Dissimilarity Representation in Multi-feature Spaces for Image Retrieval

Luca Piras and Giorgio Giacinto

Department of Electrical and Electronic Engineering University of Cagliari,
Piazza D'armi 09123 Cagliari, Italy
luca.piras@diee.unica.it, giacinto@diee.unica.it

Abstract. In this paper we propose a novel approach to combine information form multiple high-dimensional feature spaces, which allows reducing the computational time required for image retrieval tasks. Each image is represented in a "(dis)similarity space", where each component is computed in one of the low-level feature spaces as the (dis)similarity of the image from one reference image. This new representation allows the distances between images belonging to the same class being smaller than in the original feature spaces. In addition, it allows computing similarities between images by taking into account multiple characteristics of the images, and thus obtaining more accurate retrieval results. Reported results show that the proposed technique allows attaining good performances not only in terms of precision and recall, but also in terms of the execution time, if compared to techniques that combine retrieval results from different feature spaces.

1 Introduction

One of the peculiarities of content-based image retrieval is its suitability to a huge number of applications. Image retrieval and categorization is used to organize professional and home photos, in the field of fashion, for retrieving paintings from a booklet of a picture gallery, for retrieving images from the Internet, and the number of applications is growing every day. The increasing use of images in digital format causes the size of visual archives of becoming bigger and bigger, thus increasing the difficulty of image retrieval tasks. In particular, the main difficulties are related to the increase of the computational load, and to the reduction of the separation between different image categories. In fact, a larger number of images belonging to different categories may be represented as close points in each one of the low-level feature spaces that can be used to represent the visual content.

The combination of multiple image representations (colors, shapes, textures, etc.) has been proposed to effectively cope with the reduced inter-class variation. As a drawback, the use of multiple image representations with a high number of components increases the computational cost of retrieval techniques. As a consequence, the response time of the system might become an issue for interactive applications (e.g., web searching). Over the years, the pattern recognition

G. Maino and G.L. Foresti (Eds.): ICIAP 2011, Part I, LNCS 6978, pp. 139–148, 2011.

community proposed a number of solutions for combining the output of different sources of information [10]. The most popular and effective techniques for output combination are based on fusion techniques, such as the mean rule, the maximum rule, the minimum rule, and weighted means. In the field of content-based image retrieval, similar approaches can be employed by considering the value of similarity between images as the output of a classifier. In particular, combination approaches have been proposed for fusing different feature representations, where the appropriate similarity metric is computed in each feature space, and then all the similarities are fused through a weighted sum [15]. Another approach to combine different image representations is to stack all the available feature vectors into a single feature vector, and then computing the similarity between images by using this high-dimensional representation. As the computational cost increases with the size of the database and the size of the feature space, it is easy to see that the use of a unique feature vector made up by stacking different feature representations might not be a feasible solution.

The issue of combining different feature representations is also relevant when relevance feedback mechanisms are used. In this case, at each iteration, similarities have to be computed by exploiting relevance feedback information, for example by resorting to Nearest-Neighbor or Support Vector Machine [17] techniques. In particular, when the combination is attained by computing a weighted sum of distances, the cost of the estimation of the weights related to relevance feedback information have to be also taken into account. It is easy to see that the effectiveness of a given representation of the images is strictly related to the retrieval method employed. In this viewpoint, an approach that has been recently proposed in the pattern recognition field is the so called "dissimilarity space". This approach is based on the creation of a new space where patterns are represented in terms of their (dis)similarities to some reference prototypes. Thus the dimension of this space does not depend on the dimensions of the low-level features employed, but it is equal to the number of reference prototypes used to compute the dissimilarities This technique has been used recently to exploit Relevance feedback in content-based image retrieval field [5,12], where relevant images play the role of reference prototypes. In addition, dissimilarity spaces have been also proposed for image retrieval to exploit information from different multi-modal characteristic [2].

In this paper we propose a novel use of the dissimilarity representation for improving relevance feedback based on the Nearest-Neighbor approach [4]. Instead of computing (dis)similarities by using different prototypes (e.g., the relevant images) and a single feature space, we propose to compute similarities by using just one prototype, and multiple feature representations. Each image is thus represented by a very compact vector that summarizes different low-level characteristics, and allows images that are relevant to the user's goals to be represented as near points. The resulting retrieval system is both accurate and fast, because, at each relevance feedback iteration, retrieval performances can be significantly improved with a low computational time compared to the number of low-level features considered.

The rest of the paper is organized as follows. Section 2 briefly reviews the dissimilarity space approach, and introduces the technique proposed in this paper. Section 3 shows the integration of the proposed approach in the learning process of a relevance feedback mechanism based on the Nearest-Neighbor paradigm. Section 4 illustrates some approaches proposed in the literature to combine different feature spaces. Experimental results are reported in Section 5. Reported results show that the proposed approach allows outperforming other methods of feature combination both in terms of performances and execution time. Conclusions are drawn in Section 6.

2 From Multi-spaces to Dissimilarity Spaces

Dissimilarity spaces are defined as follows [13]. For a given classification task, let us consider a set $P = \{\mathbf{p}_1, \ldots, \mathbf{p}_L\}$ made up of L patterns selected as *prototypes*, and let us compute the distances $d(\cdot)$ between each pattern and the set of prototypes. These distances can be computed in a low-level feature space. Each pattern is then represented in terms of a L-dimensional vector, where each component is the distance between the pattern itself and one of the L prototypes. If we denote with $d(\mathbf{I}_i, \mathbf{p}_j)$ the distance between pattern $\mathbf{I}_i$ and the prototype $\mathbf{p}_j$, the representation of pattern $\mathbf{I}_i$ in the dissimilarity space will be:

$$\mathbf{I}_i^P = [d(\mathbf{I}_i, \mathbf{p}_1), \ldots, d(\mathbf{I}_i, \mathbf{p}_P)]. \tag{1}$$

It should be quite clear that the performances depend on the choice of the prototypes, especially when this technique is used to transform a high-dimensional feature space into a lower dimensional feature space. The literature clearly shows that the choice of the most suitable prototypes is not a trivial task [13]. In this paper we use the basic idea of dissimilarity spaces to produce a new vector from different feature spaces.

Before entering into the details of the proposed technique, let us recall that the goal is to produce an effective way of combining different feature representations of images in the context of a Nearest-Neighbor relevance feedback approach for content-based image retrieval [4]. Relevance feedback provides the systems a number of images that are relevant to the user's needs at each iteration. It is quite easy to see that if we consider different image representations, usually different sets of images are found in the nearest neighborhood of relevant images. Which strategy can be employed to assess which of the images can be considered as relevant? One solution can be the use of combination mechanisms based on the weighted fusion of similarity measures computed in different feature spaces. As an alternative, strategies based on the computation of the *max*, or the *min* similarity measure can be employed. Finally, the computation of similarity can be carried out using a vector where the components from different representations are stacked. The fusion of similarities requires some heuristics to compute the weights of the combination, while the *max* and *min* rules can be more sensitive to "semantic" errors in the evaluation of similarity due to the so-called semantic gap [9]. Finally, the use of stacked vectors can be computationally expensive, and can suffer from the so-called "curse of dimensionality", as the dimension

of the resulting space may be too large compared to the number of available samples of relevant images.

In order to provide a solution to the computation of the relevance of an image with respect to the user's goal by exploiting information from different image representations, we propose to construct a dissimilarity space by computing the dissimilarities from a single prototype using multiple feature representations. This approach results in a very compact feature space, as the dimension is equal to the number of feature representations. In order to formalize the proposed technique, let $F = \{f_1, \ldots, f_M\}$ be the set of low-level feature spaces extracted from the images, and let $d_{f_m}\left(\mathbf{I}_i^{f_m}, \mathbf{I}_j^{f_m}\right)$ be the distance between the images $\mathbf{I}_i$ and $\mathbf{I}_j$ evaluated in the feature space f_m. Given a reference image $\mathbf{q}$, the new representation of a generic image $\mathbf{I}_i$ in the *dissimilarity multi-space* is

$$\mathbf{I}'_i = \left[d_{f_1}\left(\mathbf{q}^{f_1}, \mathbf{I}_i^{f_1}\right), \ldots, d_{f_M}\left(\mathbf{q}^{f_M}, \mathbf{I}_i^{f_M}\right)\right]. \tag{2}$$

Summing up, while dissimilarity space are usually constructed by stacking dissimilarities from multiple prototypes, we propose to stack multiple dissimilarities originated by considering a single reference point, and measuring the distances from this point in different feature representations.

Let us have a close look on the choice of the reference image to be used in equation (2). When the first round of retrieval is performed, i.e., no feedback is available, we use the query image as the reference point. At each round of relevance feedback, the reference point is computed according to a "query shifting mechanisms", i.e., a mechanism designed to exploit relevance feedback by computing a new vector in the feature space such that its neighborhood contains relevant images with high probability [15]. In particular, we used a modified Rocchio formula, that has been proposed in the framework of the Bayes decision theory, namely Bayes Query Shifting (BQS) [6].

$$\mathbf{q}_{BQS} = \mathbf{m}_r + \frac{\sigma}{\|\mathbf{m}_r - \mathbf{m}_n\|}\left(1 - \frac{r-n}{\max(r,n)}\right)(\mathbf{m}_r - \mathbf{m}_n) \tag{3}$$

where $\mathbf{m}_r$ and $\mathbf{m}_n$ are the mean vectors, in each feature space, of relevant and non-relevant images respectively, σ is the standard deviation of the images belonging to the neighborhood of the original query, and r and n are the number of relevant and non relevant images retrieved after the latter iteration, respectively.

The choice of the query, and the BQS as the reference prototypes is twofold. First of all, as we are taking into account retrieval tasks in which the user performs a "query by example" search, and the BQS technique is aimed to represent the concept that the user is searching for by definition. On the other hand, the use of multiple images as prototypes can introduce some kind of "noise" because not all the images may exhibit the same "degree" of relevance to the user's needs. The second reason is that the use of a single prototype makes the search independent from the number of images in the database that are relevant to the user's query.

In the literature of content-based image retrieval, few works addressed the use of dissimilarity spaces to provide for a more effective representation. Some of the approaches proposed so far employed the original definition of dissimilarity space, where dissimilarities are computed by taking into account multiple

prototypes of relevant images [12,5]. Other authors have proposed to use the "dissimilarity space" technique for combining different feature space representations [2]. However, their approach is based on the computation of dissimilarity relationships between all the patterns in the dataset. Then, a number of prototypes are selected in each feature space, and the resulting dissimilarity spaces are then combined to attain a new multi-modal dissimilarity space. Thus the components of the resulting space are not related to the number of the original feature spaces, but they are related to the number of patterns used as prototypes to create the different dissimilarity spaces.

3 Nearest-Neighbor Relevance Feedback in the Dissimilarity Multi-space

The generation of the dissimilarity space is strictly related to the use of a Nearest-Neighbor approach to exploit relevance feedback in multiple feature spaces. In fact, the new space provides for a compact representation of patterns that ease the computation of nearest-neighbor relationships in multiple low-level feature representations. The dissimilarity representation computed with respect to a set of prototypes basically assumes that patterns belonging to the same category are represented as close points. Analogously, we expect that relevant images are represented as close points in the space made up of dissimilarities computed with respect to one reference point in multiple low-level feature spaces. The Nearest-Neighbor technique employed to exploit relevance feedback is based on the computation of a relevance score for each image according to its distance from the nearest relevant image, and the distance from the nearest non relevant image [4]. This score is further combined to a score related to the distance of the image from the point computed according to the BQS (Eq. 3), that is the likelihood that the image is relevant according to the users' feedback. The combined relevance score is computed as follows:

$$rel(\mathbf{I}'_i)_{stab} = \left(\frac{n/k}{1+n/k}\right) \cdot rel_{BQS}(\mathbf{I}'_i) + \left(\frac{1}{1+n/k}\right) \cdot rel_{NN}(\mathbf{I}'_i) \tag{4}$$

where n and k are the number of non-relevant images, and the whole number of images retrieved after the latter iteration, respectively. The two terms rel_{NN} and rel_{BQS} are computed as follows:

$$rel_{NN}(\mathbf{I}'_i) = \frac{\|\mathbf{I}'_i - NN^{nr}(\mathbf{I}'_i)\|}{\|\mathbf{I}'_i - NN^{r}(\mathbf{I}'_i)\| + \|\mathbf{I}'_i - NN^{nr}(\mathbf{I}'_i)\|} \tag{5}$$

where $NN(\mathbf{I}'_i)$ denotes the Nearest-Neighbor of $\mathbf{I}'_i$, and $\|\cdot\|$ is the Euclidean distance,

$$rel_{BQS}(\mathbf{I}'_i) = \frac{1 - e^{1-\left(d'\left(\mathbf{q}'_{BQS},\mathbf{I}'_i\right) / \max_i d'\left(\mathbf{q}'_{BQS},\mathbf{I}'_i\right)\right)}}{1-e} \tag{6}$$

where i is the index of all images in the database and $d'\left(\mathbf{q}'_{BQS},\mathbf{I}'_i\right)$ is the distance of image $\mathbf{I}'_i$ from the point computed according to Eq. 3.

The dissimilarity multi-space is included in a content-based retrieval system with Nearest-Neighbor relevance-feedback according to the following algorithm:

i) the user submits a query image $\mathbf{q}$. The distances $d_{f_m}\left(\mathbf{q}^{f_m}, \mathbf{I}_i^{f_m}\right)$, $m = 1, \ldots, M$, and $i = 1, \ldots N$ are computed, where M is the number of low-level features used to represent the images, and N is the number of the images in the database;

ii) in each feature space these distances are normalized between 0 and 1 and they are used to create, for each image $\mathbf{I}_i$, the new dissimilarity representation $\mathbf{I}'_i$, $i = 1, \ldots N$, according to Eq. 2;

iii) the Euclidean distances $d'\left(\mathbf{q}', \mathbf{I}'_i\right)$ between the dissimilarity representation of the query, and the dissimilarity representation of all the images are computed, and then sorted from the smallest to the largest;

iv) the first k images are labelled by the user as being relevant or not;

v) after the relevance feedback, the new reference point $\mathbf{q}_{BQS}$ is computed according to Eq. 3 in each feature space;

vi) the distances $d_{f_m}\left(\mathbf{q}_{BQS}^{f_m}, \mathbf{I}_i^{f_m}\right)$ are computed and normalized in each low-level feature space analogously to steps **i)** and **ii)**, where the query $\mathbf{q}$ is substituted with the new point $\mathbf{q}_{BQS}$. These distances are then used to create a new dissimilarity representation according to Eq. 2 where, again, the query $\mathbf{q}$ is substituted with the new point $\mathbf{q}_{BQS}$;

vii) in this new space, a score for all the images in the dataset is evaluated according to Eq. 4, where all the distances are computed according to the dissimilarity representation;

viii) all the images are sorted according to the value of the relevance score, and the first k images are labelled by the user as in step **iv)**;

ix) the algorithm starts again from step **iv)** until the user is satisfied.

4 Techniques for Combining Different Feature Spaces

In the previous sections we have mentioned a number of techniques that can be used to combine different image representations. In the following we will briefly review the six combination techniques that have been used in the experimental section for comparison purposes. Four combination methods aims to combine the relevance scores computed after relevance feedback, while the other two methods aim at combining distances.

The four techniques used to combine the relevance scores computed separately in each of the available feature spaces are the following:

$$score_{MAX}\left(\mathbf{I}_i\right) = \max_{f \in F}\left(score_f\left(\mathbf{I}_i\right)\right) \tag{7}$$

$$score_{MIN}\left(\mathbf{I}_i\right) = \min_{f \in F}\left(score_f\left(\mathbf{I}_i\right)\right) \tag{8}$$

$$score_{MEAN}\left(\mathbf{I}_i\right) = \frac{\sum_{f \in F} score_f(\mathbf{I}_i)}{|F|} \tag{9}$$

where F is the set of the feature spaces and $score_f\left(\mathbf{I}_i\right)$ is the relevance score evaluated in the feature space f. RR weight is the weighted sum of the relevance

scores, where the Relevance Rank Weights are obtained as in the following equation [7]

$$w_{RR_f} = \frac{\sum_{j \in R} \frac{1}{rank_f(\mathbf{I}_j)}}{\sum_{f' \in F} \sum_{j \in R} \frac{1}{rank_{f'}(\mathbf{I}_j)}} \tag{10}$$

and

$$score_{RRW}(\mathbf{I}_i) = \sum_{f \in F} w_{RR_f} \cdot score_f(\mathbf{I}_i) \tag{11}$$

where $f \in F$, $score_f(\mathbf{I}_i)$ is the relevance score evaluated in the feature space f, $rank_f(\mathbf{I}_j)$ is the rank of the image $\mathbf{I}_j$ according to $score_f$, and R is the set of the relevant images.

The other two combination methods are used to combine the distances computed in different feature spaces. One method computes the sum of the normalized distances (SUM), while the other method computes a "Nearest-Based" weighted sum (NBW) where the weights are computed in a similar way as in [14]:

$$w_f = \frac{\sum_{i \in R} \sum_{j \in R} d_f(I_i, I_j)}{\sum_{i \in R} \sum_{j \in R} d_f(I_i, I_j) + \sum_{i \in R} \sum_{h \in N} d_f(I_i, I_h)} \tag{12}$$

where $f \in F$, $d_f(\cdot)$ is a function that returns the distance between two images measured in the feature space f, and R, and N are respectively the set of the relevant and non-relevant images.

5 Experimental Results

5.1 Datasets

Experiments have been carried out using the Caltech-256 dataset, from the California Institute of Technology[1], that consists of 30607 images subdivided into 257 semantic classes [8]. Five different features have been extracted, namely the *Tamura* features [16] (18 components), the *Scalable Color* (64 components), *Edge Histogram* (80 components), *Color Layout* descriptors (12 components) [1], and the *Color and Edge Directivity Descriptor* (*Cedd*, 144 components) [3]. The open source library LIRE (Lucene Image REtrieval) has been used for feature extraction [11].

5.2 Experimental Set-Up

In order to test the performances, 500 query images have been randomly extracted from the dataset, covering all the semantic classes. The top twenty best-scored images for each query are returned to the user. Relevance feedback is

[1] http://www.vision.caltech.edu/Image_Datasets/Caltech256/

performed by marking images belonging to the same class of the query as relevant, and all other images in the top twenty as non-relevant. Performances are evaluated in terms of retrieval precision, and recall.

In order to evaluate the improvement attained by the proposed method in the next sub-section we will show the results attained separately in each feature space, and the performance related to the six combination techniques described in Section 4.

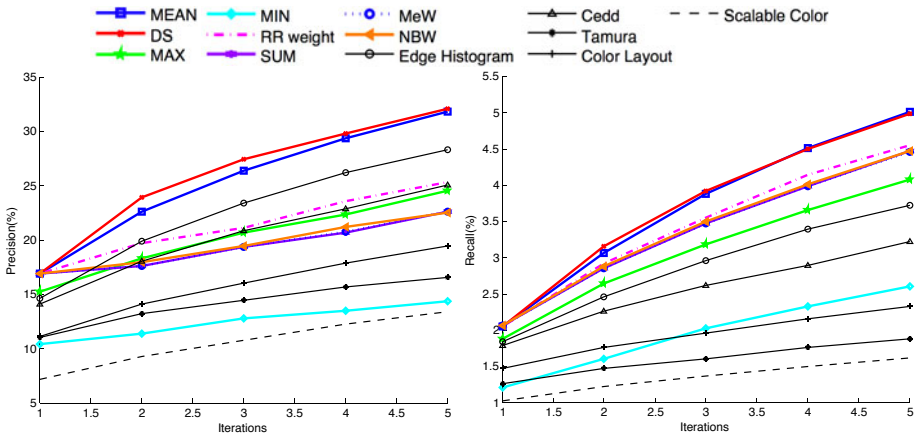
5.3 Results

Figures 1(a) and 1(b) show the performance of the proposed dissimilarity space representation compared to the six combination techniques described in Section 4, and the performance attained separately in each feature space.

By inspecting the behavior of the precision reported in Fig. 1(a), it can be easily seen that the highest performances are provided by the proposed dissimilarity (DS) based technique, and by the MEAN of the relevance scores. The combination of the relevance scores by the MAX and RR Weight rules allows attaining higher precision results than those attained by four out of the five features considered. This result is quite reasonable as typically the goal of the combination is to avoid choosing the worst problem formulation. In addition, it can be seen that the weighted combination (RR Weight) provides a lower result compared to the arithmetic MEAN, thus confirming the difficulty in providing an effective estimation of the weights. Finally, the worst result is attained by the MIN rule, that represents the logical AND function. Thus, we can conclude that, at least for the considered data set, the fusion of information from multiple feature spaces is more effective than the selection of one feature space. In addition, the results attained by the proposed DS space and the MEAN rule confirm that an unweighted combination can be more effective than weighted combination or selection. If we consider the two techniques based on the combination of the distances, namely the SUM rule and the NBW rule, we can see that their performances are lower than those of the techniques based on the combination of scores.

If we consider the recall (Fig. 1(b)), we can see that MEAN rule, and the DS approaches are still the best technique. It is worth noting that all the combination techniques, except for the MIN, provide an improvement in recall with respect to the performance attained in the individual feature spaces. In particular the fusion techniques working at the distance level provided good results, quite close to those attained by the RR Weight rule.

The proposed approach based on dissimilarity spaces not only allows attaining good performances in terms of precision and recall, but also requires a low effort in terms of computational time if compared to other combination techniques (Fig. 2). This effect can be explained by considering that all the distances from each image to the query are computed only once, during the first retrieval iteration. All the following iterations can exploit this result, and all the computation are made in the low-dimensional dissimilarity space.



(a) Precision (b) Recall

Fig. 1. Caltech-256 Dataset - Precision and Recall for 5 rounds of relevance feedback

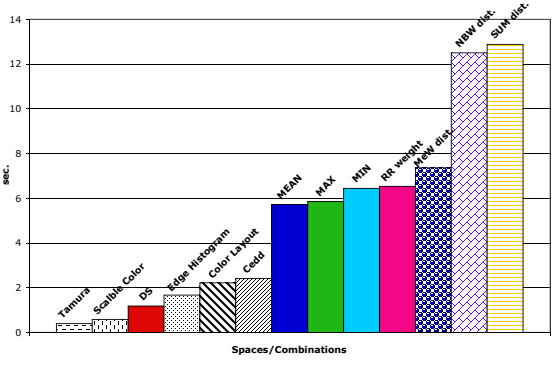


Fig. 2. Caltech-256 Dataset - Mean execution time for 1 round of relevance feedback

6 Conclusion

In this paper we proposed a technique that addresses the problem of the combination of multiple-features for image retrieval with relevance feedback. We showed that a dissimilarity representation of images allows to combine nicely and effectively a number of feature spaces. In particular, reported results show that the proposed technique allows outperforming other combination methods, both in terms of performances and computational time. In addition, this method scales well with the number of features, as the addition of one feature space adds one component to the dissimilarity vector, and the distances in the original feature spaces from the query needs to be computed only once.

References

1. Information technology - Multimedia content description interface - Part 3: Visual, ISO/IEC Std. 15938-3:2003 (2003)
2. Bruno, E., Moenne-Loccoz, N., Marchand-Maillet, S.: Learning user queries in multimodal dissimilarity spaces. In: Detyniecki, M., Jose, J.M., Nürnberger, A., van Rijsbergen, C.J. (eds.) AMR 2005. LNCS, vol. 3877, pp. 168–179. Springer, Heidelberg (2006)
3. Chatzichristofis, S.A., Boutalis, Y.S.: Cedd: Color and edge directivity descriptor: A compact descriptor for image indexing and retrieval. In: Gasteratos, A., Vincze, M., Tsotsos, J.K. (eds.) ICVS 2008. LNCS, vol. 5008, pp. 312–322. Springer, Heidelberg (2008)
4. Giacinto, G.: A nearest-neighbor approach to relevance feedback in content based image retrieval. In: CIVR 2007: Proceedings of the 6th ACM International Conference on Image and Video Retrieval, pp. 456–463. ACM, New York (2007)
5. Giacinto, G., Roli, F.: Dissimilarity representation of images for relevance feedback in content-based image retrieval. In: Perner, P., Rosenfeld, A. (eds.) MLDM 2003. LNCS, vol. 2734, pp. 202–214. Springer, Heidelberg (2003)
6. Giacinto, G., Roli, F.: Bayesian relevance feedback for content-based image retrieval. Pattern Recognition 37(7), 1499–1508 (2004)
7. Giacinto, G., Roli, F.: Nearest-prototype relevance feedback for content based image retrieval. In: ICPR (2), pp. 989–992 (2004)
8. Griffin, G., Holub, A., Perona, P.: Caltech-256 object category dataset. Tech. Rep. 7694, California Institute of Technology (2007), `http://authors.library.caltech.edu/7694`
9. Kittler, J., Hatef, M., Duin, R.P.W., Matas, J.: On combining classifiers. IEEE Trans. Pattern Anal. Mach. Intell. 20(3), 226–239 (1998)
10. Kuncheva, L.I.: Combining Pattern Classifiers: Methods and Algorithms. Wiley, Chichester (2004)
11. Lux, M., Chatzichristofis, S.A.: Lire: lucene image retrieval: an extensible java cbir library. In: MM 2008: Proceeding of the 16th ACM International Conference on Multimedia, pp. 1085–1088. ACM, New York (2008)
12. Nguyen, G.P., Worring, M., Smeulders, A.W.M.: Similarity learning via dissimilarity space in cbir. In: Wang, J.Z., Boujemaa, N., Chen, Y. (eds.) Multimedia Information Retrieval, pp. 107–116. ACM, New York (2006)
13. Pekalska, E., Duin, R.P.W.: The Dissimilarity Representation for Pattern Recognition: Foundations And Applications (Machine Perception and Artificial Intelligence). World Scientific Publishing Co., Inc., River Edge (2005)
14. Piras, L., Giacinto, G.: Neighborhood-based feature weighting for relevance feedback in content-based retrieval. In: WIAMIS, pp. 238–241. IEEE Computer Society, Los Alamitos (2009)
15. Rui, Y., Huang, T.S.: Relevance feedback techniques in image retrieval. In: Lew, M.S. (ed.) Principles of Visual Information Retrieval, pp. 219–258. Springer, London (2001)
16. Tamura, H., Mori, S., Yamawaki, T.: Textural features corresponding to visual perception. IEEE Trans. Systems, Man and Cybernetics 8(6), 460–473 (1978)
17. Zhang, L., Lin, F., Zhang, B.: Support vector machine learning for image retrieval. In: ICIP (2), pp. 721–724 (2001)