\dots,C\}$ are generated with an autoregressive process of the p-th order and are normally distributed with zero mean and an unknown covariance matrix, then the asymptotically minimax criterion of testing for speech homogeneity can be written as

$$\min_{r\in\{1,\dots,R\}} \sum_{k=1}^{K} (\rho_{IS}(X(k), \mathbf{x}_{c_{r,k}}) + \rho_{IS}(\mathbf{x}_{c_{r,k}}, X(k))). \tag{21}$$

In the experiment, utterances were recorded in the following format: PCM wav, mono, sampling rate 8000 Hz, 16 bits per sample. The vocabulary contains 1913 Russian names of drugs sold in one pharmacy of Nizhny Novgorod. A total of 10 speakers (5 men and 5 women of different age) pronounced each word from this vocabulary twice in isolated syllable mode to simplify comparison of the distances (20) and (21). To train the system, each speaker pronounced 10 vowels in isolated mode. The conventional values of parameters (Benesty *et al.*, 2008) were chosen: frame length 30 ms, frame overlap 10 ms, autoregression model order p equal to 12. To estimate the closeness of speech signals, the conventional IS (20) and the proposed criterion (21) on the basis of segment homogeneity testing were used. To compare our method with the conventional approach to ASR, CMU Pocketsphinx was tested in speaker-dependent mode with MLLR (maximum likelihood linear regression) adaptation with the available phonetic database $\{\mathbf{x}_c\}$.

Finally, we applied a speaker-dependent mode of Pocketsphinx to recognize vowels in a syllable instead of the IS distance in (20) to demonstrate the superiority of our approach in the phoneme classification task. We added an artificially generated white noise to each test signal (with the signal-to-noise ratio (SNR) 25 dB, 15 dB, 10 dB). The error rates are shown in Fig. 3. Here the isolated syllable mode allowed increasing the accuracy by 3–6% for the conventional Pocketsphinx ASR. However, the usage of the user-specific phonetic database in (20) and (21) even decreases the error rate by 7–9%. Finally, the most valuable conclusion here is the achievement of a higher accuracy with the testing for the homogeneity of speech segments (21) when compared with the testing (20) for a simple hypothesis. McNemar's test verified that this improvement is significant in all cases.

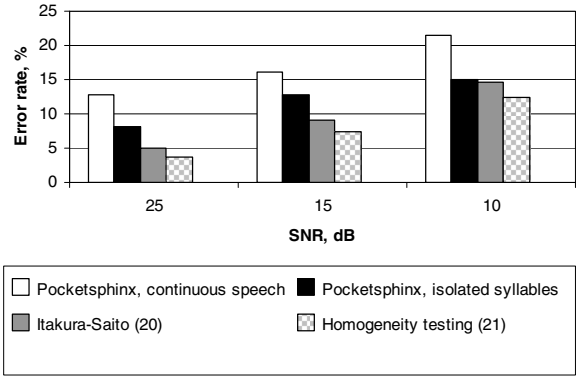


Fig. 3. ASR error rate [%].

5. Conclusion

In this paper we analyzed the methods of classification of piecewise-regular objects. The dependence of the classifier choice on the number of classes and models in the database was highlighted. Our brief survey showed that the current trends in the development of composite object recognition methods are connected with the refusal of complex algorithms of uncorrelated feature extraction and complication of the classifiers. We emphasized one of the most exciting challenges in this field, namely, a small number of models per each class. Most researchers are familiar with the insufficient accuracy of maximum likelihood criteria (e.g., (4), (15)) in this case. For instance, in Tables 2 and 3, the NN rule with the KL divergence (15) is not better than the conventional Euclidean distance in most cases. This issue is usually explained by the incorrect ("naive") assumption of statistical independence of features inside one segment (segment). However, in this paper we found another explanation. Namely, the recognition task should not be reduced to testing for a simple hypothesis with estimation of unknown densities based on the training set. We believe that this problem should be described in terms of testing for the homogeneity of a query object (or its segments for piecewise-regular objects) with available models (Borovkov, 1998). Our brief proposal of a rather simple statistical model of the composite object as a sequence of segments of i.i.d. feature vectors allowed us to present several statistical recognition criteria (4), (9), (13)–(18) for both parametric and nonparametric estimates of unknown class densities. The approach based on homogeneity testing (9), (16), (18) and (21) showed its superiority in our experimental results for various recognition tasks (Figs. 1–3) over the conventional testing for a simple hypothesis (4), (15), (17) and (20), respectively.

The main direction for further research of the proposed approach with testing of segment homogeneity

can be related to its application with modern features in various classification tasks. It is necessary to apply it with CNN based features (Zhou *et al*., 2015) in a face verification task with the LFW dataset. Another possible direction is the usage of modern approximate nearest neighbor methods (e.g., Savchenko, 2012), if the number of classes is high $(C \gg 1)$ to increase the recognition speed of exhaustive search (9), (14).

Acknowledgment

The work by A.V. Savchenko was conducted at the National Research University Higher School of Economics and supported by an RSF grant 14-41-00039.

References

Asadpour, V., Homayounpour, M.M. and Towhidkhah, F. (2011). Audio-visual speaker identification using dynamic facial movements and utterance phonetic content, *Applied Soft Computing* **11**(2): 2083–2093.

Benesty, J., Sondhi, M.M. and Huang, Y. (2008). *Springer Handbook of Speech Processing*, Springer, Berlin.

Borovkov, A.A. (1998). *Mathematical Statistics*, Gordon and Breach Science Publishers, Amsterdam.

Bottou, L., Fogelman Soulie, F., Blanchet, P. and Lienard, J. (1990). Speaker-independent isolated digit recognition: Multilayer perceptrons vs. dynamic time warping, *Neural Networks* **3**(4): 453–465.

Ciresan, D., Meier, U., Masci, J. and Schmidhuber, J. (2012). Multi-column deep neural network for traffic sign classification, *Neural Networks* **32**: 333–338.

Dalal, N. and Triggs, B. (2005). Histograms of oriented gradients for human detection, *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2005, San Diego, CA, USA*, pp. 886–893.

Gray, R., Buzo, A., Gray, A., Jr. and Matsuyama, Y. (1980). Distortion measures for speech processing, *IEEE Transactions on Acoustics, Speech and Signal Processing* **28**(4): 367–376.

Haykin, S.O. (2008). *Neural Networks and Learning Machines*, 3rd Edn., Prentice Hall, Harlow.

Hinton, G., Deng, L., Yu, D., Dahl, G., Mohamed, A., Jaitly, N., Senior, A., Vanhoucke, V., Nguyen, P., Sainath, T. and Kingsbury, B. (2012). Deep neural networks for acoustic modeling in speech recognition: The shared views of four research groups, *IEEE Signal Processing Magazine* **29**(6): 82–97.

Hinton, G.E., Osindero, S. and Teh, Y.-W. (2006). A fast learning algorithm for deep belief nets, *Neural Computation* **18**(7): 1527–1554.

Huang, J.-T., Li, J., Yu, D., Deng, L. and Gong, Y. (2013). Cross-language knowledge transfer using multilingual deep neural network with shared hidden layers, *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing, ICASSP 2013, Vancouver, BC, Canada*, pp. 7304–7308.

Janakiraman, R., Kumar, J. and Murthy, H. (2010). Robust syllable segmentation and its application to syllable-centric continuous speech recognition, *Proceedings of the National Conference on Communications, NCC 2010, Chennai, India*, pp. 1–5.

Kullback, S. (1997). *Information Theory and Statistics*, Dover Publications, New York, NY.

LeCun, Y., Bengio, Y. and Hinton, G. (2015). Deep learning, *Nature* **521**(7553): 436–444.

LeCun, Y., Bottou, L., Bengio, Y. and Haffner, P. (1998). Gradient-based learning applied to document recognition, *Proceedings of the IEEE* **86**(11): 2278–2324.

Liao, S., Zhu, X., Lei, Z., Zhang, L. and Li, S.Z. (2007). Learning multi-scale block local binary patterns for face recognition, *in* S.-W. Lee and S.Z. Li (Eds.), *Advances in Biometrics*, Lecture Notes in Computer Science, Vol. 4642, Springer, Berlin/Heidelberg, pp. 828–837.

Lowe, D.G. (2004). Distinctive image features from scale-invariant keypoints, *International Journal of Computer Vision* **60**(2): 91–110.

Martins, A.F.T., Figueiredo, M.A.T., Aguiar, P.M.Q., Smith, N.A. and Xing, E.P. (2008). Nonextensive entropic kernels, *Proceedings of the 25th International Conference on Machine Learning, ICML '2008, New York, NY, USA*, pp. 640–647.

Merialdo, B. (1988). Multilevel decoding for very-large-size-dictionary speech recognition, *IBM Journal of Research and Development* **32**(2): 227–237.

Pfau, T. and Ruske, G. (1998). Estimating the speaking rate by vowel detection, *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing, ICASSP 1998, Seattle, WA, USA*, Vol. 2, pp. 945–948.

Rutkowski, L. (2008). *Computational Intelligence: Methods and Techniques*, Springer-Verlag, Berlin/Heidelberg.

Sas, J. and Żołnierek, A. (2013). Pipelined language model construction for Polish speech recognition, *International Journal of Applied Mathematics and Computer Science* **23**(3): 649–668, DOI: 10.2478/amcs-2013-0049.

Savchenko, A.V. (2012). Directed enumeration method in image recognition, *Pattern Recognition* **45**(8): 2952–2961.

Savchenko, A.V. (2013a). Phonetic words decoding software in the problem of Russian speech recognition, *Automation and Remote Control* **74**(7): 1225–1232.

Savchenko, A.V. (2013b). Probabilistic neural network with homogeneity testing in recognition of discrete patterns set, *Neural Networks* **46**: 227–241.

Savchenko, A.V. and Khokhlova, Y.I. (2014). About neural-network algorithms application in viseme classification problem with face video in audiovisual speech recognition systems, *Optical Memory and Neural Networks (Information Optics)* **23**(1): 34–42.

Specht, D.F. (1990). Probabilistic neural networks, *Neural Networks* **3**(1): 109–118.

Świercz, E. (2010). Classification in the Gabor time-frequency domain of non-stationary signals embedded in heavy noise with unknown statistical distribution, *International Journal of Applied Mathematics and Computer Science* **20**(1): 135–147, DOI: 10.2478/v10006-010-0010-x.

Tan, X., Chen, S., Zhou, Z.-H. and Zhang, F. (2006). Face recognition from a single image per person: A survey, *Pattern Recognition* **39**(9): 1725–1745.

Theodoridis, S. and Koutroumbas, K. (2008). *Pattern Recognition*, 4th Edn., Academic Press, Burlington, MA/London.

Zhou, E., Cao, Z. and Yin, Q. (2015). Naive-deep face recognition: Touching the limit of LFW benchmark or not?, *CoRR* **abs/1501.04690**.

Andrey V. Savchenko graduated from Nizhny Novgorod State Technical University in 2008. He received the Ph.D. degree in mathematical modeling at the State University Higher School of Economics in 2010. Now he is an associate professor of the National Research University Higher School of Economics—Nizhny Novgorod, and a researcher of the Laboratory of Algorithms and Technologies for Networks Analysis at the NRU HSE.

Natalya S. Belova received her M.Sc. and Ph.D. degrees at the Moscow State University of Instrument Engineering and Computer Science in 2005 and 2009, respectively. She currently works as an associate professor in the Faculty of Computer Science at the NRU HSE.

Received: 1 November 2014
Revised: 25 March 2015