Learning User Queries in Multimodal Dissimilarity Spaces

Eric Bruno, Nicolas Moenne-Loccoz, Stéphane Marchand-Maillet*

Computer Vision and Multimedia Laboratory, University of Geneva
eric.bruno@unige.ch

Abstract. Different strategies to learn user semantic queries from dissimilarity representations of video audio-visual content are presented. When dealing with large corpora of videos documents, using a feature representation requires the online computation of distances between all documents and a query. Hence, a dissimilarity representation may be preferred because its offline computation speeds up the retrieval process. We show how distances related to visual and audio video features can directly be used to learn complex concepts from a set of positive and negative examples provided by the user. Based on the idea of dissimilarity spaces, we derive three algorithms to fuse modalities and therefore to enhance the precision of retrieval results. The evaluation of our technique is performed on artificial data and on the complete annotated TRECVID corpus.

1 Introduction

Determining semantic concepts by allowing users to iteratively refine their queries is a key issue in multimedia content-based retrieval. The relevance feedback loop allows to build complex queries made out of positive and negative documents as examples. From this training set, a learning process should then extract relevant documents from feature spaces. Many relevance feedback techniques have been developed that operate directly in the feature space [3, 11, 13, 15].

Describing content of videos requires to deal in parallel with many high-dimensional feature spaces expressing the multimodal characteristics of the audiovisual stream. This mass of data makes retrieval operations computationally expensive when dealing directly with features. The simplest task of computing the distance between a query and all other elements becomes infeasible when involving tens of thousands of documents and thousands of feature space components. This problem is even more sensible when the similarity measures are complex functions or procedures, such as prediction functions for temporal distances [2] or graph exploration for semantic similarities [10].

Another aspect to prefer similarities rather than features is the multimodal fusion problem. Dealing directly with multimodal and heterogeneous features

* This work is funded by the Swiss NCCR (IM)2 (Interactive Multimodal Information Management).

imposes to build complex learning setup that need to take into account every feature metric. On the opposite, the similarity approach provides us an homogeneous framework where the learning process consists in combining various distance measurements, whatever the features involved into the problem are.

A solution to allow on-line interaction would be to compute off-line monomodal dissimilarity relationships between elements and to use the dissimilarity matrices or distance-based indexing structures [4] as an index for retrieval operations. The problem is then to find distance-based solutions that go beyond the classical k-NN approaches [1, 8] in order to perform discriminative classification that provides effective retrieval of semantic concepts. Pekalska *et al* [9] have proposed dissimilarity spaces where objects are represented not by their features but by their relative dissimilarities to a set of selected objects. These representations seem to form a convenient approach to tackle the similarity-based indexing and retrieval problem.

In this paper, we show how dissimilarities can be used to build low-dimensional multimodal representation spaces where learning machines, (*eg* SVM), could operate and investigate several strategies to fuse the modalities. Our thorough evaluation on both artificial data and the complete TRECVID news broadcast corpus shows that multimodal dissimilarity spaces allow to perform effective retrieval of video documents using real time interaction.

2 Query-dependent dissimilarity space

In the proposed framework, users formulate complex queries by iteratively providing positive and negative examples in a relevance feedback loop. From this training data, the aim is to perform a real-time dissimilarity-based classification that will return relevant documents to user. We present in the following the dissimilarity space introduced by Pekalska *et al* in [9] and show how it can be adapted to provide us with a low-dimensional approximation of the original feature space where an efficient classification could be performed.

Let $d(\mathbf{x}_i, \mathbf{x}_j)$ be the distance between elements i and j according to their descriptors $\mathbf{x} \in \mathcal{F}$. $\mathcal{F}$ expresses the (unavailable) original feature space. The dissimilarity space is defined as the mapping $\mathbf{d}(\mathbf{z}, \Omega) : \mathcal{F} \rightarrow \mathbb{R}^N$ given by:

$$\mathbf{d}(\mathbf{z}, \Omega) = [d(\mathbf{z}, \mathbf{x}_1), d(\mathbf{z}, \mathbf{x}_2), \dots \, d(\mathbf{z}, \mathbf{x}_N)].$$

The representation set $\Omega = \{\mathbf{x}_1, \dots, \mathbf{x}_N\}$ is a subset of N objects defining the new space. The new "features" of an input element are now its dissimilarity values with the representation objects. As a consequence, learning or classification tools for feature representations are also available to deal with the dissimilarities.

The dimensionality of the dissimilarity space is directly linked to the size of Ω, which controls the approximation made on the original feature space (such an approximation could be computed using projection algorithms like classical scaling [5]). Increasing the number of elements in Ω increases the representation accuracy. On the other hand, a well-chosen space of low dimension would be more

effective for further learning processes as it avoids the *curse of dimensionality* problem and reduces computation load.

The selection of a good representation set may be driven by considerations on the particular learning problem we are dealing with. Let us denote the query as the set T of positive and negative training examples (respectively denoted $\mathcal{P}$ and $\mathcal{N}$ with $T = \mathcal{P} \cup \mathcal{N}$). As mentioned by Zhou *et al* [15], we are generally dealing with a $1 + x$ class setup with 1 class associated to positives and x to negatives. Dedicated algorithms, such as *Bias-Map* [15, 14], have been developed to tackle this classification problem. In our case, the choice of the set Ω offers us the possibility to turn the problem into a more classical formulation: Selecting the representation set as the set of positive examples $\mathcal{P}$ turns the problem into a binary classification. Indeed, assuming that the positive examples are close to each other while being far from negative examples, the vectors $\mathbf{d}(\mathbf{z}_{i\in\mathcal{P}}, \mathcal{P})$ (*within* scatter) have norms lower than vectors $\mathbf{d}(\mathbf{z}_{i\in\mathcal{N}}, \mathcal{P})$ (*between* scatter), leading to a binarization of the classification, as illustrated in figure 1 with artificial data In this particular case, the learning task does not consist anymore in estimating the circular distribution of the negative class but a simpler function that separate the positive class, close to the origin, to the rest of the space.

The second advantage of selecting $\mathcal{P}$ as the representation set is that it readily induces to work in a low dimensional space of $p = |\mathcal{P}|$ components, where online learning processes are dramatically speed-up.

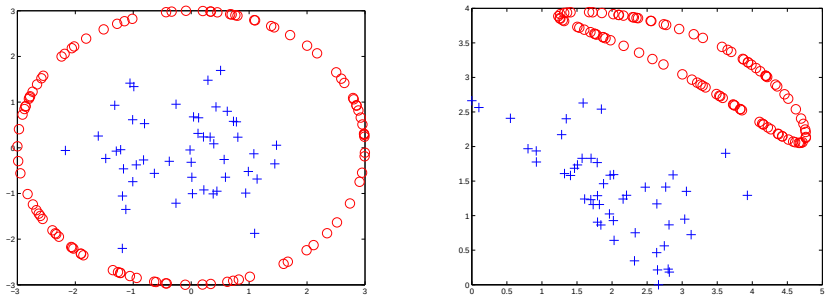

Fig. 1. The $1+x$ class problem in feature space (left) and 2D dissimilarity space (right) where the representation objects are two points from the central class (cross)

3 Multimodal dissimilarity space

A multimodal description of multimedia data provides a number of feature spaces (one or more per modality). Each of them leads to a dissimilarity matrix containing pairwise distances between all documents, which are now referred by several dissimilarity measures that could be partially dependent. The success for interpreting a user query relies on the effective use of all of these information sources as well as their interdependencies. In the following, we discuss the

different strategies to design a multimodal representation of data based on the dissimilarity spaces previously introduced.

We note d^{f_i} the distance measure applied to the feature space $\mathcal{F}_i$ and assume that dissimilarity matrices are known for M feature spaces. Then, given a set of positive examples $\mathcal{P}$, M monomodal spaces $\mathbf{d}^{f_i}$ are built.

A first way to fuse modalities is to consider that a possible multimodal dissimilarity would be the sum of all (normalized) monomodal distances. With this definition, the multimodal space $\mathbf{d}$ is simply

$$\mathbf{d} = \mathbf{d}^{f_1} + \mathbf{d}^{f_2} + \cdots + \mathbf{d}^{f_M} \in \mathbb{R}^p. \tag{1}$$

The dissimilarity space dimension is independent from the number of original feature spaces as it is always equal to p. This is a great advantage when M is large, but, on the other hand, one can object that the linear sum does not make sense to fuse features, especially when it deals with many sources of information. Moreover, the sum is sensitive to noisy and uninformative modalities, which will corrupt any classification operation.

To overcome these difficulties, we can consider that the fusion is carried out *a posteriori* by the classifier, which operates directly on multimodal components. The multimodal space is then the concatenation of all monomodal spaces, each element of which being represented by a multimodal dissimilarity vector

$$\mathbf{d} = \left[\mathbf{d}^{f_1 T}, \mathbf{d}^{f_2 T}, \ldots, \mathbf{d}^{f_M T}\right]^T \in \mathbb{R}^{p.M}. \tag{2}$$

This solution has the benefit to leave the fusion decision to the training process. However, it imposes to work in a higher-dimensional space where the estimation of the class distributions from a small training set may be less reliable.

Finally, we can adopt a more general fusion scheme where the input of the multimodal space is made of outputs of base classifiers. This solution is known as the *general combining classifier* [6, 12].

$$\mathbf{d} = \left[g_1(\mathbf{d}^{f_1}), g_2(\mathbf{d}^{f_2}), \ldots, g_M(\mathbf{d}^{f_M})\right]^T \in \mathbb{R}^M, \tag{3}$$

where $g_i(.)$ denote the output of base classifiers for the ith modality. The fusion algorithm is then splitted into two steps. First, individual classifiers are trained on their respective dissimilarity spaces. The classifier outputs are then used as input of a super classifier who takes the final fusion decision. This solution has the benefit to work in low-dimensional spaces but imposes $M+1$ classifications, leading to a computational over-head.

The choice between one of these three strategies will be discussed in the light of the results exposed in section 5. The problem now is to define which classification algorithm with which parameters we will be used to learn queries.

4 Classification

Many algorithms can be used to train a classifier that will learn semantic concepts. We have chosen to use an SVM because of its effectiveness and its flexibility in parameterization. The kernel selection and setting is a critical issue

to successfully learn queries. It actually decides upon the classical trade-off between over-fitting and generalization properties of the classifier and hence is very dependent of the considered representation space. Depending of the multimodal space used, we differentiate three setups for the classifier:

- sum of dissimilarity spaces (def. (1): An RBF kernel $k(\mathbf{x}, \mathbf{y}) = e^{-(\mathbf{x}-\mathbf{y})^T \mathbf{A}(\mathbf{x}-\mathbf{y})}$ with $\mathbf{A} = \sigma^{-1}\mathbf{Id}$ is used. The estimation of σ is based on a heuristic adapting the model to the query

$$\sigma = C \cdot \text{median}_i(\min_j ||\mathbf{d}_i^+ - \mathbf{d}_j^-||^2) \tag{4}$$

  The scale value is tuned to the median of all the minimum distances between the negative and the positive examples. In that way, the kernel becomes tighter as the two classes become closer to each other. The parameter C has been empirically set to 2.0.

- concatenation of spaces (def. (2): The same RBF kernel is used but $\mathbf{A} = \text{diag}[\boldsymbol{\sigma}_{f_1}, \cdots, \boldsymbol{\sigma}_{f_M}]$ so as to allow independent scaling for each modality. The scale vector $\boldsymbol{\sigma}_{f_i} \in \mathbb{R}^p$ is constant with all values equal to the scale parameter σ_{f_i} computed for each monomodal space $\mathbf{d}^{f_i}$ using the formula (4).

- hierarchical classification (def. (3): Independent classification of monomodal spaces are done with an RBF kernel and a scale σ computed with (4). As super classifier, a sigmoid kernel is used with a scaling parameter C set to 0.1.

The three above definitions become clearly equivalent when only one modality is considered. In the following, for the clarity of the results, we denote them respectively as (1) SUM, (2) CONC and (3) HIER.

5 Experimentations

The following experimentations have been conducted to compare the three proposed multimodal fusion algorithms and to measure the efficiency of the approach in a real video retrieval application. For both artificial and real annotated data, the experimentation consists in making queries corresponding to concepts and measuring the average precision, AP as the sum of the precision at each relevant hit in the retrieved list divided by the minimum between the number of relevant documents in the collection and the length of the list. The retrieved list has 100 entries and the measures are averaged over 50 queries. The annotated positive examples are removed from the hit-list so that they are not taken into account when measuring performances.

5.1 The video database

We use the complete annotated video corpus TRECVid-2003 composed of 133 hours of CNN and ABC news. Videos are segmented into shots and every shot has been annotated by several concepts. The speech transcripts extracted by Automatic Speech Recognition (ASR) at LIMSI laboratory [7] are also available.

We extracted the three following features from the 37'500 shots composing the corpus: Color histogram, Motion vector histogram and Word occurrence histogram (after stemming and stopping). The distance measures used are Euclidean for Color and Motion histogram and intersection for Word occurrence histogram.

5.2 Results

The two following experiments consist in characterizing and comparing the fusion strategies on artificial data. The features are generated from statistical distributions and Euclidean distance is used to compute pairwise dissimilarities.

We first evaluate the discriminative power of the three strategies. The features are drawn from two multivariate centered Gaussian distributions so as to generate three dissimilarity spaces simulating three modalities. The positive class has a variance fixed to 1 whereas the variance of the negative class varies from 1 to 3.5. That way, the negative class is gradually surrounding the positive one. We then measure how fast the fusion algorithms are able to isolate the positive elements from the rest of the data. The figure 2 plots the Average Precision for our three fusion strategies SUM, CONC and HIER.

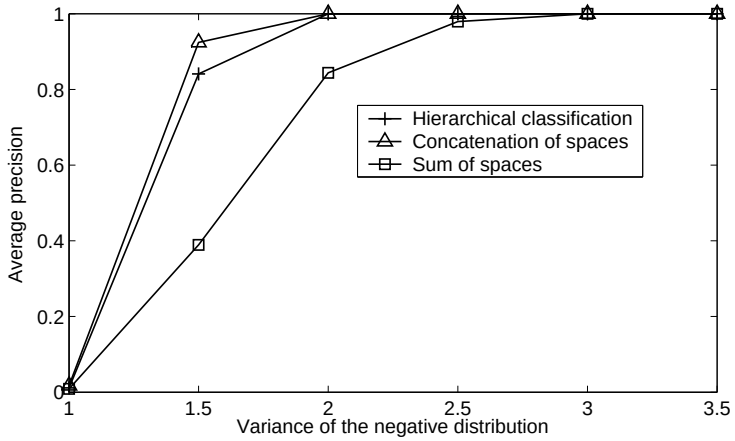

Fig. 2. Average precision vs. the variance of the Gaussian distribution corresponding to negative examples (the positive Gaussian distribution's variance is set to 1). The query is composed of 10 positive and 10 negative examples randomly chosen.

We observe that the CONC and HIER strategies show similar results and outperform the SUM scheme to discriminate between the two classes. It validates our prior idea that the SUM fusion scheme is clearly not optimal to combine dissimilarities.

We now evaluate the strategies in term of robustness to uninformative modalities. The problem is simulated by 10 modalities where one is purely informative (the dissimilarities are set to 0 for elements belonging to the sought concept and 1 for the rest) and the nine others are just uniform noise. Figure 3 shows

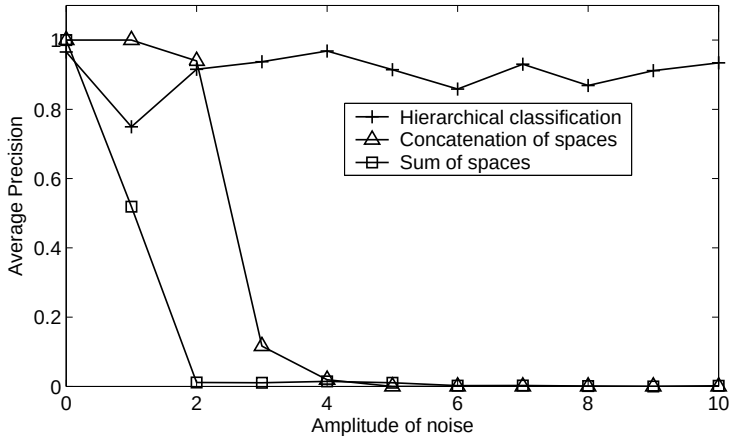

Fig. 3. Average precision on artificial data where one modality is purely informative while the 9 others are purely non-informative. The query is composed of 10 positive and 10 negative examples randomly chosen.

AP results when the amplitude of the noise rises from 0 to 10. Again we can observe that the SUM algorithm performs worse than the two others. But only the hierarchical classification is actually robust to noisy modalities.

The following results have been obtained on the video database. It is worth noting that the features we have extracted from videos are too weak to obtain a full video retrieval system with high performances. Our goal here is to demonstrate the potential capabilities of the approach for content-based indexing and retrieval on large multimedia databases.

We evaluate how the combination of modalities can improve the retrieval efficiency. The task is to retrieve shots that are annotated *"Weather news"*. This concept is consistent with the low-level features we use (coherence of colors, motion and speech) and enables us to evaluate the classification task into the dissimilarity spaces rather than the suitability of the features to characterize semantics. The graph 4 compares the AP when monomodal sources and multimodal information are considered. Whatever the strategy used, the retrieval precision is better when multimodal information is taken into account. We can observe that the speech has a great importance to characterize the concept, but

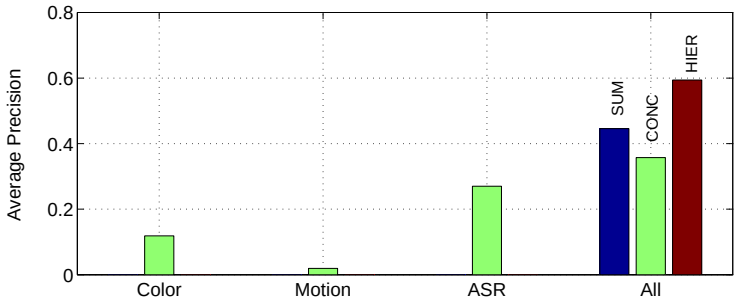

Fig. 4. Average precision for monomodal and multimodal retrieval and for the three multimodal fusion strategies. The queries are composed of 10 positive and 10 negative examples.

the addition of less relevant features such as motion and color still improves the performances.

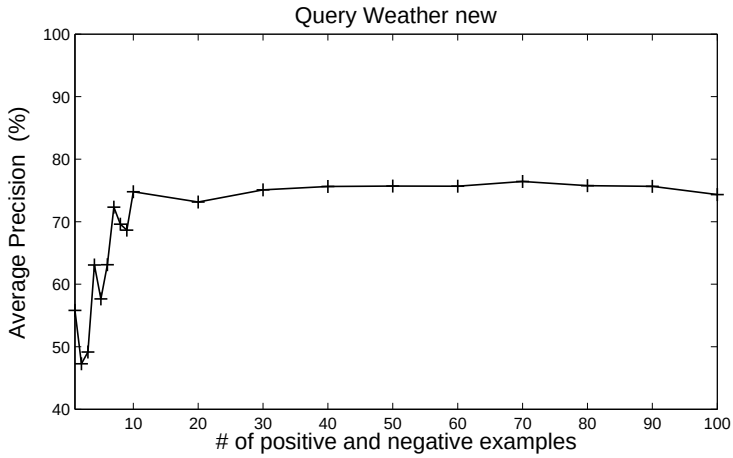

Fig. 5. Hierarchical classification results when positive examples and negative examples increase ($N_p = N_n$).

The following experiment tests how the retrieval precision evolves when the number of positive and negative documents grows (results have been obtained with the HIER strategy). As figure 5 shows, the AP increases with the size of the training set until a maximum value is reached from where the addition of new examples does not improve anymore the classification accuracy. This saturation behavior illustrates how the users, by providing more and more examples (in the

relevance feedback loop), can refine their queries until reaching the optimum of the classifier.

Next, we conducted experiments for several other annotated concepts (Table 1). These results are also compared to a random guess (e.g seeking hits at random within the database) to illustrates the capability of the algorithm to use low-level multimodal information to create models of semantic concepts defined by the user. This improves dramatically the performance of the search. However, with the reading of this table, it is rather difficult to form an opinion about the best fusion algorithm. While the HIER and CONC approach perform alternatively the best, the three algorithm performances are within the same range of precision. Surprisingly, the mean Average Precision (mAP) obtained for the SUM algorithm is higher than that calculated for the two other strategies. It is indeed disappointing that the less discriminative and robust approach gives globally the best results on real data. It is however worth noting that it also provides the simplest solution by minimizing both the dimensionality of the representation space and the number of classifications to operate. This simplicity is probably the reason of the good behavior of the SUM approach.

Table 1. AP for concept retrieval

Concept	SUM	CONC	HIER	Random
Studio Setting	0.56	0.57	**0.60**	0.01
Basketball	0.21	**0.22**	0.12	0.0024
Weather news	0.51	0.37	**0.55**	0.0014
Hockey	0.05	**0.06**	0.02	$1.85e^{-4}$
mean AP (mAP)	**0.33**	0.30	0.32	0.0035

Table 2. Computation load (in second)

	Training set		Algorithms		
	Positive ex.	Negative ex.	SUM	CONC	HIER
3 modalities	10	10	0.3	0.4	1.2
	20	10	0.5	0.8	1.9
	40	10	1.4	1.7	5.9
	10	40	1.2	1.3	4.2
10 modalities	10	10	0.3	0.6	3.4

Finally, we give the computation load of each algorithm (Table 2) for various training sets and for 3 and 10 modalities. We see that the SUM strategy is effectively the fastest while the computation load of the hierarchical classification increases linearly with the number of modalities used.

6 Discussion

Our three dissimilarity-based multimodal fusion strategies are able to take benefit from low-level audio-visual descriptions of video documents and, as a consequence, to learn semantic queries from a limited number of input examples. Moreover, the fusion of the information sources performs better than considering modalities independently.

The design of the dissimilarity space as been achieved so as to simplify the classification problem while building a low-dimensional representation of the data. As a result, queries on large databases are processed in near real-time which authorizes the use of feedback loop as a search paradigm.

In the light of the results, it is however difficult to decide which approach is optimal for the retrieval task. Tests on artificial data have demonstrated the superiority of the hierarchical classification in terms of class discrimination and robustness to corrupted modalities. The computation over-head is however not negligible, especially when a large number of features is used. On the other hand, the experiments carried out on real data do not exhibit clearly any superiority between the three fusion strategies. Actually, the weakness the low-level features do not allow us to definitely conclude this study.

Our future work will therefore consist in extracting new features that better characterize audiovisual content. It will enable us to properly evaluate the fusion algorithms and to determine the limits of the fusion schemes when a large number of features are used.

References

1. Liudmila Boldareva and Djoerd Hiemstra. Interactive content-based retrieval using pre-computed object-object similarities. In *Conference on Image and Video Retrieval, CIVR'04*, pages 308–316, Dublin, Ireland, 2004.
2. Eric Bruno, Nicolas Moenne-Loccoz, and Stephane Marchand-Maillet. Unsupervised event discrimination based on nonlinear temporal modelling of activity. *Pattern Analysis and Application, special issue on Video Event Mining*, 2005. (to appear).
3. E. Y. Chang, B. Li, G. Wu, and K. Go. Statistical learning for effective visual information retrieval. In *Proceedings of the IEEE International Conference on Image Processing*, 2003.
4. E. Chávez, G. Navarro, R. Baeza-Yates, and J.L. Marroquin. Searching in metric spaces. *ACM Computing Surveys*, 33(3):273–321, September 2001.
5. T.F. Cox and M.A.A. Cox. *Multidimensional scaling.* Chapman & Hall, London, 1995.
6. R.P.W. Duin. The combining classifier: To train or not to train? In *Proceedings of the 16th International Conference on Pattern Recognition, ICPR'02*, volume II, pages 765–770, Quebec City, 2004. IEEE Computer Socity Press.
7. J.L. Gauvain, L. Lamel, and G. Adda. The limsi broadcast news transcription system. *Speech Communication*, 37(1-2):89–108, 2002.
8. D Heesch and S Rger. Nnk networks for content-based image retrieval. In *26th European Conference on Information Retrieval*, Sunderland, UK, 2004.

9. E. Pekalska, P. Paclík, and R.P.W. Duin. A generalized kernel approach to dissimilarity-based classification. *Journal of Machine Learning Research*, 2:175–211, December 2001.
10. Philip Resnik. Using information content to evaluate semantic similarity in a taxonomy. In *14th International Joint Conference on Artificial Intelligence, IJCAI*, pages 448–453, Montreal, Canada, 1995.
11. J. R. Smith, A. Jaimes, C.-Y. Lin, M. Naphade, A. Natsev, and B. Tseng. Interactive search fusion methods for video database retrieval. In *IEEE International Conference on Image Processing (ICIP)*, 2003.
12. Y. Wu, E. Y. Chang, K.C-C Chang, and J.R Smith. Optimal multimodal fusion for multimedia data analysis. In *Proceedings of ACM Int, Conf. on Multimedia*, New York, 2004.
13. R. Yan, A. Hauptmann, and R. Jin. Negative pseudo-relevance feedback in content-based video retrieval. In *Proceedings of ACM Multimedia (MM2003)*, Berkeley, USA, 2003.
14. X.S. Zhou, A. Garg, and T.S. Huang. A discussion of nonlinear variants of biased discriminant for interactive image retrieval. In *Proc. of the 3rd Conference on Image and Video Retrieval, CIVR'04*, pages 353–364, 2004.
15. X.S. Zhou and T.S. Huang. Small sample learning during multimedia retrieval using biasmap. In *Proceedings of the IEEE Conference on Pattern Recognition and Computer Vision, CVPR'01*, volume I, pages 11–17, Hawaii, 2004.