An Optimal Greedy Approximate Nearest Neighbor Method in Statistical Pattern Recognition

Andrey V. Savchenko(✉)

Laboratory of Algorithms and Technologies for Network Analysis,
National Research University Higher School of Economics,
Nizhny Novgorod, Russia
avsavchenko@hse.ru

Abstract. The insufficient performance of statistical recognition of composite objects (images, speech signals) is explored in case of medium-sized database (thousands of classes). In contrast to heuristic approximate nearest-neighbor methods we propose a statistically optimal greedy algorithm. The decision is made based on the Kullback-Leibler minimum information discrimination principle. The model object to be checked at the next step is selected from the class with the maximal likelihood (joint density) of distances to previously checked models. Experimental study results in face recognition task with FERET dataset are presented. It is shown that the proposed method is much more effective than the brute force and fast approximate nearest neighbor algorithms, such as randomized kd-tree, perm-sort, directed enumeration method.

Keywords: Statistical pattern recognition · Approximate nearest neighbor · Kullback-Leibler divergence · Directed enumeration method · Face recognition

1 Introduction

The problem of small sample size is crucial in pattern recognition of complex objects (e.g., images) [1]. In fact, most of known algorithms in this case are equivalent to the nearest neighbor (NN) method with appropriate similarity measure [2]. If the number of classes is large (hundreds or even thousands of classes), the performance of NN's exhaustive search is not enough for real-time processing. It seems, conventional fast approximate NN image *retrieval* methods [3] can be applied, e.g. AESA (Approximating and Eliminating Search Algorithm) [4], composite kd-tree [5], randomized kd-tree [6], recent variations of Locality-Sensitive Hashing [7], etc. Unfortunately, these techniques usually cannot be efficiently used in *recognition* tasks as the latter are significantly different from retrieval in terms of

(1) quality indicators (accuracy in recognition and recall in retrieval): 3–5 % losses in accuracy/recall of retrieval techniques are inappropriate for many recognition tasks;

(2) similarity measures in recognition tasks are much more complex [8] in comparison with conventional Minkowski or cosine distance in retrieval. Image retrieval

M. Kryszkiewicz et al. (Eds.): PReMI 2015, LNCS 9124, pp. 236–245, 2015.
DOI: 10.1007/978-3-319-19941-2_23

methods are with similarity measures which satisfy metric properties (sometimes, triangle inequality and, usually, symmetry) [4, 9]. They are known to show good performance only if the first NN is quite different from other models;

(3) classification methods (1-NN in recognition and k-NN in retrieval);

(4) database size (medium in recognition and very-large in retrieval). Performance of approximate NN algorithms is comparable with brute-force for medium-sized training sets (thousands of classes). To decrease the recognition speed for such training sets, other methods, e.g., ordering permutations (perm-sort) [10] and directed enumeration method (DEM) [11] has recently been proposed.

Final issue is the heuristic nature of most approximate NN methods. It is usually impossible to prove that particular algorithm is optimal and nothing can be done to improve it. In this paper we propose an alternative solution on the basis of the statistical approach - while looking for the NN for particular query object, conditional probability of belonging of previously checked models to each class is estimated. The next model from the database is selected from the class with the maximal probability.

The rest of the paper is organized as follows. In Sect. 2 we recall the Kullback-Leibler minimum discrimination principle [12] in statistical pattern recognition. In Subsect. 2.2 we briefly review the baseline method (DEM). In Sect. 3 the novel Maximum-Likelihood DEM (ML-DEM) is proposed. In Sect. 4 experimental study results are presented in face recognition with FERET dataset. Finally, concluding comments are given in Sect. 5.

2 Materials and Methods

2.1 Statistical Pattern Recognition

In the pattern recognition task it is required to assign the query object X to one of $R > 1$ classes [2]. Each class is specified by the given model object X_r, $r \in \{1, \ldots, R\}$. First stage is feature extraction. In this paper we use the statistical approach and assume that each class is characterized with its own probabilistic distribution of appropriate features. Let's focus on the most popular discrete case, in which the features can take $N > 1$ different values. Hence, the distribution of rth class is defined as a histogram $H_r = [h_{r;1}, h_{r;2}, \ldots, h_{r;N}]$ estimated based on the X_r. The same procedure of histogram evaluation $H = [h_1, h_2, \ldots, h_N]$ is repeated for the query object X.

If the prior probabilities of each class are equal, the maximal likelihood criterion [2] can be used to test statistical hypothesis W_r, $r \in \{1, \ldots, R\}$ about distribution H:

$$\max_{r \in \{1,\ldots,R\}} f_r(X), \tag{1}$$

where the likelihood of rth class $f_r(X)$ is estimated as follows

$$f_r(X) = \prod_{i=1}^{N} (h_{r;i})^{n \cdot h_i}. \tag{2}$$

Here it is assumed that the query object X contains n simple features to estimate the histogram H. Thus, the decision (1) is equivalent to the Kullback-Leibler minimum information discrimination principle [12]

$$\min_{r \in \{1,\ldots,R\}} \rho(X, X_r), \tag{3}$$

where

$$\rho(X, X_r) = \rho_{KL}(H, H_r) = \sum_{i=1}^{N} h_i \cdot \ln \frac{h_i}{h_{r;i}}. \tag{4}$$

is the Kullback-Leibler divergence between densities H and H_r.

2.2 Baseline: Directed Enumeration Method

It is known that the performance of brute force implementation of criterion (3) can be rather low. To speed-up recognition process, fast approximate NN algorithms can be used. As a baseline approximate NN method we use the DEM [11] which was based on the metric properties of the Kullback-Leibler divergence and regards the models' similarity $\rho_{i,j} = \rho(X_i, X_j)$ as an average information from an observation to distinct class i from an alternative class j. At the preliminarily step, the model distance matrix $\mathrm{P} = [\rho_{i,j}]$ is calculated as it is done in the AESA [3]. This time-consuming procedure should be repeated only once for a particular task and training set.

Original DEM used the following heuristic: if there exists a model X_ν for which $\rho(X, X_\nu) < \rho_0 \ll 1$, then condition holds $|\rho(X, X_r) - \rho_{\nu,r}| \ll 1$ with high probability for an arbitrary r-th model. Hence, criterion (3) can be simplified

$$\rho(X, X_\nu) < \rho_0 = const. \tag{5}$$

This equation defines the termination condition of the approximate NN method. If false-accept rate (FAR) β is fixed, then ρ_0 is evaluated as a β-quantile of the distances between images from distinct classes $\{\rho_{i,j} | i \in \{1, \ldots, R\}, j \in \{1, \ldots, i-1, i+1, \ldots, R\}\}$ [11].

According to the DEM [11], at first, the distance $\rho(X, X_{r_1})$ to randomly chosen model X_{r_1} is calculated. Next, it is put into the priority queue of models sorted by the distance to X. The highest priority item X_i is pulled from the queue and the set of models $X_i^{(M)}$ is determined from

$$\left(\forall X_j \notin X_i^{(M)}\right) \left(\forall X_k \in X_i^{(M)}\right) \quad \Delta\rho(X_j) \geq \Delta\rho(X_k) \tag{6}$$

where $\Delta\rho(X_j) = |\rho_{i,j} - \rho(X, X_j)|$ is the deviation of $\rho_{i,j}$ relative to the distance between X and X_j. For all models from the set $X_i^{(M)}$ the distance to the query object is calculated and the condition (5) is verified. After that, every previously unchecked model from this set is put into the priority queue. The method is terminated if for one model object condition (5) holds or after checking for $E_{\max} = const$ models.

As we stated earlier, this method is heuristic as most popular approximate NN algorithms. However, the probability that the model is the NN of X can be directly calculated for the Kullback-Leibler discrimination by using its asymptotic properties. Let's describe this idea in detail in the next section.

3 Maximum-Likelihood Directed Enumeration Method

In this section we primarily focus on *greedy* algorithms: it explores an each step the model which is the NN of the query object X with the highest probability. It is known [12] that if an object X has distribution H_v, then the distance $2n \cdot \rho(X, X_v)$ is asymptotically distributed as a χ^2 with (N - 1) degrees of freedom. Similarly, $2n \cdot \rho(X, X_r)$, $r \neq v$ has asymptotic non-central χ^2 distribution with (N - 1) degrees of freedom and noncentrality parameter $2nK \cdot \rho_{v,r}$. If N is high, then, by using the central limit theorem, we obtain the normal distribution of the distance $\rho(X, X_r)$:

$$N\left(\rho_{v,r} + (N-1)/(2n); \left(\sqrt{8n \cdot \rho_{v,r} + 2(N-1)}/(2n)\right)^2\right). \tag{7}$$

At first, based on the asymptotic distribution (7) we replace the step (6) of the original DEM to the procedure of choosing the maximum likelihood model. Let's assume that the models $X_{r_1}, \ldots, X_{r_l}$ have been examined before the l-th step. We choose the next most probable model $X_{r_{l+1}}$ with the maximum likelihood method:

$$r_{l+1} = \underset{v \in \{1,\ldots,R\} - \{r_1,\ldots,r_l\}}{\arg\max} \prod_{i=1}^{l} f(\rho(X, X_{r_i})|W_v). \tag{8}$$

where $f(\rho(X, X_{r_i})|W_v)$ is the conditional density (likelihood) of the distance $\rho(X, X_{r_i})$ if the hypothesis W_v is true. By using asymptotic distribution (7), the likelihood in (8) can be written in the following form

$$\begin{aligned} f(\rho(X, X_{r_i})|W_v) &= \frac{2n}{\sqrt{2\pi \cdot (8n \cdot \rho_{v,r_i} + 2(N-1))}} \\ &\quad \times \exp\left[-\frac{\left(2n \cdot \left(\rho(X, X_{r_i}) - \rho_{v,r_i}\right) - (N-1)\right)^2}{8n \cdot \rho_{v,r_i} + 2(N-1)}\right] \\ &= \frac{2n}{\sqrt{2\pi}} \exp\left[-\frac{1}{2}\ln\left(8n \cdot \rho_{v,r_i} + 2(N-1)\right)\right] \\ &\quad \times \exp\left[-\frac{\left(2n \cdot (\rho(X, X_{r_i}) - \rho_{v,r_i}) - (N-1)\right)^2}{8n \cdot \rho_{v,r_i} + 2(N-1)}\right] \end{aligned} \tag{9}$$

By several transformations of (9) and assuming that the number of simple features is much higher the number of parameters $n \gg N$, expression (8) is written as

$$r_{l+1} = \underset{\mu \in \{1,\ldots,R\} - \{r_1,\ldots,r_l\}}{\arg\min} \sum_{i=1}^{l} \varphi_{\mu}(r_i). \tag{10}$$

where

$$\varphi_{\mu}(r_i) \approx \left(\rho(X, X_{r_i}) - \rho_{\mu,r_i}\right)^2 / (4\rho_{\mu,r_i}). \tag{11}$$

This equation is in good agreement with the heuristic from the original DEM [11] - the closer are the distances $\rho(X, X_{r_i})$ and ρ_{μ,r_i} and the higher is the distance between models X_{μ} and X_{r_i}, the lower is $\varphi_{\mu}(r_i)$.

Next, the termination condition (5) is tested for the model $X_{r_{l+1}}$. If the distance $\rho\left(X, X_{r_{l+1}}\right)$ is lower than a threshold ρ_0 or the number of calculated distances exceed $E_{\max}$, then the search procedure is stopped on the $L_{checks} = l + 1$ step. Otherwise the model $X_{r_{l+1}}$ is put into the set of previously checked models and the procedure (10) and (11) is repeated.

Let us return to the initialization of our method. We would like to choose the first model X_{r_1} to obtain the decision (5) in a shortest (in terms of number of calculations L_{checks}) way. An average probability to obtain the decision is maximized at the second step:

$$r_1 = \underset{\mu \in \{1,\ldots,R\}}{\arg\max} \frac{1}{R} \sum_{\nu=1}^{R} P\left(\varphi_{\nu}(\mu) \leq \min_{r \in \{1,\ldots R\}} \varphi_r(\mu) \middle| W_{\nu} \right). \tag{12}$$

To estimate the conditional probability in (12) we use again the asymptotic distribution (7). The first model to check X_{r_1} is obtained from the following expression

$$r_1 = \underset{\mu \in \{1,\ldots,R\}}{\arg\max} \sum_{\nu=1}^{R} \prod_{r=1}^{R} \left(\frac{1}{2} + \Phi\left(\frac{\sqrt{n}}{2} \left| \sqrt{\rho_{r,\mu}} - \sqrt{\rho_{\nu,\mu}} \right| \right) \right). \tag{13}$$

where $\Phi(\cdot)$ is the cumulative density function of the normal distribution.

Thus, the proposed ML-DEM (10), (11) and (13) is an optimal (maximal likelihood) greedy algorithm for an approximate NN search. The ML-DEM is different from the baseline DEM in initialization (14) and in the rule of choosing the next model (10) and (11). In the DEM $M > 1$ models are chosen (6) and in the proposed ML-DEM only one model is selected (10). Only the termination condition (5) is the same for both DEM and ML-DEM. In fact, the proposed method can be applied not only with the Kullback-Leibler discrimination (4), but with an arbitrary similarity measure. However, the property of statistical optimality is preserved only for similarity measure (4).

4 Experimental Results

Our experimental study deals with face recognition problem [13] with color FERET dataset. All 2720 frontal photos were converted to grayscale intensity images. Random cross-validation repeated 100 times was applied. Each time $R = 1420$ randomly chosen images of 994 persons populate the database (i.e. a training set), other 1300 photos of the same persons formed a test set. Faces were detected with the OpenCV library. After that the median filter with window size (3×3) is applied to remove noise in detected faces. The facial image is divided into a regular grid of $K \times K$ blocks, where $K = 10$. Next the HOGs (histograms of oriented gradients) $H_r(k_1, k_2)$ with $N = 8$ bins are calculated for each block (k_1, k_2) [14]. We assume, that each HOG is normalized, so that it may be treated as a probability distribution [14] in (4). The distance in the nearest neighbor rule (3) is calculated as follows [9]

$$\rho(X, X_r) = \sum_{k_1=1}^{K} \sum_{k_2=1}^{K} \min_{|\Delta_1| \le \Delta, |\Delta_2| \le \Delta} \rho(H(k_1, k_2), H_r(k_1 + \Delta_1, k_2 + \Delta_2)) \qquad (14)$$

with the mutual alignment of the histograms in the Δ-neighborhood in order to take into account the small spatial deviations due to misalignment after face detection. In (14) we use the Kullback-Leibler divergence (4) between the HOGs and the homogeneity-testing probabilistic neural network (HT-PNN) which showed high face recognition rate and is equivalent to the statistical approach if the pattern recognition problem is referred as a task of testing for homogeneity of segments [15].

The error rate obtained with the NN rule and similarity measure (1) with Kullback-Leibler and the HT-PNN distances is shown in Table 1 in the format average error rate ± its standard deviation. Here, first, alignment of HOGs (22) with $\Delta = 1$ improves the recognition accuracy. And, second, we experimentally support the claim [15] that the error rate for the Kullback-Leibler distance (4) is higher when compared with the HT-PNN.

Table 1. Face recognition error rate, criterion (3) and (14)

	$\Delta = 0$	$\Delta = 1$
Kullback-Leibler divergence	8.9 ± 1.3	7.0 ± 1.3
HT-PNN	7.8 ± 1.2	6.6 ± 1.3

In the next experiment we compare the performance of the proposed ML-DEM with brute force (3), original DEM [11], and several approximate NN methods from FLANN [5] and NonMetricSpaceLib [16] libraries showed the best speed, namely

1. Randomized kd-tree from FLANN with 4 trees [6]
2. Composite index from FLANN which combines the randomized kd-trees (with 4 trees) and the hierarchical k-means tree [5].
3. Ordering permutations (perm-sort) from NonMetricSpaceLib which is known to decrease the recognition speed for medium-sized databases [10].

We evaluate the error rate (in %) and the average time (in ms) to recognize one test image with a modern laptop (4 core i7, 6 Gb RAM) and Visual C++ 2013 Express compiler (×64 environment) and optimization by speed. We explore an application of parallel computing [17] by dividing the whole training set into T = const non-overlapped parts. Each part is processed in its own thread implemented by using the Windows ThreadPool API. We analyze both nonparallel ($T = 1$) and parallel ($T = 8$) cases. After several experiments the best (in terms of recognition speed) value of parameter M of original DEM (6) was chosen $M = 64$ for nonparallel case and $M = 16$ for parallel one. To obtain threshold ρ_0, the FAR is fixed to be $\beta = 1$ %. Parameter $E_{\max}$ was chosen to achieve the recognition accuracy which is not 0.5 % less than the accuracy of brute force (Table 1). If such accuracy can not be achieved, $E_{\max}$ was set to be equal to the count of models R. The average recognition time per one test image (in ms) is shown in Table 2.

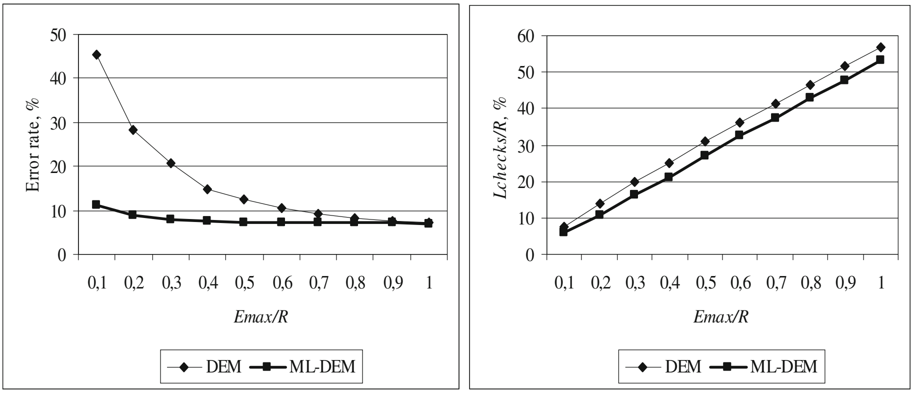
Table 2. Average recognition time (ms.)

Distance/ features	**Kullback-Leibler divergence**				**HT-PNN**			
	$\Delta = 0$		$\Delta = 1$		$\Delta = 0$		$\Delta = 1$	
	$T = 1$	$T = 8$	$T = 1$	$T = 8$	$T = 1$	$T = 8$	$T = 1$	$T = 8$
Brute force	12.9	2.8	99.1	26.6	19.4	5.5	146.1	38.7
Randomized KD tree	11.9	2.6	91.2	21.4	16.4	4.3	129.4	30.4
Composite	12.0	2.6	91.5	22.5	16.7	4.3	129.9	35.2
Perm-sort	4.0	2.1	31.0	12.9	7.8	2.4	43.7	14.3
DEM	5.34	1.3	52.7	12.7	7.3	2.3	52	16.1
ML-DEM	2.8	0.8	24.2	10.0	5.3	1.4	24.9	5.8

Here randomized and composite kd-trees do not show superior performance even over brute force as the number of models in the database is not very high. However, as it was expected, perm-sort method is characterized with 2–3.5 times lower recognition speed in comparison with an exhaustive search. Moreover, perm-sort seems to be better than the original DEM for nonparallel case ($T = 1$), though the DEM's parallel implementation is a bit better. The most important conclusion here is that the proposed ML-DEM shows the highest speed in all experiments. The results of the HT-PNN's usage are very similar, though the error rate here is 0.5–1 % lower (see Table 1). In this case the original DEM is slightly faster than the perm-sort for conventional distance ($\Delta = 0$) but is not so effective for alignment ($\Delta = 1$). FLANN's kd-trees are 10–15 % faster than the brute force. And again, the proposed ML-DEM is the best choice here especially for most complex case ($T = 8$, $\Delta = 1$) for which only 6 ms (in average) is necessary to achieve 93 % accuracy.

To clarify the difference in performance of the original DEM and the proposed ML-DEM, we show the dependence of the error rate and the number of checked models $L_{checks}/R \cdot 100$ % on $E_{\max}$ in Fig. 1a, b respectively. Here the speed of convergence to an optimal solution for the ML-DEM is much higher than the same indicator of the

DEM (Fig. 1a). Even when $E_{\max} = 0.1 \cdot R$ we can get an appropriate solution. Figure 1b proves that, as expected, the proposed ML-DEM is better than the DEM in terms of the number of calculated distances L_{checks}. However, additional computations of the ML-DEM which include the calculations for every non previously checked model, are quite complex. Hence, the difference in performance with the DEM and other approximate NN methods is high only for very complex similarity measures (e.g., in case of $\Delta = 1$).



a b

Fig. 1. Dependence of: (a) error rate; and (b) count of models checks per database size $L_{checks}/R \cdot 100$ % on $E_{\max}/R$, Kullback-Leibler discrimination, $\Delta = 1$

5 Conclusion

In this paper we demonstrated that using the asymptotic properties (7) of the Kullback-Leibler divergence in the proposed ML-DEM gives very good results in image recognition with medium-sized database, reducing the recognition speed by more than 2.5–6.5 times in comparison with brute force and by 1.2–2.5 times in comparison with other approximate NN methods from FLANN and NonMetricSpaceLib libraries. In contrast to the most popular fast algorithms, our method is not heuristic (except the termination condition (5)). Moreover, it does not build data structure based on an algorithmic properties of applied similarity measure (e.g., triangle inequality of Minkowski metric in the AESA [4], Bregman ball for Bregman divergences [9]). The proposed ML-DEM is an optimal (maximum likelihood) greedy method in terms of the number of distance calculations for NN rule (3) with the sum (14) of Kullback-Leibler discriminations (4). Moreover, the ML-DEM can be successfully applied with other distances, e.g. the HT-PNN [15].

The main direction for further research of the proposed method is its modification in case of simple similarity measures. For now, the complexity of extra computation at each step of the ML-DEM (10) and (11) is rather high. Hence, the difference in performance with original DEM and popular approximate NN methods is significant only for very complex similarity measure. One possible solution is a pivot-based indexing [10] and ordering all models with respect to their log-likelihoods (10) and (11).

Acknowledgements. The article was prepared within the framework of the Academic Fund Program at the National Research University Higher School of Economics (HSE) in 2015–2016 (grant № 15-01-0019) and supported within the framework of a subsidy granted to the HSE by the Government of the Russian Federation for the implementation of the Global Competitiveness Program.

References

1. Tan, X., Chen, S., Zhou, Z.H., Zhang, F.: Face recognition from a single image per person: a survey. Pattern Recogn. **39**(9), 1725–1745 (2006)
2. Theodoridis, S., Koutroumbas, K.: Pattern Recognition, 4th edn. Elsevier Inc., Burlington (2009)
3. Arya, S., Mount, D.M., Netanyahu, N.S., Silverman, R., Wu, A.Y.: An optimal algorithm for approximate nearest neighbor searching fixed dimensions. J. ACM **45**(6), 891–923 (1998)
4. Vidal, E.: An algorithm for finding nearest neighbours in (approximately) constant average time. Pattern Recogn. Lett. **4**(3), 145–157 (1986)
5. Muja, M., Lowe, D.G.: Fast approximate nearest neighbors with automatic algorithm configuration. In: 4th International Conference on Computer Vision Theory & Applications (VISAPP), pp. 331–340 (2009)
6. Silpa-Anan, C., Hartley, R.: Optimised KD-trees for fast image descriptor matching. In: International Conference on Computer Vision & Pattern Recognition, pp. 1–8 (2008)
7. He, J., Kumar, S., Chang, S.: On the difficulty of nearest neighbor search. In: 29th International Conference on Machine Learning (ICML-2012), pp. 1127–1134 (2012)
8. Savchenko, A.V.: Nonlinear transformation of the distance function in the nearest neighbor image recognition. In: Zhang, Y.J., Tavares, J.M.R.S. (eds.) CompIMAGE 2014. LNCS, vol. 8641, pp. 261–266. Springer, Heidelberg (2014)
9. Cayton, L.: Efficient Bregman range search. In: Bengio, Y., Schuurmans, D., Lafferty, J.D., Williams, C.K.I., Culotta, A. (eds.) Advances in Neural Information Processing Systems, vol. 22, pp. 243–251. (2009)
10. Gonzalez, E.C., Figueroa, K., Navarro, G.: Effective proximity retrieval by ordering permutations. IEEE Trans. Pattern Anal. Mach. Intell. **30**(9), 1647–1658 (2008)
11. Savchenko, A.V.: Directed enumeration method in image recognition. Pattern Recogn. **45** (8), 2952–2961 (2012)
12. Kullback, S.: Information Theory and Statistics. Dover Publications, Mineola (1997)
13. Chellappa, R., Du, M., Turaga, P., Zhou, S.K.: Face tracking and recognition in video. In: Li, S.Z., Jain, A.K. (eds.) Handbook of Face Recognition, pp. 323–351. Springer, London (2011)
14. Dalal, N., Triggs, B.: Histograms of oriented gradients for human detection. In: International Conference on Computer Vision & Pattern Recognition, pp. 886–893 (2005)
15. Savchenko, A.V.: Probabilistic neural network with homogeneity testing in recognition of discrete patterns set. Neural Netw. **46**, 227–241 (2013)

16. Boytsov, L., Naidan, B.: Engineering efficient and effective non-metric space library. In: Brisaboa, N., Pedreira, O., Zezula, P. (eds.) SISAP 2013. LNCS, vol. 8199, pp. 280–293. Springer, Heidelberg (2013)
17. Savchenko, A.V.: Real-time image recognition with the parallel directed enumeration method. In: Chen, M., Leibe, B., Neumann, Bernd (eds.) ICVS 2013. LNCS, vol. 7963, pp. 123–132. Springer, Heidelberg (2013)