Classification-based multimodality fusion approach for similarity ranking

Emilia López-Iñesta, Miguel Arevalillo-Herráez, Francisco Grimaldo
Department of Computer Science
University of Valencia
Avda. de la Universidad s/n. 46100-Burjassot (SPAIN)
Email: eloi@alumni.uv.es,{miguel.arevalillo,francisco.grimaldo}@uv.es

***Abstract*—The need for similarity rankings is common to a wide diversity of Pattern Recognition problems. When multiple modalities are available, effective combination methods that exploit the information contained in the different representations are required. In this paper, a method for effectively combining the information in the different modalities is presented. The method adopts the common framework used in metric learning and assumes that training samples are available, in the form of pairs of objects labeled as similar or dissimilar. For each pair, one or more distance measures are computed in each representation space, and these are used to train a soft classifier. Estimated class conditional probabilities are then used as scores for ranking purposes. The approach has been tested and compared to other existing combination methods in an image retrieval context, showing competitive results.**

I. Introduction

Similarity rankings are a common instrument used in retrieval problems (e.g., image, video). In the single modality case, distance functions can directly be used for ranking purposes e.g., Manhattan, Euclidean, Mahalanobis, Canberra, etc. When multiple representations are available, it is usually more effective to compute distances in each space and then combine them to produce a single score.

One typical approach to combining distances consists of using a weighted linear combination, sometimes preceded by some kind of normalization e.g., Gaussian [1], [2] or re-scaling [3]. Weights can be directly indicated by the user or automatically calculated by using a training set. In the latter case, some type of optimization algorithm is used to find a solution that best fits the training data. Genetic algorithm and linear optimization techniques have commonly been used for this purpose e.g., [1], [4]. Other more elaborate approaches have searched for non-linear combinations, by using genetic programming [5] or probabilistic frameworks [6].

Existing proposals use training data in different formats. For example, each training element in [5] is composed of a query and the set of images that should ideally be retrieved. In [1], the best match image for each query needs also to be provided. In [4], triplets of images are used instead. Each triplet is composed of a focal image and two other images, with a label indicating which one is more similar to the focal image.

In this paper, we first present a classification based technique to combine multiple modalities into a single similarity measure. As many classical metric learning approaches [7], [8], [9], binary pairwise comparisons between samples are used for training. Available samples are used to train a soft classifier, which is then used to obtain a score for any new pair. The approach has been tested in a Content-Based Image Retrieval (CBIR) context, and compared to other existing methods in three databases with distinct characteristics.

As a second contribution, we have extended the technique by using several distance measures in each subspace. A significant improvement has also be obtained in this case.

II. The Problem

A. Notation

Let us assume a repository $\mathcal{X}$ containing a set of objects $x_i, i = 1, 2, \ldots, m$ conveniently represented in a particular feature space where the whole set of available (vector) features is designated as $\mathbb{F}$ and $\{\mathbb{F}^{(u)}\}_{u=1}^{n}$ is a family of subspaces of $\mathbb{F}$.

Let us also assume that a similarity/dissimilarity measure has been defined in each subspace u using the subset of features $\mathbb{F}^{(u)}$ and let this be expressed as

$$s_u : \mathbb{F}^{(u)} \times \mathbb{F}^{(u)} \longrightarrow \mathbb{R} \tag{1}$$

The feature based representation used and the family of similarity measures $\{s_u\}_{u=1}^{n}$ make it possible to define the function s

$$s : \mathcal{X} \times \mathcal{X} \longrightarrow \mathbb{R}^n \tag{2}$$

by which a pair of objects $pair_k = (x_i, x_j)$ is associated with a tuple of n values $< v_{k,1} \ldots v_{k,n} >$, where each value $v_{k,u}$ represents the similarity between the pair of objects $pair_k$ as produced by function s_u (see Figure 1).

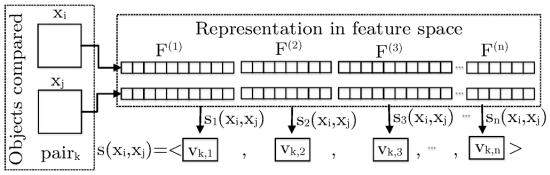


Fig. 1. Function s maps two objects x_i and x_j to an n-tuple of real values. Each value represents the similarity between the two objects in a particular feature space.

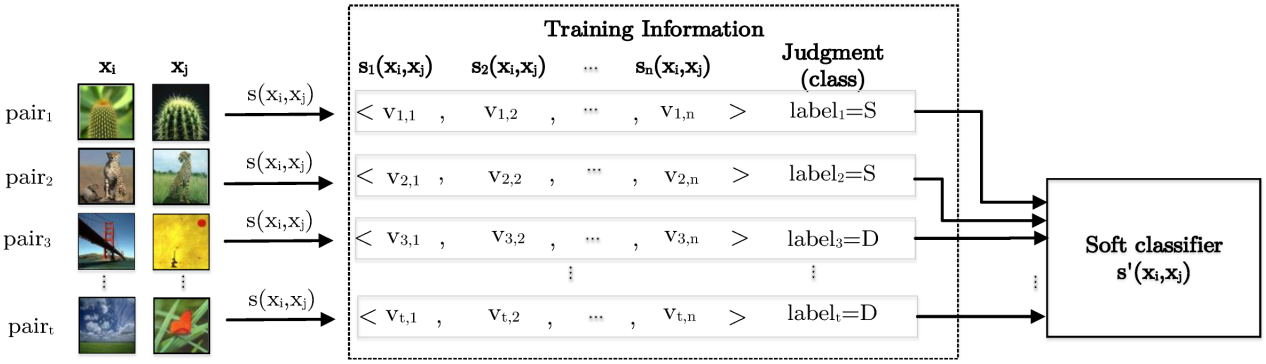


Fig. 2. Training of the soft classifier to compute the function s'. Each labeled image pair $pair_k$ results in a labeled training instance. To this end, the function s is used to compute a tuple of n values, each using a similarity measures defined in the corresponding subspace.

B. Problem formulation

Let us assume that a similarity relation among the objects in $\mathcal{X}$ exists, so that any pair of objects $pair_k = (x_i, x_j)$ can be considered either as similar or dissimilar. Let us also assume that a list of t binary judgments is available and that each entry in this list is composed of a pair and a label $(pair_k, label_k)$, which takes value S (similar) or D (dissimilar).

By using the notation above, our objective is to use the information available to learn a new function s'

$$s' : \mathbb{R}^n \longrightarrow [0, 1] \tag{3}$$

that is able to convert the n-tuple produced by function s into a single value in the closed interval $[0, 1]$, which represents the similarity between the object pair according to the same (and maybe unknown) criteria that was used in the list of judgments provided for training.

Once the function s' has been obtained, the composition of the functions s and s' allows one to associate any pair of objects with a continuous similarity value, which can be used for ranking purposes.

$$s' \circ s : \mathcal{X} \times \mathcal{X} \longrightarrow [0, 1] \tag{4}$$

C. Antecedents

Some typical approaches to tackle this problem assume that all measures s_u have equal importance, are statistically independent, and/or the similarity values they produce follow a Gaussian distribution. These assumptions make it possible to use relatively simple probabilistic frameworks to compute an estimate for the probability $p_u(similar|x_i, x_j)$ in each feature space, and then combine the estimates into a single score. The use of normalization approaches, the assignment of different weights to each set of descriptors [10], [11], [3], and the use of the product and sum rules [12] are common in this context.

In [2], a Gaussian normalization of the distance values obtained in each subspace was proposed. This consists in a mapping $s_u(x_i, x_j) \rightarrow (s_u(x_i, x_j) - \mu_u)/3\sigma_u$, where μ_u and σ_u represent the mean and standard deviation of the values produced by the function s_u. Under the Gaussianity assumption, this normalization ensures a 99% probability that the normalized value be in $[-1, 1]$. Then, the results obtained in each subspace are added to yield a single score. A major limitation of this method is that equal emphasis is placed in all descriptors.

To reflect the user's different emphasis of each representation in the overall similarity, the replacement of the standard sum rule by a linear combination of the distance values has also been proposed. In [10], a Gaussian normalization was also used, but it was combined with a method that dynamically updated the query weights by using relevance feedback information.

In a more recent work [6], the Gaussianity assumption was made unnecessary. In this case, training samples are used to build a probability mapping function $p_u(similar|s_u(x_i, x_j))$ in each subspace $\mathbb{F}^{(u)}$. Functions p_u are computed independently in each subspace, and relate any similarity value produced by the function s_u for a pair of images $pair_k$ to the probability that images i and j would be considered similar by a general user. To this end, kernel density estimation [13] is used, and the distribution of similarity values for positive and negative pairs in the training set are both taken into account. Finally, the desired function s' is given as a two step process. First, the values s_u obtained by each similarity function are mapped to probabilities by using the corresponding function p_u. Then, statistical independence is assumed and the values in each subspace are multiplied to get the final similarity value.

D. Proposal

To remove the need for the statistical independence assumption, this problem can also be approached from a classical supervised learning perspective, and solved by using a standard soft classifier in the similarity space defined by the function s. At a first stage, labeled pairs are used to train the classifier. To this end, the similarity function s is used to convert each training pair $pair_k = (x_i, x_j)$ into a labeled input vector $< s_1(x_i, x_j), s_2(x_i, x_j), \cdots, s_n(x_i, x_j) >$. The inferred function represents the desired function s' (see Figure 2).

Once the classifier has been trained, a similarity score can easily be computed for any new unseen pair of objects (x_i, x_j). This is done by providing the vector $< s_1(x_i, x_j), s_2(x_i, x_j), \cdots, s_n(x_i, x_j) >$ as an input to the

classifier. The conditional probability of the similar class S represents $p(similar|x_i, x_j)$ and hence can be used as a similarity estimate (see Figure 3).

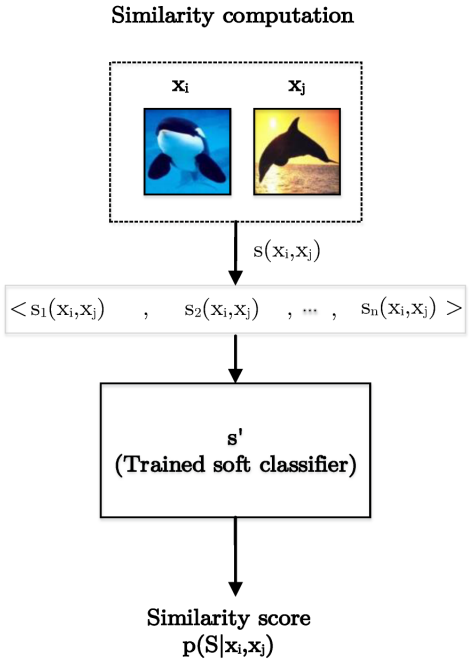


Fig. 3. Computation of a similarity score. To compute a similarity score between a new pair of objects, the new pair is provided as an input to the classifier. The conditional probability of the similar class S computed by the classifier is used as a similarity score.

Note that the approach presented does not depend on whether the individual functions $s_u, u = 1 \dots n$ are similarity or dissimilarity measures. In addition, the final similarity score can be turned into a dissimilarity (distance) score by considering the conditial probability of the dissimilar class (D).

E. Extension

Other than the combination method used, the quality of the final score is highly sensitive to the family of similarity measures $\{s_u\}_{u=1}^{n}$ used in the definition of the function s. The selection of the most appropriate measure in each particular subspace $\mathbb{F}^{(u)}$ is context dependent and varies across different data sets.

Given that different similarity measures may result in significant variations in performance, additional gains can potentially be achieved by embedding several similarity functions per subspace into the approach.

For illustrative purposes, we have used different L_p norms (Minkowski distances). These are widely used dissimilarity measures. In addition, the relatively large differences in performance of different L_p norms on the same data [14], [15] suggest that they may be combined to obtain improved results.

III. Empirical Evaluation

A number of comparative experiments to validate the success of the proposal have been carried out in a CBIR context. In a first experiment, results obtained with the plain proposal by using Euclidean distances at each subspace are compared to the ones obtained by using: (a)

1) the probabilistic method presented in [6]. This method uses a similar problem formulation based on exactly the same training information.
2) a standard Gaussian normalization as described in [2], [10]. This consists of a mapping function $d_i \rightarrow (d_i - \mu)/3\sigma$, where μ and σ represent the mean and the standard deviation of the distance d_i.
3) a linear combination of the Euclidean distances computed at each subspace, with all distances contributing equally.

In the remaining of this section, these three approaches will be referred to as probabilistic, Gaussian normalization and linear combination, respectively.

In a second experiment, the additional benefits obtained by using the proposed extension are evaluated. In this case, results are compared to the ones obtained by using the plain proposal on different L_p norms.

TABLE I. Details of the three databases used in the experiments.

Name	Size	Descriptors	Dimensions	Categories
Small	1508	10	12-7-3-4-11-11-6-30-10-10	29
Art	5476	10	12-7-3-4-11-11-6-30-10-10	63
Corel	30000	4	32-32-9-16	71

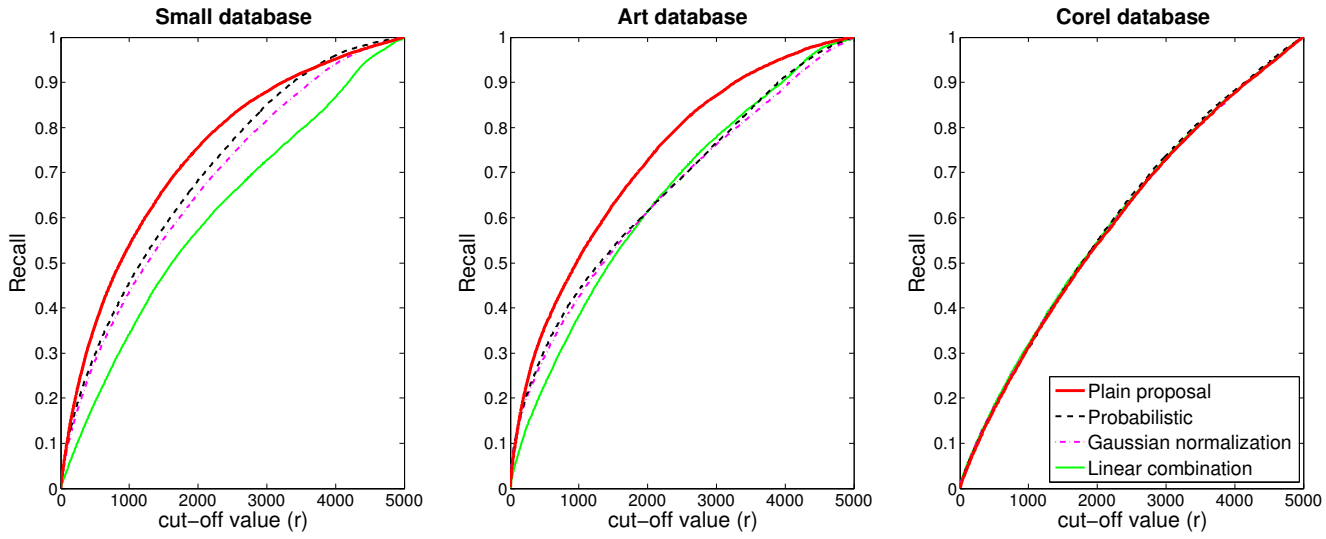
A. Databases

Experimental results have been validated in three different databases that have been previously used in other similar studies e.g., [16], [17]. Table I shows a summary of the details of each database. The size indicates the number of images in the repository. The descriptors column refers to the number of different subspaces n. Dimensions refers to the number of features in each subspace $\mathbb{F}^{(u)}$. The categories column represents the number of classes in each repository, according to the database classification provided. These repositories, along with further details about their contents can be found in http://kdd.ics.uci.edu/databases/CorelFeatures and http://www.uv.es/arevalil/dbImages/, for the Corel and the other two databases, respectively.

B. Implementation details

Despite that several classification methods have been attempted, only results for the Support Vector Machine (SVM) [18] are reported in this paper. These were consistently better than those obtained by using other classification methods, namely Naive Bayes [19] and ID3 [20]. The kernel chosen has been a Gaussian radial basis function. The parameters γ and C have been tuned by using an exhaustive grid search on a held out validation set composed of a 30% partition of the training data ($C \in \{0.1, 1, 10\}$ and $\gamma \in \{0.0001, 0.001, 0.01, 0.1, 1, 10\}$).

To test the extended method, we have used different p-Norms ($p \in \{0.5, 1, 1.5, 2\}$) in each subspace $\mathbb{F}^{(u)}$. This implies that the dimension of the training vectors is multiplied



(a) (b) (c)

Fig. 4. Comparative performance between the plain proposal and the probabilistic, gaussian normalization and linear combination approaches in a) the small database; b) the Art database; and c) the Corel database.

by a factor of 4. Fractional values of p have been included because they have been reported to provide more meaningful results for high dimensional data, both from the theoretical and empirical perspective [14]. In addition, these results have been confirmed in a CBIR context [15].

C. Experimental setting

Results obtained with each distance combination are evaluated by using a ranking-based method. At a first stage, the same 5 000 pairs of images (half similar and half non-similar) are supplied to all methods as training data (except for the Gaussian normalization and linear combination methods, which do not use any training information). These are randomly selected from the set of all possible pairs in the database. The available categories at each repository are used to simulate user judgments, so that images under the same category are considered subjectively similar.

Results are evaluated on a new set composed of another 5 000 image pairs. To this end, all methods evaluated are used to rank the 5 000 pairs in descending order of similarity. Then, the fraction of similar pairs that appear in between the first r positions of the ranking (recall) is measured for every value of $r = [1, 5\,000]$, and used to visually compare the performance of all methods.

To increase the reliability of the results, all experiments have been repeated 100 times and results have been averaged.

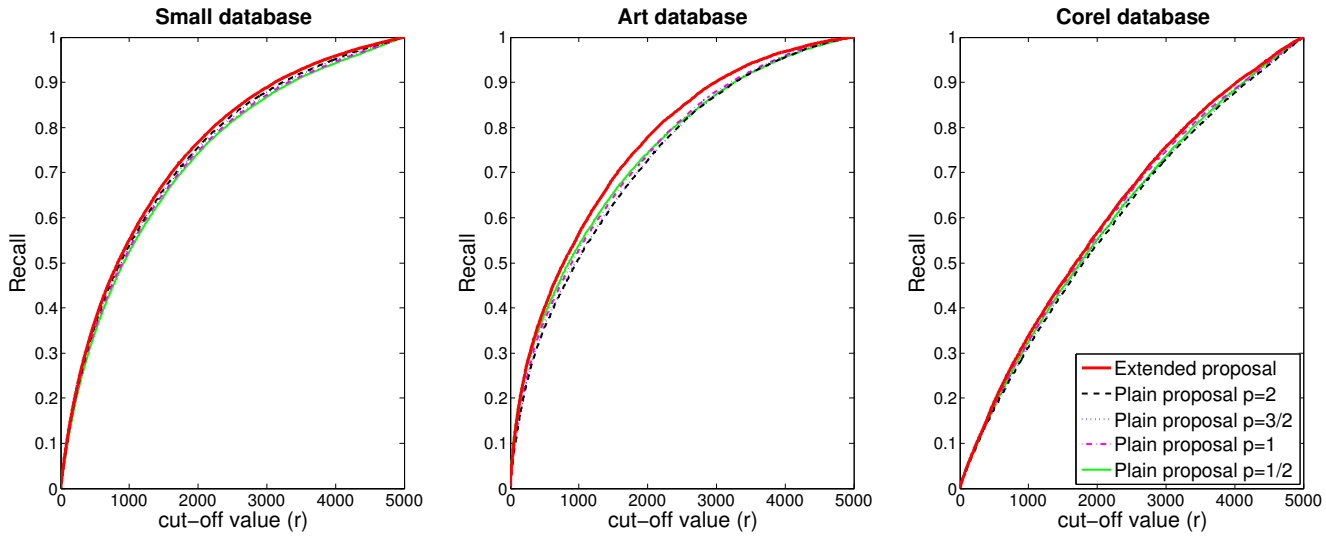
D. Results

Fig. 4 shows the comparative performance of plain proposal against the probabilistic, gaussian normalization and linear combination approaches. This plot reveals noticeable performance differences in recall in the Small and Art databases, clearly in favor of the proposed approach. In the Corel database, very similar recall results can be observed for the three methods in all databases (only very small differences in favour of the probabilistic method can be observed for ranking positions $r > 2\,000$). This may be due to the higher subjectivity involved in the classification of the images in this large-sized database, where different criteria might have been applied (also by different people), that results in a classification that considers possibly similar concepts under different labels. As an example, the Corel database includes Insects and Insects II as two different categories. Hence, our experimental setting would consider images in these two groups as dissimilar.

In Fig. 5, the extended method using multiple p-norms is compared to the plain combination proposal, when using different L_p norms to define the similarity functions s_u used in each subspace $\mathbb{F}^{(u)}$. A first interesting result relates to the comparative performance of the different L_p norms used. No consistent behavior is observed, and the best norm in one database can be the worst when used with another data set. For example, norm 0.5 performs the best in the Art database but the worst in the small repository.

A second and more relevant result relates to the fact that the performance when the four L_p norms are combined is always equal or better than when using any of the single norms. This is an interesting result because despite the large literature on using different norms, most approaches aim at selecting the one that offers the highest performance, rather than combining them to improve retrieval results e.g., [14], [15], [21]. The results obtained in this experiment suggest that important performance gains can be obtained by combining L_p norms.

For clarity reasons, Fig. 6 shows the performance of the extended approach when compared to the other competing methods. It can be observed that the extended proposal consistently performs better than the original one, and also outperforms the other three methods in all databases. Again, differences are specially relevant in the Small and the Art databases. In the Corel repository, only a small difference in favor of the extended proposal can be observed. This is in



(a) (b) (c)

Fig. 5. Performance gains by using the proposed extension in a) the small database; b) the Art database; and c) the Corel database.

contrast to the result offered by the plain proposal, which was slightly worse than the Probabilistic method for values of $r > 2000$.

IV. Conclusions

In this paper, a distance combination method based on binary pairwise comparisons between samples has been presented. In addition, an extension that uses several distances in each subspace has been introduced. In particular, four L_p norms per subspace have been used. Results by using this approach have clearly outperformed the ones obtained by using two other competitive methods. In addition, they have provided an evidence of the potential of combining L_p norms for retrieval tasks. The most appropriate L_p norm to be used with the approach is an important topic that deserves further investigation. In addition, the combination of other similarity/distance measures is currently an issue under study.

The utility of the methods proposed depend on the particular application context. One typical application of distance combination methods is as part of classification approaches. However, the proposed methods result in a non-metric estimate. Although this is not an inconvenience for ranking purposes, the effect when used in conjunction with classification algorithms (e.g., nearest neighbor) needs to be investigated in more depth.

Further research that could potentially improve the results obtained with the extended approach include a) fine tuning the SVM by trying a wider range of parameter values or other different kernels; b) using cross validation to determine the most convenient set of distances to be used in each subspace and c) using classification methods other than the ones used in this work.

V. Acknowledgements

This work has been supported by the Spanish Ministry of Science and Innovation through project TIN2011-29221-C03-02

References

[1] H. Shao, J.-W. Zhang, W.-C. Cui, and H. Zhao, "Automatic feature weight assignment based on genetic algorithm for image retrieval," in *IEEE International Conference on Robotics, Intelligent Systems and Signal Processing*, vol. 2, 2003, pp. 731–735.

[2] Q. Iqbal and J. Aggarwal, "Combining structure, color and texture for image retrieval: A performance evaluation," in *16th International Conference on Pattern Recognition (ICPR)*, Quebec City, QC, Canada, August 2002, pp. 438–443.

[3] G. Giacinto and F. RoIi, "Nearest-prototype relevance feedback for content based image retrieval," in *Proceedings of the 17th International Conference on Pattern Recognition (ICPR)*, vol. 2, 2004, pp. 989–992.

[4] A. Frome, Y. Singer, and J. Malik, "Image retrieval and classification using local distance functions," in *Neural Information Processing Systems Foundation (NIPS)*, Vancouver, Canada, December 2006.

[5] R. da S. Torres, A. X. Falcão, M. A. Gonalves, J. P. Papa, B. Zhang, W. Fan, and E. A. Fox, "A genetic programming framework for content-based image retrieval," *Pattern Recognition*, vol. 42, no. 2, pp. 283 – 292, 2009.

[6] M. Arevalillo-Herráez, J. Domingo, and F. J. Ferri, "Combining similarity measures in content-based image retrieval," *Pattern Recognition Letters*, vol. 29, no. 16, pp. 2174–2181, 2008.

[7] E. P. Xing, A. Y. Ng, M. I. Jordan, and S. Russell, "Distance metric learning, with application to clustering with side-information," in *Advances in Neural Information Processing Systems 15*. MIT Press, 2002, pp. 505–512.

[8] J. V. Davis, B. Kulis, P. Jain, S. Sra, and I. S. Dhillon, "Information-theoretic metric learning," in *International Conference on Machine Learning*, Corvalis, Oregon, USA, June 2007, pp. 209–216.

[9] A. Pérez-Suay, F. J. Ferri, and M. Arevalillo-Herráez, "Passive-aggressive online distance metric learning and extensions," *Progress in AI*, vol. 2, no. 1, pp. 85–96, 2013.

[10] Y. Rui, T. S. Huang, M. Ortega, and S. Mehrotra, "Relevance feedback: a power tool for interactive content-based image retrieval," *IEEE Trans. Circuits Syst. Video Techn.*, vol. 8, no. 5, pp. 644–655, 1998.

[11] Q. Zhang and E. Izquierdo, "Optimizing metrics combining low-level visual descriptors for image annotation and retrieval," in *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, vol. 2, 2006, pp. 405–408.

[12] J. Kittler, M. Hatef, R. P. W. Duin, and J. Matas, "On combining classifiers," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 20, no. 3, pp. 226–239, 1998.

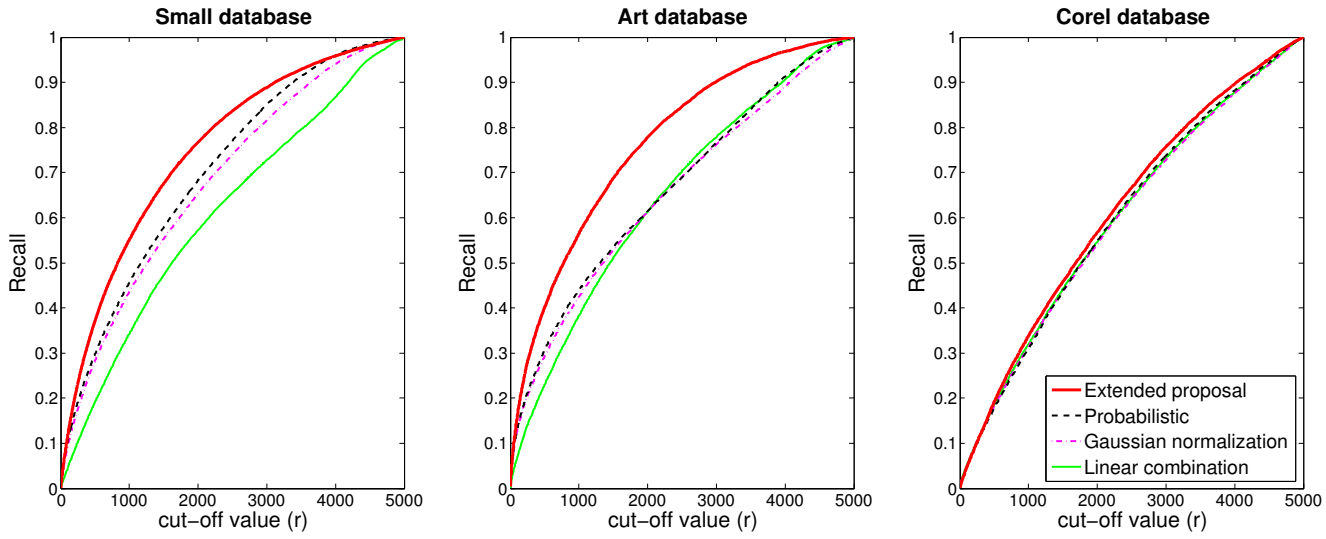


Fig. 6. Comparative performance between the extended approach and the sum and probabilistic methods in a) the small database; b) the Art database; and c) the Corel database.

[13] R. O. Duda, P. E. Hart, and D. G. Stork, *Pattern Classification (2nd Edition)*. Wiley-Interscience, 2000.

[14] C. Aggarwal, A. Hinneburg, and D. Keim, "On the surprising behavior of distance metrics in high dimensional space," in *Database Theory ICDT 2001*, ser. Lecture Notes in Computer Science, J. Bussche and V. Vianu, Eds. Springer Berlin Heidelberg, 2001, vol. 1973, pp. 420–434.

[15] P. Howarth and S. Rüger, "Fractional distance measures for content-based image retrieval," in *Proceedings of the 27th European conference on Advances in Information Retrieval Research (ECIR)*. Berlin, Heidelberg: Springer-Verlag, 2005, pp. 447–456.

[16] M. Arevalillo-Herráez, M. Zacarés, X. Benavent, and E. de Ves, "A relevance feedback CBIR algorithm based on fuzzy sets," *Signal Processing: Image Communication*, vol. 23, no. 7, pp. 490–504, 2008.

[17] M. Arevalillo-Herráez and F. J. Ferri, "An improved distance-based relevance feedback strategy for image retrieval," *Image and Vision Computing*, vol. 31, no. 10, pp. 704 – 713, 2013.

[18] A. Christmann and I. Steinwart, *Support Vector Machines*. Springer New York, 2008.

[19] G. H. John and P. Langley, "Estimating continuous distributions in bayesian classifiers," in *Proceedings of the Eleventh conference on Uncertainty in artificial intelligence*, ser. UAI'95. San Francisco, CA, USA: Morgan Kaufmann Publishers Inc., 1995, pp. 338–345.

[20] J. R. Quinlan, "Induction of decision trees," *Mach. Learn.*, vol. 1, no. 1, pp. 81–106, Mar. 1986.

[21] J. Zhang and L. Ye, "An unified framework based on p-norm for feature aggregation in content-based image retrieval," in *Ninth IEEE International Symposium on Multimedia (ISM)*, 2007, pp. 195–201.