On Improving Dissimilarity-Based Classifications Using a Statistical Similarity Measure*

Sang-Woon Kim[1] and Robert P.W. Duin[2]

[1] Dept. of Computer Science and Engineering, Myongji University,
Yongin, 449-728 South Korea
kimsw@mju.ac.kr

[2] Faculty of Electrical Engineering, Mathematics and Computer Science,
Delft University of Technology, The Netherlands
r.p.w.duin@tudelft.nl

Abstract. The aim of this paper is to present a dissimilarity measure strategy by which a new philosophy for pattern classification pertaining to dissimilarity-based classifications (DBCs) can be efficiently implemented. In DBCs, classifiers are not based on the feature measurements of individual patterns, but rather on a suitable dissimilarity measure among the patterns. In image classification tasks, such as face recognition, one of the most intractable problems is the distortion and lack of information caused by the differences in illumination and insufficient data. To overcome the above problem, in this paper, we study a new way of measuring the dissimilarity distance between two images of an object using a statistical similarity metric, which is measured based on intra-class statistics of data and does not suffer from the insufficient number of the data. Our experimental results, obtained with well-known benchmark databases, demonstrate that when the dimensionality of the dissimilarity representation has been appropriately chosen, DBCs can be improved in terms of classification accuracies.

1 Introduction

Dissimilarity-based classifications (DBCs) [11] are a way of defining classifiers among the classes; and the process is not based on the feature measurements of individual patterns, but rather on a suitable dissimilarity measure among the individual patterns. The characteristic of the dissimilarity approach is that it offers a different way to include expert knowledge on the objects. The three major questions we encountered when designing DBCs are summarized as follows: (1) How can prototype subsets be selected (or created) from the training samples? (2) How can the dissimilarities between samples be measured? (3) How can a classifier in the dissimilarity space be designed?

Several strategies have been used to explore these questions. First, various methods have been proposed in the literature [11], [12] as a means of selecting a representation subset of data that is both compact and capable of representing the entire data set. In

these methods, however, it is difficult to find the optimal number of prototypes and, furthermore, selecting prototype stage may potentially lose some useful information for discrimination. To avoid these problems, Bunke and his colleagues [12] and Kim and Gao [8] prefer not to directly select the representative prototypes from the training samples; rather, they use a dimension reduction scheme after computing the dissimilarity matrix with the *entire* training samples.

With regard to the second question, investigations have focused on measuring the appropriate dissimilarity by using various L_p norms, modified Hausdorff norms [7], and traditional measures, such as those used in template matching and correlation-based analysis [1], [11]. On the final question, the learning paradigms, Pekalska and Duin [11] reported the use of many traditional decision classifiers, including the k-NN rule and the linear/quadratic normal-density-based classifiers. Recently, in [4], they tried to refine the dissimilarity matrix by employing a pseudo-Euclidean embedding algorithm [3]. In addition, optimizing DBCs through combining dissimilarity matrices generated with different measures has been investigated in the literature [9], [15].

On the other hand, when designing a specific classification system, sometimes we suffer from the difficulty of collecting sufficient data for each object. In face recognition, for example, there are many kinds of variations based on such factors as pose (direction), expression, and illumination [1], [6]. However, as mentioned above, collecting sufficient facial data is difficult. To solve this problem, Lee and Park [10] proposed a measuring scheme to extract more robust and essential information of data distributions in biometric problems and applied it to developing a similarity measure. The information obtained with the scheme does not depend on the distribution of each class for each object, but depends on all the data. From this point of view, they claimed to get a more reliable similarity measure.

The major task of this study is to deal with how the dissimilarity measure can be effectively computed. However, when a limited number of object samples are available or the representational capability is insufficient to cover the possible variations of data, it is difficult to improve the performance of DBCs in the dissimilarity space. To overcome this limitation and thereby improve the classification performance of DBCs, in this paper, we study a new way of enriching the representational capability of dissimilarity measures. In particular, this goal can be achieved by using a statistical similarity measure based on intra-class statistics of data [10].

The main contribution of this paper is to demonstrate that the classification performance of DBCs can be improved by employing a similarity measure based on the intra-class statistics of all the training samples. Here, the measuring system has been used to accommodate some useful information for discrimination and to avoid the difficulty of collecting sufficient training data. The remainder of the paper is organized as follows: In Section 2, after providing a brief introduction to DBCs, we present an explanation of the statistical similarity measure and an improved DBC. In Section 3, we present the experimental results obtained with real-life benchmark data sets. In Section 4, we present our concluding remarks.

2   Related Work

Dissimilarity Representation: A dissimilarity representation of a set of samples, $T = \{\boldsymbol{x}_i\}_{i=1}^n \in \mathbb{R}^{n \times d}$, is based on pair-wise comparisons, and is expressed, for example, as an $n \times m$ dissimilarity matrix, $D_{T,Y}[\cdot, \cdot]$, where $Y = \{\boldsymbol{y}_j\}_{j=1}^m \in \mathbb{R}^{m \times d}$, a prototype set, is extracted from T, and the subscripts of D represent the set of elements, on which the dissimilarities are evaluated. Thus, each entry, $D_{T,Y}[i, j]$, corresponds to the dissimilarity between the pairs of objects, $\langle \boldsymbol{x}_i, \boldsymbol{y}_j \rangle$, where $\boldsymbol{x}_i \in T$ and $\boldsymbol{y}_j \in Y$. Consequently, an object, $\boldsymbol{x}_i$, is represented as a column vector as follows:

$$[d(\boldsymbol{x}_i, \boldsymbol{y}_1), d(\boldsymbol{x}_i, \boldsymbol{y}_2), \cdots, d(\boldsymbol{x}_i, \boldsymbol{y}_m)]^T, 1 \le i \le n. \tag{1}$$

Here, the dissimilarity matrix, $D_{T,Y}[\cdot, \cdot]$, is defined as a *dissimilarity space*, on which the d-dimensional object, $\boldsymbol{x}$, given in the feature space, is represented as an m-dimensional vector, $\delta(\boldsymbol{x}, Y)$, where if $\boldsymbol{x} = \boldsymbol{x}_i$, $\delta(\boldsymbol{x}_i, Y)$ is the i-th row of $D_{T,Y}[\cdot, \cdot]$. In this paper, the column vector, $\delta(\boldsymbol{x}, Y)$, is simply denoted by $\delta_Y(\boldsymbol{x})$, where the latter is an m-dimensional vector, while $\boldsymbol{x}$ is d-dimensional.

On the basis of what we have briefly discussed, we assert that the state-of-the-art strategy for DBCs involves the following steps:

1. Select the representation subset, Y, from the training set, T, by using one of the selection methods described in the literature.

2. Using Eq. (1), compute the dissimilarity matrix, $D_{T,Y}[\cdot, \cdot]$, in which each dissimilarity is computed on the basis of the measures described in the literature.

3. For a testing sample, $\boldsymbol{x}$, compute a dissimilarity column vector, $\delta_Y(\boldsymbol{x})$, by using the same measure used in Step 2.

4. Achieve the classification by invoking a classifier built in the dissimilarity space and operating it on the dissimilarity vector $\delta_Y(\boldsymbol{x})$.

Here, we can see that the performance of DBCs relies heavily on how well the dissimilarity space, which is determined by the dissimilarity matrix, is constructed. To improve the performance, we need to ensure that the matrix is well designed.

A Statistical Similarity Measure [10]: To define a new similarity measure based on the statistics of data, let us represent the data as a random variable, $\boldsymbol{x}_i = (x_{i1}, x_{i2}, \cdots, x_{id})^T$. The data set, T, can be decomposed into subsets, T_k, as follows: $T = \bigcup_{k=1}^c T_k, T_i = \{\boldsymbol{x}_1, \cdots, \boldsymbol{x}_{n_i}\}$, with $n = \sum_{i=1}^c n_i, T_i \cap T_j = \phi, \forall i \ne j$.

First, we introduce a random variable, $\boldsymbol{z}_i$, which is defined by using a pair of data, $\langle \boldsymbol{x}_k, \boldsymbol{x}_l \rangle$, where $\boldsymbol{x}_k, \boldsymbol{x}_l \in T_i$, from the same class ω_i. We then try to estimate a multivariate Gaussian distribution, $p_i(\boldsymbol{z})$, instead of $p_i(\boldsymbol{x})$, $i = 1, \cdots, c$, and use it to define the similarity measure. Let us define the random variable, $\boldsymbol{z}_i$, as follows:

$$\boldsymbol{z}_i = \boldsymbol{x}_k - \boldsymbol{x}_l, \forall \boldsymbol{x}_k, \boldsymbol{x}_l \in T_i, (k \ne l). \tag{2}$$

This is given under the assumption that the difference between each pair of samples from the same class originates from some additive Gaussian noises [10].

To define a distance measuring system, we first construct a representation set, $S = \{\boldsymbol{z}_i\}_{i=1}^m$, from the training set, $T = \{\boldsymbol{x}_j\}_{j=1}^n$, as follows: (1) For every class

ω_i, compute $S_i = \{z_i\}_{i=1}^{m_i} \in \mathbb{R}^{m_i \times d}$, $m_i = n_i(n_i - 1)$, using Eq. (2). (2) Return $S = S_1 \cup S_2 \cup \cdots \cup S_c \in \mathbb{R}^{m \times d}$, where $m = \sum_{i=1}^{c} m_i$, as the final representation set.

After obtaining the representation set S, we estimate intra-class statistics of the set, such as mean values and standard deviations, as follows:

$$\mu_i = \frac{1}{m} \sum_{j=1}^{m} z_{ji}, \ (1 \leq i \leq d), \tag{3}$$

$$\sigma_i = \left(\frac{1}{m} \sum_{j=1}^{m} (z_{ji} - \mu_i)^2 \right)^{1/2}, \ (1 \leq i \leq d). \tag{4}$$

Using these statistics, we define a similarity measure between two points, $\boldsymbol{x}_i$ and $\boldsymbol{y}_j$, as follows:

$$s(\boldsymbol{x}_i, \boldsymbol{y}_j) = \sum_{k=1}^{d} \left(\frac{x_{ik} - y_{jk} - \mu_k}{\sigma_k} \right)^2, \ \forall \boldsymbol{x}_i, \boldsymbol{y}_j \in T. \tag{5}$$

Here, since the cardinality of S, m_i, is much larger than that of each T_i, n_i, the intra-class statistics computed are more accurate and more robust against some additive Gaussian noises. Consequently, the similarity measure obtained with these statistics is also robust against the noises and works well with the insufficient data.

Optimized Dissimilarity-Based Classification: As mentioned earlier, there are a few ways by which the classification efficiency of DBCs can be improved. To overcome the limitation caused by the variations in illumination and the insufficient number of data, in this paper, we used the similarity function of Eq. (5). The proposed approach, which is referred to as an optimized DBC (ODBC), is summarized in the following:

1. Select the whole training set, T, as the representation subset Y.

2. After constructing S from T, compute the intra-class statistics, $\boldsymbol{\mu}$ and $\boldsymbol{\sigma}$, of S using Eqs. (3) and (4), respectively.

3. Using Eq. (1), compute the dissimilarity matrix $D_{T,Y}[\cdot, \cdot]$, in which each individual dissimilarity is computed using the similarity function of Eq. (5), rather than using one of the Euclidean measures.

4. This step is the same as Step 3 in the conventional DBC.

5. This step is the same as Step 4 in the conventional DBC.

The time complexities of the above algorithm, ODBC, can be analyzed as follows: As in the case of DBC, almost all the processing CPU-time of ODBC is consumed in computing the dissimilarity matrices. So, the difference in magnitude between the computational complexities of DBC and ODBC depends on the computational costs associated with the dissimilarity matrix. More specifically, in DBC, Step 2 of computing the $n \times n$ dissimilarity matrix requires $O(dn^2)$ time. On the other hand, the computation of that of ODBC needs $O(dn^2 + cn^2)$ time[1] in executing Steps 2 and 3. Here, n, d, and c are the numbers of total samples, dimensions, and classes, respectively.

---

[1] For large data sets, to reduce the CPU-time of ODBC, a sample reduction technique can be considered.

3   Experimental Results

Experimental Data: The proposed method has been tested and compared with the conventional ones. This was done by performing experiments on well-known face databases, namely, AT&T [13], Yale [6], and CMU-PIE [14], and other multivariate data sets cited from UCI Machine Learning Repository [2].

AT&T database consists of ten different images of 40 distinct objects, for a total of 400 images. The size of each image is 112×92 pixels, for a total dimensionality of 10304 pixels. Yale database contains 165 gray scale images of 15 individuals. The size of each image is 178×236 pixels, for a total dimensionality of 42008 pixels. To obtain a different data set, a partial image, which is 64×78 pixels in size and contains only facial components without background, was extracted from each of the Yale images. This database is referred to as "Yale2" in the following sections. CMU-PIE database involves 41368 images of 68 people. To reduce the computational complexity of this experiment and, furthermore, to investigate the run-time characteristics of the method, a number of subsets, such as 175 (= 25 people $\times$ 7 pictures), 350 (=25 people $\times$ 14 pictures), and 525 (=25 people $\times$ 21 pictures) images, were selected from the database and down-sampled into 92×80 pixels, for a total dimensionality of 7360 pixels.

Experimental Method: In this experiment, first, data sets are randomly split into training sets and test sets in the ratio of 75 : 25. Then, the training and testing procedures are repeated 15 times and the results obtained are averaged. To evaluate the classification accuracies of DBCs and ODBCs, different classifiers, such as k-nearest neighbor classifiers and support vector machines [5], are employed and implemented with PRTools[2], and will be denoted as *knnc* and *libsvm*, respectively, in subsequent sections.

Experimental Results: First, the experimental results obtained with *knnc* and *libsvm* trained in DBCs and ODBCs for the face databases, namely, AT&T, Yale (Yale2), and CMU-PIE, were probed into. Here, we first reduced the dimensionality of the image vectors by performing a principal component analysis (PCA). Then, we constructed the dissimilarity matrix, D, with respect to *all* the training samples. Fig. 1 shows a comparison of the error rates of *knnc* and *libsvm* trained in DBCs and ODBCs for the four databases, where x and y axes are those of the reduced dimensions (which are obtained with a PCA) and the estimated error rates, respectively.

The observations obtained from the figure are the followings: First, it should be pointed out that the difference in the estimated error rates between DBCs and ODBCs increases as the dimension of the subspaces increases. This is clearly shown in the error rates represented with two lines marked with $\diamond$ and $\times$, or $\circ$ and $+$, respectively, in the pictures. This comparison shows that the classification accuracy of ODBCs is higher than that of DBCs when the dimensionality of the subspace is appropriately chosen (refer to Fig. 1 (b), (c), and (d)). Next, for the experiment of CMU-PIE, the two error rates of ODBCs, obtained with the two subsets of 175 (= 25×7) and 350 (= 25×14) images, respectively, are almost the same, while those of DBCs are different; the error rates of the latter subset are lower than that of the former. However, for the experiment of AT&T, in which all of the facial images have a uniform background, the classification

---

[2] PRTools is a Matlab toolbox for pattern recognition(refer to http://prtools.org/).
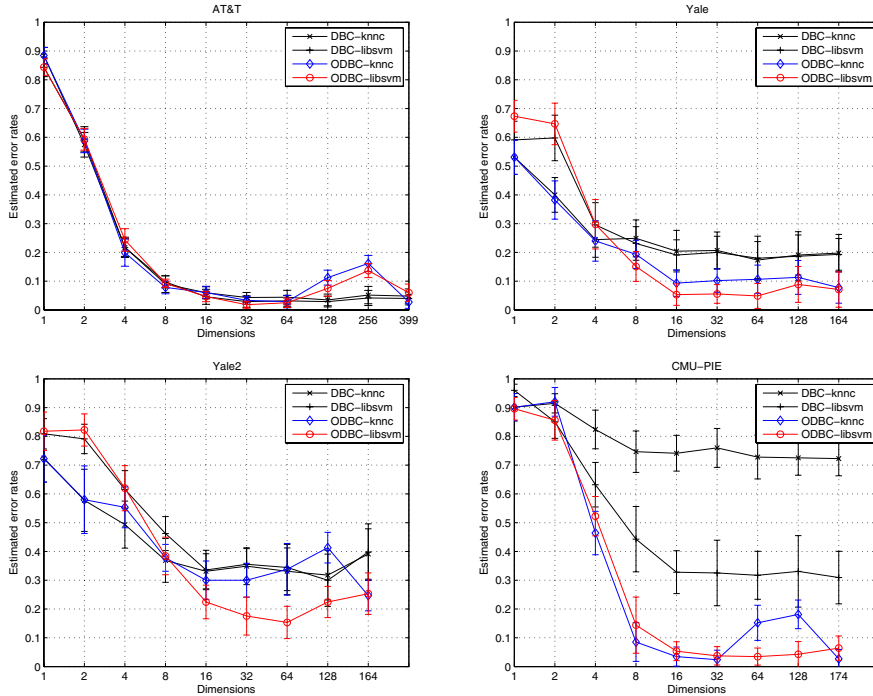
Fig. 1. A comparison of the estimated error rates of DBCs and ODBCs: (a) top left, (b) top right, (c) bottom left, and (d) bottom right; (a) - (d) are obtained with AT&T, Yale, Yale2, and CMU-PIE 175 (25 people × 7 pictures) databases, respectively. Here, the error rates are evaluated with *knnc* and *libsvm* trained in the subspaces of different dimensions. (The error rates of CMU-PIE 350(25 × 14) and 525(25 × 21) subsets are omitted here in the interest of compactness.)

accuracies have not increased with the proposed method (refer to Fig. 1 (a)). From these considerations, the reader should observe that DBCs can be improved by employing the statistical similarity measure and the resultant DBCs, i.e., ODBCs, do not suffer from the insufficient number of data.

In addition, the scaling of the features by σ_k may work in two directions. It works well for features that contribute to class differences by their separateness, but that have a small variability and that thereby have a too small contribution to the dissimilarities if there is no scaling. In case of bad features with a small variability, however, the noise of these features is emphasized. This happens for the peaking points in Fig. 1 (a) and (d) as the smallest eigenvectors are most emphasized.

To further investigate the advantage of using the proposed method, and, especially, to find out which kinds of significant data set are more suitable for the scheme, we repeated the experiment with the UCI benchmark databases. Table 1 shows a numerical comparison of the averaged error rates and their standard deviations for the benchmark databases. Here, the estimated error rates that increase and/or decrease more than the sum of the standard deviations are underlined.

Table 1. A numerical comparison of the estimated error rates for the benchmark databases. Here, the numbers in brackets in each row represent the standard deviations. Also, the estimated error rates that increase and/or decrease more than the sum of the standard deviations are underlined.

data sets	parameters (c, d, n)	*knnc*		*libsvm*	
		DBCs	ODBCs	DBCs	ODBCs
apect	2, 44, 80	0.3400 (0.0761)	0.3367 (0.0876)	0.1900 (0.0660)	0.2067 (0.0842)
chromo	24, 30, 1143	0.4285 (0.0215)	0.4662 (0.0254)	<u>0.3860</u> (0.0341)	0.4754 (0.0244)
cmc	3, 9, 1473	0.5241 (0.0225)	0.5755 (0.0256)	0.4816 (0.0173)	0.4652 (0.0148)
dermatology	6, 34, 366	0.2418 (0.0381)	<u>0.0681</u> (0.0228)	0.0520 (0.0240)	0.0476 (0.0237)
diabetes	2, 8, 768	0.2161 (0.0413)	<u>0.0542</u> (0.0201)	0.0623 (0.0293)	0.0505 (0.0238)
ecoli	3, 7, 272	0.0756 (0.0279)	0.0796 (0.0274)	0.0527 (0.0202)	0.0925 (0.0556)
glass	4, 9, 214	0.3167 (0.0645)	0.3033 (0.0421)	0.2967 (0.0647)	0.3450 (0.0649)
heart	2, 13, 297	0.4396 (0.0402)	<u>0.2360</u> (0.0442)	0.3270 (0.0567)	<u>0.1793</u> (0.0308)
lung-cancer	3, 56, 32	0.5429 (0.1724)	0.4571 (0.1344)	0.5238 (0.1034)	0.5238 (0.1285)
malaysia	20, 8, 291	0.4757 (0.0606)	<u>0.2531</u> (0.0402)	0.7141 (0.0466)	<u>0.2802</u> (0.0543)
sonar	2, 60, 208	0.1961 (0.0385)	0.1882 (0.0479)	0.1725 (0.0414)	0.2575 (0.0662)
wine	3, 13, 178	0.2837 (0.0458)	<u>0.0202</u> (0.0149)	0.2558 (0.0596)	<u>0.0186</u> (0.0219)

We observed the same characteristics in Table 1 as in Fig. 1 (see the underlined numbers). This improvement can be seen by observing how the estimated error rates (%) change. For example, for the last data set, wine, the error rates of *knnc* (and *libsvm*) designed with DBC and ODBC significantly decrease from 28.37% to 2.02% (and 25.58% to 1.86%), respectively. The same characteristics could also be observed in the underlined data sets. However, for the other data sets, the error rates of the both of DBCs and ODBCs are *almost* the same; the increase and/or decrease of the error rates is not significant. Additionally, what can be observed in the table for the feature based data sets can for a large deal be explained by scaling differences between the features. Malaysia and diabetes are typically data sets with entirely different features that are not scaled properly. The sonar data set consists of spectra. By scaling the tails of the spectra are emphasized.

4 Conclusions

In our efforts to improve the classification performance of DBCs, we used a statistical measuring technique based on intra-class statistics of data. To achieve this improvement of DBCs, we first computed the intra-class statistics, such as the mean and the standard deviation, of the training data set. Using these statistics, we then constructed the dissimilarity matrices, where the dissimilarity was measured with the similarity function. This measuring technique has been employed to solve the problems caused by the differences in illumination and the insufficient number of data. The proposed method was tested on four face databases and some UCI data sets, and the results were compared with those of a Euclidean method. Our experimental results demonstrate that the classification accuracies of DBCs were improved significantly when the dimensionality of the dissimilarity representation has been appropriately chosen. Although we have shown that DBCs can be improved by employing the statistical measuring scheme, many tasks remain open. One of them is to improve the classification efficiency by combining the

optimized DBCs in the dissimilarity space. Also, it is not yet clear that which kinds of significant data sets are more suitable for the scheme. Therefore, the problem of theoretically investigating the measuring method developed for the proposed DBCs remains to be done. Future research will address these concerns.

References

1. Adini, Y., Moses, Y., Ullman, S.: Face recognition: the problem of compensating for changes in illumination direction. IEEE Trans. Pattern Anal. and Machine Intell. 19(7), 721–732 (1997)
2. Asuncion, A., Newman, D.J.: UCI Machine Learning Repository. University of California, School of Information and Computer Science, Irvine, CA (2007), Can also be downloaded as of `http://www.ics.uci.edu/~mlearn/MLRepository.html` (February 2010)
3. Borg, I., Groenen, P.: Morden Mutlidimensional Scaling: Theory and Applications. Springer, New York (1997)
4. Duin, R.P.W., Pekalska, E., Harol, A., Lee, W.-J., Bunke, H.: On Euclidean corrections for non-Euclidean dissimilarities. In: da Vitoria Lobo, N., Kasparis, T., Roli, F., Kwok, J.T., Georgiopoulos, M., Anagnostopoulos, G.C., Loog, M. (eds.) SS+SSPR 2008. LNCS, vol. 5342, pp. 664–673. Springer, Heidelberg (2008)
5. Fan, R.-E., Chen, P.-H., Lin, C.-J.: Working set selection using the second order information for training SVM. Journal of Machine Learning Research 6, 1889–1918 (2005)
6. Georghiades, A.S., Belhumeur, P.N., Kriegman, D.J.: From few to many: Illumination cone models for face recognition under variable lighting and pose. IEEE Trans. Pattern Anal. and Machine Intell. 23(6), 643–660 (2001)
7. Hu, Y., Wang, Z.: A similarity measure based on Hausdorff distance for human face recognition. In: Proceedings of 18th International Conference on Pattern Recognition (ICPR 2006), Hong Kong, vol. 3, pp. 1131–1134 (2006)
8. Kim, S.-W., Gao, J.: On using dimensionality reduction schemes to optimize dissimilarity-based classifiers. In: Ruiz-Shulcloper, J., Kropatsch, W.G. (eds.) CIARP 2008. LNCS, vol. 5197, pp. 309–316. Springer, Heidelberg (2008)
9. Kim, S.-W., Duin, R.P.W.: On optimizing dissimilarity-based classifier using multi-level fusion strategies. Journal of The Institute of Electronics Engineers of Korea 45-CI(5), 15–24 (2008) (in Korean); A preliminary version of this paper was presented at Kobti, Z., Wu, D. (eds.): Canadian AI 2007. LNCS (LNAI), vol. 4509, pp. 110–121. Springer, Heidelberg (2007)
10. Lee, K., Park, H.: A new similarity measure based on intraclass statistics for biometric systems. ETRI Journal 25(5), 401–406 (2003)
11. Pekalska, E., Duin, R.P.W.: The Dissimilarity Representation for Pattern Recognition: Foundations and Applications. World Scientific Publishing, Singapore (2005)
12. Riesen, K., Kilchherr, V., Bunke, H.: Reducing the dimensionality of vector space embeddings of graphs. In: Perner, P. (ed.) MLDM 2007. LNCS (LNAI), vol. 4571, pp. 563–573. Springer, Heidelberg (2007)
13. Samaria, F., Harter, A.: Parameterisation of a stochastic model for human face identification. In: Proceedings of 2nd IEEE Workshop on Applications of Computer Vision, Sarasota FL, pp. 215–220 (1994)
14. Sim, T., Baker, S., Bsat, M.: The CMU pose, illumination, and expression(PIE) database of human faces, Technical report CMU-RI-TR-01-02, Robotics Institute, Carnegie Mellon University, Pittsburgh, PA (2001)
15. Woznica, A., Kalousis, A., Hilario, M.: Learning to combine distances for complex representations. In: Proceedings of the 24th International Conference on Machine Learning, Corvallis OR, pp. 1031–1038 (2007)