Design of multimodal dissimilarity spaces for retrieval of video documents

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

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

Abstract

This paper proposes a novel representation space for multimodal information, enabling fast and efficient retrieval of video data. We suggest describing the documents not directly by selected multimodal features (audio, visual or text), but rather by considering cross-document similarities relatively to their multimodal characteristics. This idea leads us to propose a particular form of *dissimilarity space* that is adapted to the asymmetric classification problem, and in turn to the *query-by-example* and *relevance feedback* paradigm, widely used in information retrieval. Based on the proposed dissimilarity space, we then define various strategies to fuse modalities through a kernel-based learning approach. The problem of automatic kernel setting to adapt the learning process to the queries is also discussed. The properties of our strategies are studied and validated on artificial data. In a second phase, a large annotated video corpus, (*ie* TRECVID-05), indexed by visual, audio and text features is considered to evaluate the overall performance of the dissimilarity space and fusion strategies. The obtained results confirm the validity of the proposed approach for the representation and retrieval of multimodal information in a real-time framework.

Index Terms

H.2.4.e Multimedia databases, H.5.1 Multimedia Information Systems, H.5.1.f Image/video retrieval, I.2.6.g Machine learning, I.2.6.b Concept learning

I. INTRODUCTION

Determining semantic concepts by allowing users to iteratively and interactively refine their queries is a key issue in multimedia content-based retrieval. The Relevance Feedback loop allows to build complex queries made out of documents marked as positive and negative examples. From this training set, a learning process has to create a model of the sought concept from a set of data features to finally provide relevant documents to the user. The success of this search strategy relies mainly on the representation spaces within which the data is embedded as well as on the learning algorithm operating in those spaces. These two issues are also intrinsically related to the problem of adequately fusing information arising from different sources. Various aspects of these problems have been studied with success for the last few years. This includes works on machine learning strategies such as active learning [6], imbalance classification algorithms [41], automatic kernel setting [40] or automatic labelling of training data [37]. Theoretical and experimental investigations have been achieved to determine optimal strategies for multimodal

fusion: Kittler *et al* and R. Duin studied different rules for classifier combination [18], [11]; Wu *et al* propose the super-kernel fusion to determine optimal combination of features for video retrieval [35]. In [16], Maximum Entropy, Boosting and SVM algorithms are compared to fuse audio-visual features. A number of further relevant references may be found into the Lecture Notes series on Multiple Classifier Systems [25].

All these studies have in common the fact to consider feature spaces to represent knowledge on the multimedia content. This representation requires to deal in parallel with many high-dimensional spaces expressing the multimodal characteristics of the documents. This mass of data makes retrieval operations computationally expensive when dealing directly with features. For instance, the simple task of computing the distance between a query element and all other elements becomes infeasible in a reasonable time when involving hundred of thousands of documents and thousands of feature space components. This problem is even more sensible when similarity measures are complex functions or procedures such as prediction functions for temporal distances [4] or graph exploration for semantic similarities [28]. The diversity of the features involved is also a difficulty when dealing with fusion and learning. Indeed, the multimedia descriptors may be extracted from visual, audio or transcript streams using various operators providing outputs such as histograms, filter responses, statistical measures or symbolic labels. This heterogeneity imposes building complex learning setup that need to take into account all the variety of the features' mathematical and semantic properties [29][38].

An alternative solution is to represent documents according to their similarities (related to one or several features) to the other documents rather than to a feature vector. Considering a collection of documents, the similarity-based representation, stored in (dis)similarity matrices or some distance-based indexing structures [7], characterizes the content of an element of the collection relatively to a part of or the whole collection. Recent studies have been published for document retrieval and collection browsing by using pre-computed similarities. In [2], Boldareva *et al* proposed to index elements relatively to their closest neighbors, *i.e.* those who have the best probabilities to belong to the same class. This provides them with a sparse association graph structuring the multimedia collection and allowing fast retrieval of data. In [15], the idea of nearest neighbor networks is extended by creating edges for every combination of features. The resulting graph, called NN^k, allows to browse the data collection from various viewpoints corresponding to the multiple features. As pointed out by authors, the similarity

approach provides a convenient way for multimodal data fusion, since adding new features simply consists in adding new distances to the same representation framework. It is also noted that the off-line computation of similarities enables fast accesses and scalable content-based multimedia retrieval systems.

Keeping the advantages offered by similarity-based representations, we wish to go further and propose solutions that go beyond the classical nearest neighbor approaches. As user's judgements are supposed to be taken into account, it is indeed crucial to let this feedback modify inter-document similarities accordingly. Our goal is then to introduce adequate non-linear learning techniques (such as SVM or boosting) in order to benefit from their good ability to adapt low-level representation spaces to semantic concepts ([30], [31]).

In [27], Pekalska *et al* introduced the idea of dissimilarity spaces where objects are no longer represented by their features but by their relative dissimilarities with respect to a set of selected objects. This set actually defines a new basis where dissimilarities are considered as features. As a consequence, standard machine learning techniques, originally designed for features, are also available in dissimilarity spaces. The technique has been already used for object recognition [26] or image retrieval [22]. In [3], we proposed to use dissimilarity spaces to fuse multimodal information and to retrieve video documents. This communication constitutes a preliminary study of the work reported here. Precisely, we investigate here more deeply various aspects of the design of a dissimilarity-based multimedia retrieval system. These aspects may be divided into three main contributions:

1) definition of low-dimensional dissimilarity spaces. We aim at deriving from the general definition of dissimilarity spaces a specific construction adapted to the retrieval problem. In particular, we concentrate on the *Small Sample Learning* and *Asymmetric Learning* problems while maintaining the low dimensionality of the representation to allow fast retrieval.
2) learning in dissimilarity spaces. We consider the user queries being formulated in terms of positive and negative examples that then form a training set. A machine learning, a kernel-SVM in our case, is used to learn the sought concepts from this training set. A major difficulty arises from the fact that the proposed dissimilarity space is query-dependent, which implies to adapt SVM parameters online (particularly the kernel parameters). We propose an empirical scale measure automatically tuning the kernel parameters to the user

query content.

3) multimodal fusion strategies. Dealing with multimedia objects leads to building a set of dissimilarity spaces related to the multimodal signals composing them (*eg* visual, audio and textual signals). The optimal fusion of these spaces is an open issue. We develop several strategies that combine both dissimilarities and SVM outputs to effectively determine which of the fusion schemes provides an effective multimodal search engine.

All these points are developed in detail organized by the following outline: Section II exposes the general idea of the dissimilarity spaces and our solution to adapt it to the problem of content-based retrieval of multimodal data. In section III we address more specifically the machine learning issue. After having observed how linear and non-linear learning behave both in dissimilarity and feature spaces, kernel-based SVM is proposed to solve the retrieval task. Consequently, an automatic kernel parameter setting is derived to enable the SVM to adapt on-line to the user feedback. In section IV, various strategies for multimodal fusion are detailed. Section V proposes an evaluation of the strategies on both artificial data and real videos, *eg* the TRECVID 2005 corpus, and assesses the usability of the algorithms for video retrieval. Finally, section VI opens perspectives on the newly proposed techniques.

II. Dissimilarity representation for interactive retrieval

In a *query by example* retrieval system, users formulate complex queries by iteratively providing positive and negative examples in a Relevance Feedback (RF) loop. From this training data, the aim is to perform, at each step of the RF loop, a real-time dissimilarity-based classification that will select the most relevant documents from within the entire collection.

In this section, we recall the dissimilarity space initially introduced by Pekalska *et al* in [27] and show how it may be adapted to provide us with a low-dimensional approximation of the original feature space where an efficient classification may be performed.

A. Dissimilarity space

Let $d(\mathbf{x}_i, \mathbf{x}_j)$ be the distance between elements i and j according to their descriptors $\mathbf{x} \in \mathcal{F}$. The space $\mathcal{F}$ expresses the original feature space. The dissimilarity space $\mathcal{D}_\Omega$ is defined relatively to a set of elements $\Omega = \{\mathbf{x}_1, \ldots, \mathbf{x}_N\}$ by the mapping $\mathbf{d}(\mathbf{x}, \Omega) : \mathcal{F} \rightarrow \mathbb{R}^N$

$$\mathbf{d}(\mathbf{x}, \Omega) = [d(\mathbf{x}, \mathbf{x}_1), d(\mathbf{x}, \mathbf{x}_2), \ldots \, d(\mathbf{x}, \mathbf{x}_N)]. \quad (1)$$

The representation set Ω is a subset of N objects against which any element of the entire collection will be evaluated. The new "features" of an input element are now the values of dissimilarity with the representation objects Ω. 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 equal to N (the cardinality of Ω) and controls the quality of the approximation made on the original feature space. For instance, assuming an Euclidean dissimilarity measure, it is possible to recover up to a rotation the original m-dimensional space $\mathcal{F}$ from $\mathcal{D}_\Omega$ whenever $N \geq m$ [9]. On the contrary, the representation is incomplete when $N < m$ in the sense that only an approximation of $\mathcal{F}$ can be recovered. In our problem however, we are interested in detecting relevant documents and not in computing exactly $\mathcal{F}$ from $\mathcal{D}_\Omega$. A well-chosen space of low dimension would be probably more effective for learning processes as it avoids the *curse of dimensionality* problem and reduces the computation load. The selection of a "good" representation set may be driven by considerations on the particular learning problem we are dealing with, as shown in the next section.

B. A query-based representation set

As mentioned before, the RF strategy consists in gathering user's judgements indicating, for some documents, whether they are relevant or irrelevant to the user request. This set, denoted T, is called the *query* and is composed of positive and negative subsets, respectively

$$\mathcal{P} = \{\mathbf{x}_i^+, i = 1, \ldots, p\} \text{ and } \mathcal{N} = \{\mathbf{x}_i^-, i = 1, \ldots, n\}.$$

The query T is used to train a machine that will produce a ranking of documents relatively to their relevance to the query. This ranking is then presented to the user who, in turn, enriches the training set by adding new positive and negative examples chosen among the hit-list. These two steps are iterated until the search converges to a satisfactory result.

At first glance, the training stage at each step of the RF strategy seems to consist in solving a traditional learning problem. However, as mentioned by Zhou *et al* [41], specific difficulties arise from the RF protocol:

a) Asymmetric learning: The class configuration in feature space is generally *asymmetric*. This situation is known as the $(1 + x)$ class setup where the one class, presumably well-clustered in the feature space, corresponds to the sought documents (positive class), while an

unknown number x of classes, partially represented by negative examples, is supposed to model all irrelevant documents. Classical learning approaches, by applying a symmetric treatment to all classes are not really efficient for such a setup. As displayed in figure (1.a), learning the negative classes, while being feasible using traditional non-linear learning machines, becomes challenging when only few samples are available. Facing this particular situation then leads to developing dedicated algorithms, such as the *Biased Discriminant Analysis* (BDA) [41] or one-class SVM [8]. These approaches take into account that only the positive class is of interest, and involve discriminant criterion enforcing only the positive class compactness while just keeping negative samples away.

b) Small Sample Learning: The training set fedback by the user is generally small and incomplete. The given examples are more likely to be only partially representative of the class distributions. This especially concerns the negative classes which might be severely under-sampled. In that context, generalizing the learning over unlabeled areas of the space is quite hazardous and should encourage us to enforce *precision* rather than *recall*, *i.e.* to prevent from *false positive* rather than *miss-detections* so as to minimize the number of negative documents populating the top of the hit list.

We now show how these two problems may be addressed by using dissimilarity representation. The mapping of data in a dissimilarity space offers us the possibility to turn the asymmetric configuration into a more classical binary setup. By selecting the representation set Ω as the set of positive examples $\mathcal{P}$, we obtain a representation space $D_{\mathcal{P}}$ where elements are only considered from the point of view of their distances to the positive class representatives

$$\mathbf{d}(\mathbf{x}, \mathcal{P}) = [d(\mathbf{x}, \mathbf{x}_1^+), d(\mathbf{x}, \mathbf{x}_2^+), \dots \, d(\mathbf{x}, \mathbf{x}_p^+)]. \tag{2}$$

We can easily show that a built-in property of $\mathcal{D}_{\mathcal{P}}$ is to transform the asymmetric classification setup such that it becomes linearly separable. Let us consider the Fisher *class separability criterion* measuring how well $\mathcal{P}$ and $\mathcal{N}$ are *linearly separable* in the original feature space $\mathcal{F}$,

$$J_{\mathcal{F}} = \frac{\mathrm{tr} S_b}{\mathrm{tr} S_w} = \frac{||\bar{\mathbf{x}}^- - \bar{\mathbf{x}}^+||^2}{\mathrm{tr} S_w},$$

where covariances S_b and $S_w = S_w^+ + S_w^-$ are respectively *between-* and *within*-scatter matrices for the positive and negative classes, while $\bar{\mathbf{x}}^-,\ \bar{\mathbf{x}}^+$ are negative and positive class centroids.

For a given covariance S_w, the criterion $J_{\mathcal{F}}$ trivially says that the data are linearly separable whenever the two classes do not overlap ($J_{\mathcal{F}} > 1$). Alternatively, when the two centroids tend

to coincide ($J_{\mathcal{F}} \rightarrow 0$), we either face an asymmetrical setup (in that case $\text{tr}S_w^+ \ll \text{tr}S_w^-$), or a non-separable problem ($\text{tr}S_w^+ \approx \text{tr}S_w^-$) where any classification based on second order statistics will fail. In the following, the asymmetric setup ($\text{tr}S_w^+ \ll \text{tr}S_w^-$) is always assumed.

Let us now consider the same criterion in the dissimilarity space $\mathcal{D}_{\mathcal{P}}$, where class centroids are denoted $\bar{\mathbf{d}}^+$ and $\bar{\mathbf{d}}^-$, and *within* scatter matrices $\Sigma_w = \Sigma_w^+ + \Sigma_w^-$

$$J_{\mathcal{D}} = \frac{||\bar{\mathbf{d}}^- - \bar{\mathbf{d}}^+||^2}{\text{tr}\Sigma_w}.$$

Considering the triangle inequality, $J_{\mathcal{D}}$ is lower bounded by

$$J_{\mathcal{D}} \geq \frac{\left[||\bar{\mathbf{d}}^-|| - ||\bar{\mathbf{d}}^+||\right]^2}{\text{tr}\Sigma_w}. \tag{3}$$

The separability criterion $J_{\mathcal{D}}$ increases as $||\bar{\mathbf{d}}^+||$ decreases and/or $||\bar{\mathbf{d}}^-||$ increases. Then, noting that in feature space $\mathcal{F}$

$$||\bar{\mathbf{d}}^+|| = \frac{1}{p}\left[\sum_{i\in\mathcal{P}}\left(\sum_{j\in\mathcal{P}} d(\mathbf{x}_i^+, \mathbf{x}_j^+)\right)^2\right]^{1/2},$$

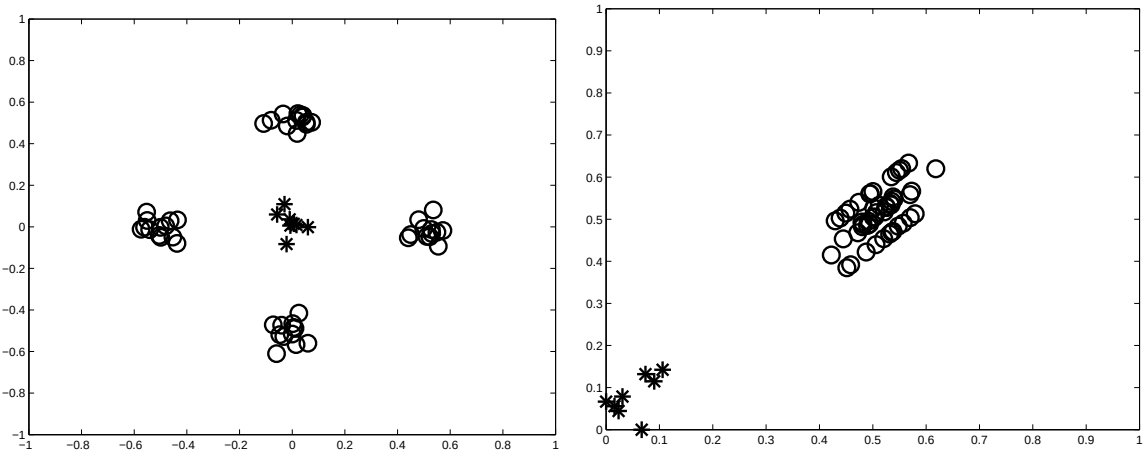
measures the compactness of $\mathcal{P}$, while

$$||\bar{\mathbf{d}}^-|| = \frac{1}{n}\left[\sum_{i\in\mathcal{N}}\left(\sum_{j\in\mathcal{P}} d(\mathbf{x}_i^-, \mathbf{x}_j^+)\right)^2\right]^{1/2},$$

measures the spread of $\mathcal{N}$ around $\mathcal{P}$, an asymmetric configuration ($||\bar{\mathbf{d}}^-|| \gg ||\bar{\mathbf{d}}^+||$) is linearly separable in the corresponding dissimilarity space $\mathcal{D}_{\mathcal{P}}$ as much as $\mathcal{P}$ has a compact distribution and/or the negative samples are spreaded around positive samples in the original feature space (Figure 1.b).

Another interesting feature of the space $\mathcal{D}_{\mathcal{P}}$ is its dimensionality. As the positive examples are provided by the user, their number is inherently limited to few dozens of documents. Consequently, it readily induces to work in a low dimensional space of $p = |\mathcal{P}|$ components, more suited for Small Sample Learning.

Finally, from an implementation point of view, it worth noting that $\mathcal{D}_{\mathcal{P}}$ was to be rebuild for every new user query. One may think this on-line indexing is prohibitive for real-time retrieval. However, as long as the complete dissimilarity matrix is known, the re-indexing consists only in reading all dissimilarity entries associated with elements from $\mathcal{P}$. This operation is linear in the size of the database whenever those dissimilarities are contiguous in memory. The storage space



(a) (b)

Fig. 1. The $(1 + x)$ class problem in feature space (a) and 2D dissimilarity space (b) where the representation objects are two points from the central class (stars). The asymmetric setup in $\mathcal{F}$ becomes linearly separable in $\mathcal{D}_{\mathcal{P}}$. Euclidean distance is considered as a dissimilarity measure.

required for the dissimilarity matrix is however effectively quadratic. While we clearly look at reducing this cost, empirical calculations show that it still permits indexing for approximatively 500'000 entries using current standard HD. Beyond this size, distance-approximating embedding such as FastMap [12], MetricMap [34] or BoostMap [1] may be considered as a potential solution to dramatically reduce the size of the dissimilarity indexes. Metric trees, such as M-trees [39], might be also of interest whenever the triangle inequality applies to the dissimilarity measures. How to couple these techniques with our dissimilarity spaces and the potential negative or positive effects on the retrieval efficiency has not been yet investigated and thus will not be reported in this article.

III. Classification in dissimilarity space

We have shown that linear machine learning is able to solve the $(1 + x)$ setup when data are projected in the dissimilarity space $\mathcal{D}_{\mathcal{P}}$. However, the $(1 + x)$ setup is an ideal case, while a more realistic configuration is rather that positive instances are distributed over several clusters surrounded by negative elements. We may note this configuration as $(c + x)$, where c is the unknown number of positive clusters (though strictly speaking there is only one positive class distributed over the c clusters). The following section displays various class configuration setups

(linear, $(1+x)$, $(c+x)$) and show how linear and non-linear SVM algorithms behave in feature and dissimilarity spaces.

A. Linear and non-linear SVM learning

For any input vector $\mathbf{z}$, the SVM decision function is the following weighted sum over the support vectors $\mathbf{z}_i$ [32]

$$f(\mathbf{z}) = \sum_{i=1}^{sv} \alpha_i k(\mathbf{z}, \mathbf{z}_i) + b. \tag{4}$$

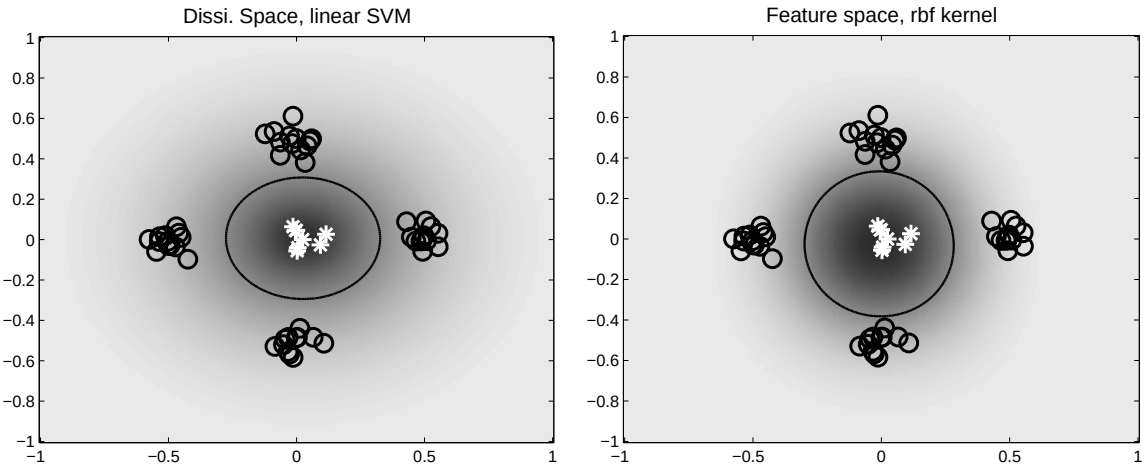
The kernel k defines the non-linearity of the classifier (simplified to a dot product for the linear case). For non-linear kernels, we restrict our study to the generic RBF kernel, $k(\mathbf{x}, \mathbf{y}) = e^{-\frac{|\mathbf{x}-\mathbf{y}|^2}{2\sigma^2}}$.

A mere classification task would involve to take the sign of $f(\mathbf{z})$, but because we are dealing with retrieval and are interested in obtaining a ranked list, the decision function is directly used to sort data according to their relevance to the positive class.

As shown by the Fisher criterion $J_{\mathcal{D}}$ (3), a linear classifier is sufficient to separate data when we assume a strict $(1+x)$ setup. The figure 2 displays a toy example depicting such a configuration. The decision function estimated with a linear SVM in the *dissimilarity space* (figure 2.a) is similar to that obtained with a RBF-SVM in *feature space*. Both approaches are able to separate the positive cluster from the negative examples. Let us now have a look to situations differing from the $(1+x)$ configuration.

Figure 3 presents results on a linearly separable problem. The decision functions obtained with a linear SVM in *dissimilarity space* and a RBF SVM in *feature space* show again similarities (figure 3.a and b). In both cases, the positive area is localized around the positive examples, restricting relevance to regions close to the positive examples (over-fitting behavior). Unsurprisingly, the optimal hyperplane generalizing the learning over the whole space is estimated through a simple linear SVM operated in *feature space* (figure 3.c). However, the decision function actually predicts the most relevant areas at infinity within the positive half space, which is not satisfactory from a retrieval point of view.

The $(c+x)$ configuration, more representative to real cases, is depicted in figure 4. In this setup, the positive examples are located in c distinct clusters. A linear discrimination in *dissimilarity space* assumes only one positive cluster, providing a decision function miss-interpreting the real distribution of the positive elements (Figure 4.a). It is worth noting that, contrary to the two



(a) (b)

Fig. 2. Toy example: Decision function for the cross configuration in 2D feature space. Circles are negative examples, stars are the positive samples. The gray level indicates relevance to the positive class (black = relevant, white = irrelevant). Black line indicates the separating hyperplane.

previous examples, this last result differs totally with the one obtained through a RBF SVM in *feature space* (Figure 4.b). In this case, the two positive classes are effectively distinguished. A similar result might be obtain in *dissimilarity space* by learning with a RBF SVM (Figure 4.c). In that case the *one positive class* assumption is alleviated by the non-linearity of the classifier.

In the light of these three examples, it appears that non-linear SVM is a reasonable choice to learn in *dissimilarity space* real-world distributions that do not conform to the $(1 + x)$ setup. In the following, we expose our motivations to use RBF kernel and propose solution for its automatic setting.

B. Automatic scale setting

A major consequence of the query-induced dissimilarity space is its variability: query after query, the data representation changes as the set $\mathcal{P}$ is augmented by the user. From a machine learning point of view, this requires to set algorithm parameters every time a new dissimilarity space is generated, that is to say, on the fly at each RF loop.

In the following, we define how to automatically adapt the RBF kernel scale parameter to the dissimilarity space modifications. This parameter has a direct geometric interpretation as it fixes the scale of integration of information within the feature space. This property allow us to

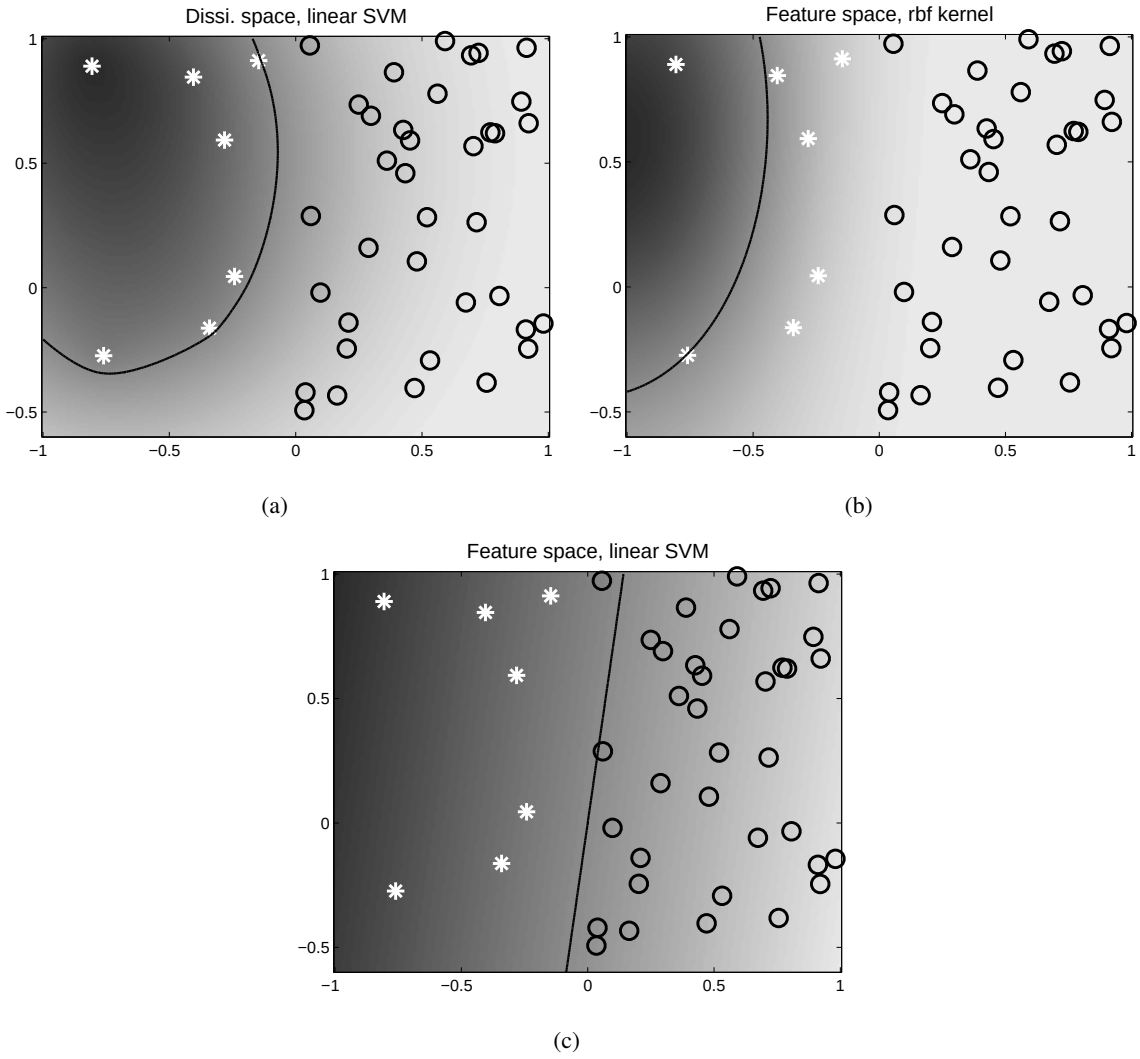


Fig. 3. Toy example: Decision function for the linear configuration in 2D feature space. Circles are negative examples, stars are the positive samples. The gray level indicates relevance to the positive class (black = relevant, white = irrelevant). Black line indicates the separating hyperplane.

determine a heuristic strategy for kernel setting from training data, as explained below.

The kernel selection and setting is a critical issue to successfully learn semantic models from 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. This difficulty has sparked growing interest in last years and several methods have been proposed to automatically select optimized kernel, such as Kernel Alignment [10], Hyperkernel [23] or Empirical Feature Space [36]. These methods rely on the optimization of criteria (called *quality*

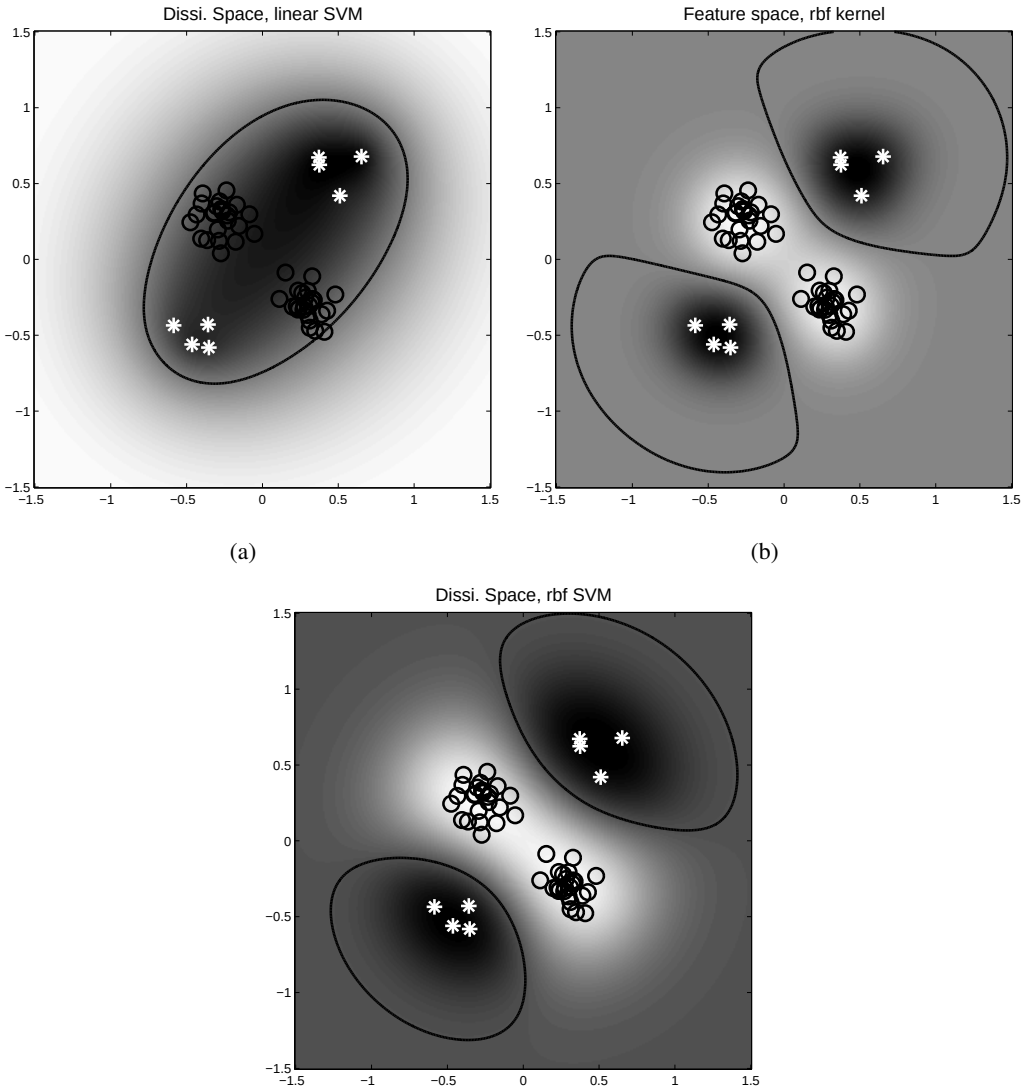


Fig. 4. Toy example: Decision function for the XOR configuration in 2D feature space. Circles are negative examples, stars are the positive samples. The gray level indicates relevance to the positive class (black = relevant, white = irrelevant). Black line indicates the separating hyperplane.

functionals in [23]) in kernel spaces according to the given training set. Therefore, they impose a non-negligible computational overhead for a result strongly dependent on the size and the quality of the training set. In the RF framework, these optimization-based approaches are thus prohibited because of the small number of training examples and the real-time constraint, but may inspire some empirical approximation of the *quality functional* objectives.

Directly derived from the Kernel Alignment is the Kernel Partial Alignment (KPA), specifically designed to cope with the class asymmetry problem [40]. It can be viewed as a measurement of the clusterization of the positive class in the kernel-induced space,

$$A^P \backsimeq \frac{1}{p^2} \sum_{i,j \in \mathcal{P}} k(\mathbf{x}_i, \mathbf{y}_j) - \frac{1}{np} \sum_{i \in \mathcal{P}, j \in \mathcal{N}} k(\mathbf{x}_i, \mathbf{y}_j). \tag{5}$$

The KPA criterion has two objectives: first, it enforces generalization within positive class by maximizing $\sum_{i,j \in \mathcal{P}} k(\mathbf{x}_i, \mathbf{y}_j)$. In that sense, it favors the *recall* of the retrieval. The second objective consists in separating clearly the positive from the negative elements by minimizing $\sum_{i \in \mathcal{P}, j \in \mathcal{N}} k(\mathbf{x}_i, \mathbf{y}_j)$, which, in turn, should enhance the *precision* of the search. However, as already said, optimizing directly (5) implies a computational effort not compatible with the RF protocol. In place, an empirical scale setting may be derived that follows the same goals as the KPA criterion.

Assuming the kernel k be a RBF function, the KPA criterion (5) tell us that a "good" scale magnitude is upper-bounded by a value proportional to the class margin. The bound ensures that kernels centered on positive samples would not overlap over negative elements. Thus setting the scale to this bound provides us with a kernel width as large as possible to cover positive examples without overlapping the negative class.

A possible approximation of this margin may be

$$\xi = \text{median}_i(\min_j ||\mathbf{d}_i^+ - \mathbf{d}_j^-||), \tag{6}$$

where $\{\mathbf{d}_i^+, i = 1, \ldots, p\}$ denotes the positive examples and $\{\mathbf{d}_i^-, i = 1, \ldots, n\}$ the negative examples in $\mathcal{D}_\mathcal{P}$. This measure considers the median of all distances separating each positive sample to its closest negative. The median operator is preferred to the mean as it introduces robustness against positive or negative outlier elements. A small value of ξ means that the two classes are close to each other or overlap . Therefore, the scale parameter has to be reduced so as to minimize the right hand term in eq. (5). On the other hand, a large value of ξ indicates a large separation between the two classes, meaning that the kernel has to be broaden to maximize the *within* term in (5).

From the class margin measurement ξ, an empirical scale may be derived

$$\sigma_{emp} = C \cdot \xi, \tag{7}$$

where the parameter C controls the trade-off between over-fitting and generalization, *i.e.* precision versus recall. As stated before, we would like to enforce precision. We fix $C = 1/2$ so as to impose that the majority of negative samples remains out of the bandwidth of kernels centered on positive examples and thus are pushed away in the rank list.

IV. 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 information sources as well as their inter-dependencies. 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_k} the distance measure applied to the feature space $\mathcal{F}_k$ and assume that dissimilarity matrices are known for M feature spaces. Then, given a set of positive examples $\mathcal{P}$, M monomodal spaces $\mathcal{D}_{\mathcal{P}}^{f_k}$ are built. The vector $\mathbf{d}^{f_k}$ denotes an element in the space $\mathcal{D}_{\mathcal{P}}^{f_k}$,

$$\mathbf{d}^{f_k}(\mathbf{x}, \mathcal{P}) = [d^{f_k}(\mathbf{x}, \mathbf{x}_1^+), d^{f_k}(\mathbf{x}, \mathbf{x}_2^+), \dots \, d^{f_k}(\mathbf{x}, \mathbf{x}_p^+)].$$

A. SUM strategy

A first way to fuse modalities would be to make the summation of all monomodal distances resulting in building a multimodal dissimilarity space. A SVM classification could then be directly applied in that space. However, the individual dissimilarity spaces need first of all to be properly scaled in order to sum homogeneous data. Following the discussion in section III, the multimodal dissimilarity vector $\mathbf{d}$ is defined as

$$\mathbf{d} = \sum_{k=1}^{M} \frac{\mathbf{d}^{f_k}}{\sqrt{2}\sigma_{f_k}} \in \mathbb{R}^p, \tag{8}$$

where σ_{f_k} is the empirical scale (see equation 7) computed in dissimilarity space $\mathbf{d}^{f_k}$. Consequently to rescale all monomodal spaces, the scale of the RBF kernel embedding $\mathbf{d}$ is set to 1.

It is worth noting that the dissimilarity space dimension is independent from the number of original feature spaces as it is always equal to $p = |\mathcal{P}|$. This is a definite advantage when M is

large, but, on the other hand, one can object that the linear sum does not make sense for fusing features, especially when it deals with many sources of information. Moreover, the sum operator is sensitive to noisy and uninformative modalities, which will corrupt any further classification operations.

B. CONC strategy

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

$$\mathbf{d} = \left[\mathbf{d}^{f_1 T}, \mathbf{d}^{f_2 T}, \ldots, \mathbf{d}^{f_M T}\right]^T \in \mathbb{R}^{pM}. \tag{9}$$

Again, a SVM with RBF kernel is used, but considering the heterogeneity of the multimodal components, the scalar scale parameter is replaced with a covariance matrix, *eg* $k(\mathbf{x}, \mathbf{y}) = e^{-(\mathbf{x}-\mathbf{y})^T \mathbf{A}^{-1}(\mathbf{x}-\mathbf{y})}$, with $\mathbf{A}$ the diagonal matrix

$$\mathbf{A} = \begin{pmatrix} \Sigma_1 & & 0 \\ & \Sigma_2 & \\ & \ddots & \\ 0 & & \Sigma_M \end{pmatrix} \in \mathbb{R}^{pM}, \text{ where } \Sigma_k = \begin{pmatrix} \sigma_{f_k} & & 0 \\ & \sigma_{f_k} & \\ & \ddots & \\ 0 & & \sigma_{f_k} \end{pmatrix} \in \mathbb{R}^{p},$$

so as to allow independent scaling for each p-dimensional dissimilarity space. The scale σ_{f_k} is estimated in $\mathcal{D}_{\mathcal{P}}^{f_k}$ with equation (7).

This solution has the advantage of leaving 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.

C. HIER strategy

Finally, we can adopt a more general fusion scheme where the input of the multimodal space is made of outputs of base classifiers. This hierarchical solution is known as the *general combining classifier* [11], [35].

$$\mathbf{d} = \left[g_1(\mathbf{d}^{f_1}), g_2(\mathbf{d}^{f_2}), \ldots, g_M(\mathbf{d}^{f_M})\right]^T \in \mathbb{R}^M, \tag{10}$$

where $g_k(\cdot)$ denotes the decision function of the base classifier for the kth modality. The fusion algorithm is then split into two steps. First, individual classifiers are trained on their respective dissimilarity spaces: a RBF SVM with automatic scale selection computed with (7) is used. The real-valued classifier outputs are then used as input of a super classifier which takes the final fusion decision. Its role consists in combining the monomodal classifiers so as to extend the learning function according to all the soft decisions taken within each modality. Here, a Gaussian kernel is used again. Setting dynamically the scale parameter remains an issue not addressed in this study. However, as the output of each monomodal classifier ranges from around -1 to 1, we consider in a first approximation that the training examples' distribution is not varying drastically from a query to an other. In that respect, the scale adaption is less crucial and might be set to a fixed value. We choose $\sigma = 1$ as we empirically observe that such a setting provides an average satisfactory performance, though results may not be optimal individually.

For the completeness of evaluation, we also propose a linear variation of the HIER strategy, called HIERLin, where the base classifiers are linear SVM. This restrict the classification problem to be $(1+x)$ in each monomodal space, but avoids the empirical scale setting. The super-classifier however remains a RBF-SVM as the fusion of modalities does clearly not conform to a linear setup.

The general combining classifier presents the advantage to work in low-dimensional spaces, where decisions are first taken independently without being polluted by possible uninformative features and then combined to aggregate multimodal information. However, it imposes $M + 1$ classifications, leading to a non-negligible computational overhead which may penalize the workflow during the on-line interactive retrieval.

V. EXPERIMENTATIONS

The following experiments have been carried out to evaluate all aspects of the proposed strategies for Relevance Feedback multimodal information retrieval. For that purpose, we have considered artificial feature and dissimilarity spaces in order to evaluate and assess the algorithms in a controlled setup. In a second stage, the algorithms were confronted to a real and difficult video retrieval benchmark, namely the TRECVID 05 benchmark (detailed in next section).

The experimentation section is composed of two parts: First, the monomodal query-based dissimilarity representation is compared to the classical feature representation in term of learning

ability using standard classifiers. The automatic kernel setting is evaluated as well. Then, the multimodal fusion strategies are evaluated, so as to determine for each of them their effectiveness, learning efficiency and robustness.

A. Video database and features

Evaluations on real data are based on the international benchmark for video information retrieval TRECVID. Year after year, the TRECVID workshops propose large corpora of video that are manually annotated so that shot and story segmentations and labelling are available. In the sequel, TRECVID-05 is considered [24]. This corpus is composed of 169 hours of multilingual TV broadcast news (English, Chinese, Arabic), roughly equally divided into a development set and a test set. The development set comes with annotations at the shot level, drawn from a list of semantic concepts (listed on table I). The two sets are used to perform various tests such as shot segmentation, low level and high level semantic feature extraction, or searching for particular topics (ad-hoc search).

In our setup, video documents are segmented into around 89'500 segments using the common shot reference. These shots are considered as individual and independent documents. This means that no contextual information is taken into account and that shot description is restricted to its audiovisual content (*eg* visual, audio and speech[1] information).

The Search Task, as defined in TRECVID-05, consists in retrieving shots that are relevant to some predefined queries (called topics). There are 24 topics concerning people (person-X queries), objects (specific or generic), locations, sports and combinations of the former. For each topic, keywords, pictures and several video shots (4-10) from the development set are provided as positive examples, while the groundtruth (obtained by pooling) concerns only documents from the test set. Further details about the Search Task may be found in [24].

During the experimentations, we only considered video shots of the development set as positive examples (keywords and additional pictures were not used). The positive examples are completed with ten negative examples randomly selected from within the test set. Starting with this initial query, a relevance feedback loop is initiated by adding to the query up to 10 new positive and negative examples returned in the 1000-entries hit-list. The process is repeated ten times.

[1]the English transcripts extracted by Automatic Speech Recognition (ASR) are provided by NIST.

Following the TRECVID evaluation protocol, the performance was measured at every iteration by Average Precision (AP) at 1000. The Average Precision is 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 (1000 in our case). The Mean Average Precision (MAP) is the AP averaged over several topics, and is used when appropriate. Additionally to the algorithm performance, a baseline consisting in retrieving randomly documents is always provided.

The multimodal feature and dissimilarity spaces are derived from the six following textual and audiovisual features:

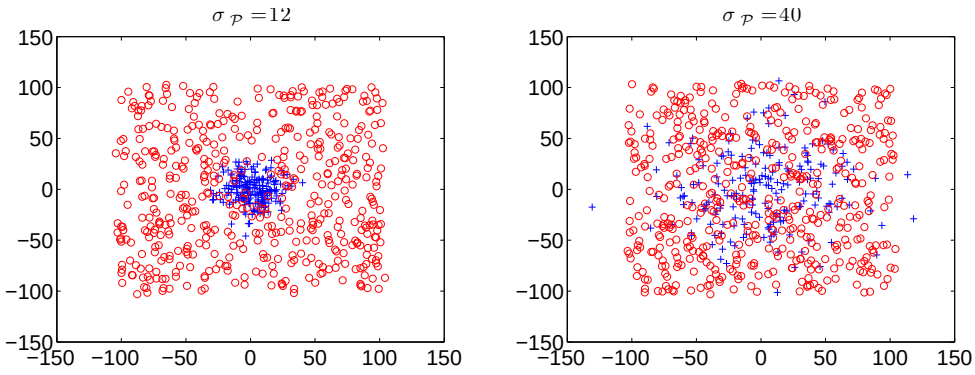
- Color histogram, $4 \times 4 \times 4$ bins in HSV space
- Motion vector histogram, 66 bins quantization of the MPEG block motion vectors [17]
- Local features, SIFT descriptors extracted around the Lowe salient points [19],
- Face detection [33],
- Word occurrence histogram (vector space model),
- Dominant audio features [13].

The distance measure used over color and motion histograms is the Euclidean distance. An approximation of the minimal matching distance is applied over local features to determine partial similarities [20]. Euclidean distance in the 30-dimensional eigenface space provides the similarity between the detected faces. Cosine distance is used for the vector space model and finally the audio similarity measure proposed in [13] is used for audio features.

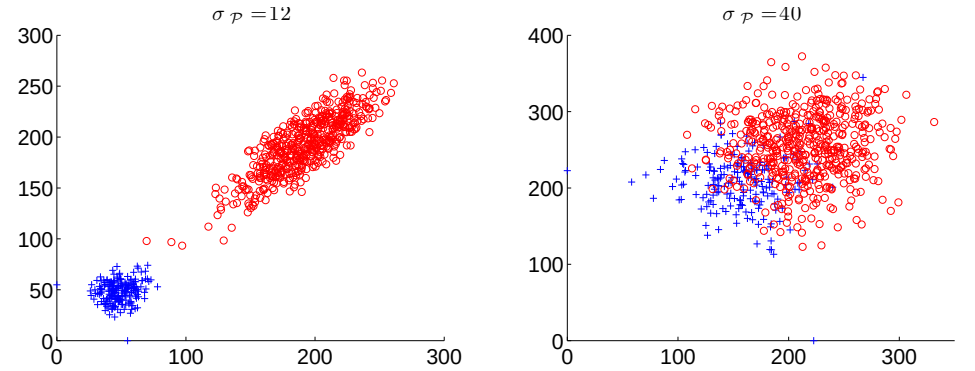
B. Monomodal dissimilarity space evaluation

This section concentrates on evaluating the dissimilarity space in itself. We have considered both artificial and real data to test various learning machines in dissimilarity spaces, to compare with feature space and to evaluate the automatic scale selection for RBF SVM.

1) Artificial data: Two baseline learning techniques (ie *k*-NN and linear SVM) are compared both in the original space and in the derived dissimilarity space. For that purpose, a 10-dimension feature space is generated with a positive class drawn from a centered Gaussian distribution $N(0, \sigma_{\mathcal{P}})$ and a negative class uniformly distributed around the positive class (see figure 5 for 2D view). The queries are made of the same number of positive elements x_i^+ and negative elements x_i^- randomly picked from the two classes. Figure 6 displays the average precision



(a) Feature space



(b) Dissimilarity space

Fig. 5. Artificial data: Blue crosses represent positive elements drawn from $\mathcal{N}(0, \sigma_{\mathcal{P}})$, while red circles represent the negative samples uniformly distributed around the center of the space.

obtained for every method in each space and for a growing number of examples. The curves support the discussion in section (III): in the original space, the problem cannot be solved using a linear classifier (linear SVM in feature space), and requires a large number of examples to be solved with k-NN. On the other hand, learning in dissimilarity space may be efficiently done using either linear SVM or k-NN and few training samples as the nature of the classification problem has changed to a simpler binary one.

Next, we consider a RBF SVM to learn the query in dissimilarity space. Our goal now is to evaluate the empirical scale setting strategy proposed in section III-B. Several feature and dissimilarity spaces are generated by varying the spread $\sigma_{\mathcal{P}}$ of the positive class to make the data more and more intricate. Figure 7 compares retrieval performances and scale values for various automatic scale setting strategies: Kernel partial alignment (σ_{KPA}), leave-one-out cross-validation

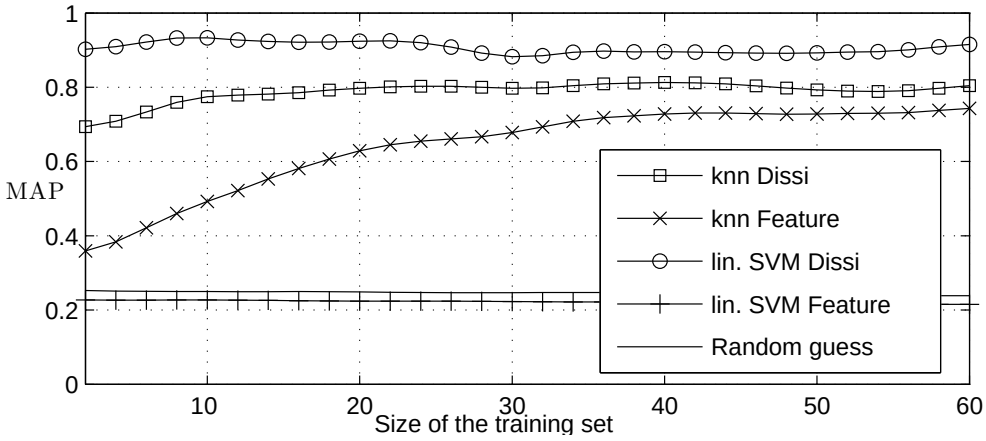
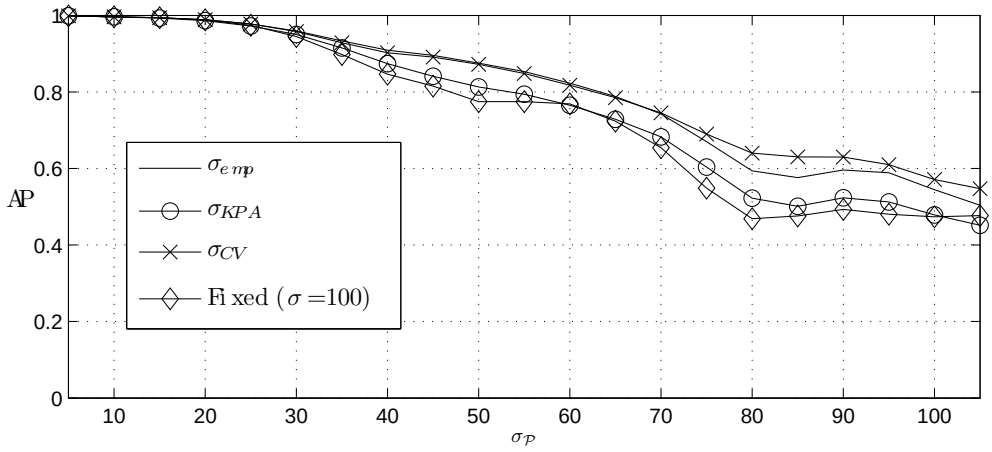


Fig. 6. Artificial data: Average precision for linear-SVM and *k*-NN in dissimilarity and feature space for a growing train set. A 10-dimension artificial space with $\sigma_{\mathcal{P}} = 80$ is used.

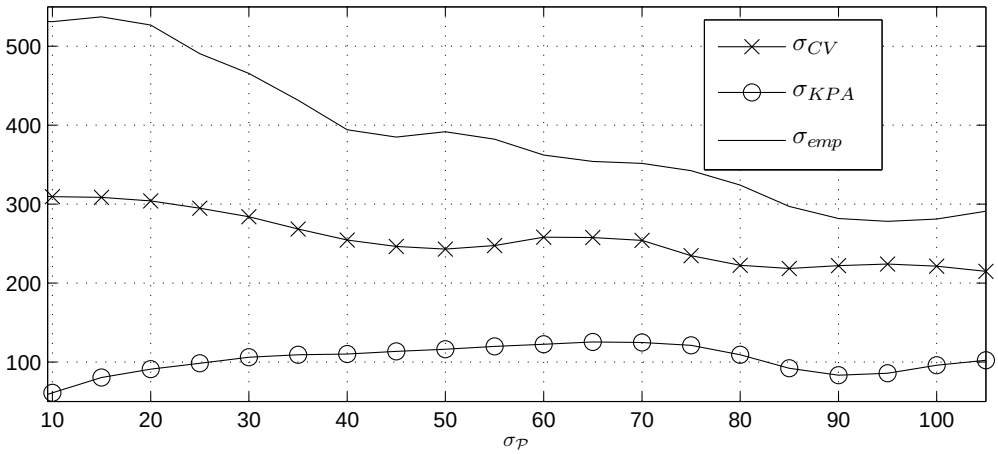
(σ_{CV}), empirical scale measure σ_{emp} (7) and fixed scale ($\sigma = 100$). The queries are composed of 10 positive and 10 negative examples. Unsurprisingly, the performance decreases as the classes overlap, and the time consuming leave-one-out crossvalidation yields better results in difficult situations (Figure 7.a). We observe however that empirical scale estimation and cross-validation provide similar performance, and outperform kernel alignment or fixed scale. Moreover, we notice in Figure 7.b that σ_{emp} and σ_{CV} have a similar decreasing trend as the two classes become less and less separable. In contrast, σ_{KPA} takes into account both the effective spread of the positive class (width increases as $\sigma_{\mathcal{P}}$ grows) and the positive/negative class separation (decreases as class separation becomes tight). This results in a less effective scale estimation of lower magnitude.

2) Real data: Finally, we propose to evaluate the approaches on the TRECVID-05 Search Task. Among the six video descriptors listed in section V-A, we consider the 64-bin color histogram to evaluate monomodal dissimilarity space. Though color information in itself is obviously inadequate to retrieve efficiently the high level topics of the benchmark, it is however sufficiently meaningful to retrieve fractions of the relevant documents and to figure out trends and behaviors of the proposed learning methods. Moreover, the use of color histogram allows to compare easily feature and dissimilarity space since a simple RBF kernel may be directly used in feature space.

The real data evaluation is conducted in two stages: First we assess RBF SVM learning in



(a) Retrieval performance



(b) Estimated scale magnitude

Fig. 7. Artificial data: Automatic scale setting of the RBF kernel using kernel partial alignment (σ_{KPA}), cross validation (σ_{CV}) and empirical estimation (σ_{emp})

dissimilarity space using automatic scale selection. Then we compare this method to a linear SVM in dissimilarity space and a RBF SVM in feature space.

Figure 8 shows MAP performance when the scale of the kernel is set to σ_{emp} (7) or to some fixed value. This result clearly indicates that adapting the scale to the queries (and to the derived dissimilarity space) permits to keep high precision results whatever the query. On the other hand, fixing σ to predefined values does not allow to reach such level of precision for all RF iterations: In some configurations ($\sigma = 10$ and $\sigma = 0.01$) and after several iterations,

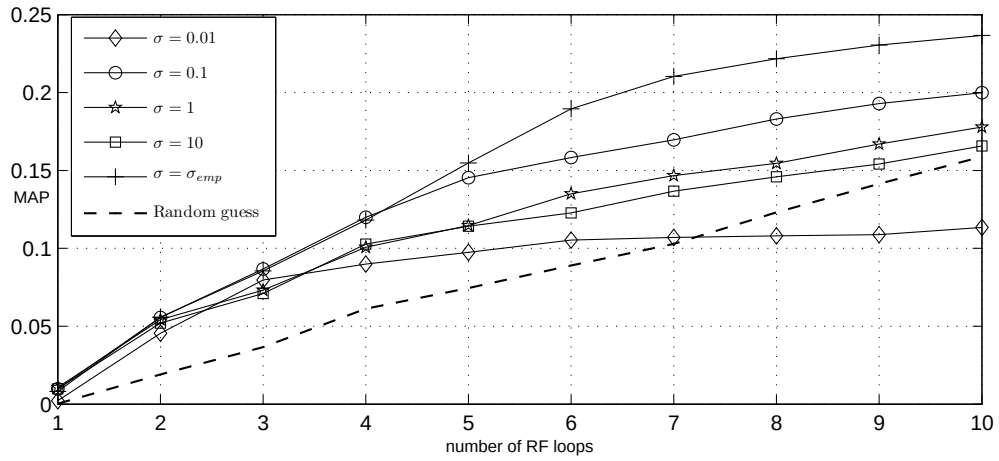


Fig. 8. TRECVID 05 Search Task, color dissimilarity space. Automatic scale setting: Comparison of adapted vs. fixed scale value.

the performance drops to a level equal or even below the random guess baseline, making the retrieval process completely inefficient.

In the next experiment (Figure 9), we compare the dissimilarity-based approach with linear and RBF kernel, to the feature-based method with RBF kernel. In both spaces, the RBF kernel parameter is set to σ_{emp}. To calculate σ_{emp} in feature space, $\mathbf{d}^+$ and $\mathbf{d}^-$ are replaced with the corresponding feature vectors $\mathbf{x}^+$ and $\mathbf{x}^-$ in equation (7).

The three approaches behave similarly for the first iterations (up to iteration 3), when few examples are available. At the first iteration (no feedback provided), we observe that linear SVM performs slightly better (MAP$= 0.012$) than RBF SVM (MAP$= 0.010$) and RBF SVM in feature space (MAP=0.008). These results confirm the fact that a simple linear learning in dissimilarity space is able to handle a complex asymmetric setup in feature space, and even outperforms feature-based non-linear SVM when scarce training samples are provided. As the training set grows, non-linear learning performance increases. We note that RBF SVM in dissimilarity space clearly outperforms both the linear learning and the feature-based approach. It is interesting also to note that linear learning in dissimilarity space tends to provide results comparable to the RBF SVM in feature space as the number of RF loops increases.

To sum up this evaluation section, all results obtained seem to confirm the advantage of using dissimilarity spaces to learn asymmetrical setup rather than feature space. Moreover, we

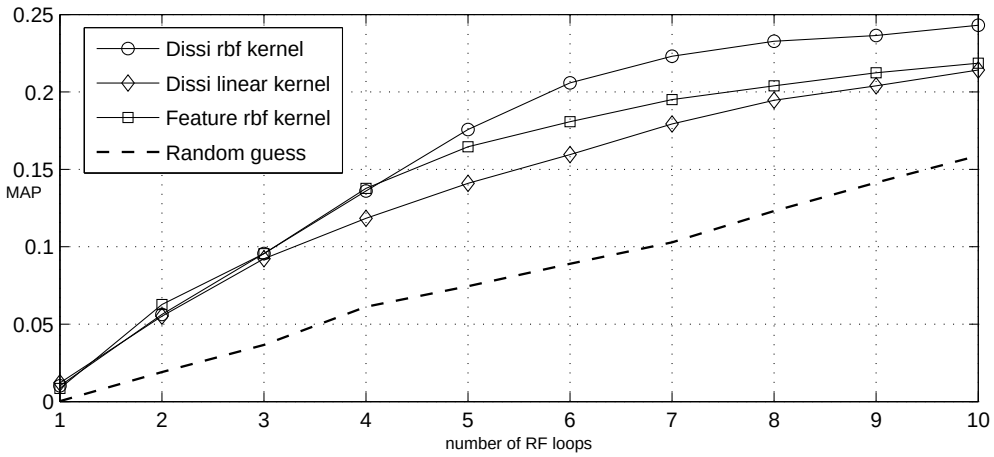


Fig. 9. TRECVID 05 Search Task, color dissimilarity space. Dissimilarity vs. feature space: Linear and RBF SVM in dissimilarity space compared to RBF SVM in feature space.

validated the fact that adapting the kernel to the dissimilarity space is crucial when learning with RBF SVM, and we validated the proposed empirical scale setting as an effective and tractable adaptive solution.

C. Multimodal fusion evaluation

The next step of the evaluation consists in studying in detail how the combination of modalities might improve the retrieval efficiency. In particular, we study how fusion algorithms behave when the number of modalities varies, when classes become less and less separable, when corrupted modalities are mixed to informative source channels and finally what their respective performance are when faced to a real video retrieval problem.

1) Artificial Data: To better understand the implemented fusion processes in various situations, tests are first performed on artificial data. Using the toy example described in figure 5, M monomodal feature spaces are produced with various width values of the positive class extent ($\sigma_{\mathcal{P}}$). Recalling that the magnitude of $\sigma_{\mathcal{P}}$ determines the separability of the two classes, we can modulate the amount of information carried by the different spaces.

For the following experiments, the training set is always composed of 5 positive and 5 negative examples. It is worth noting that the three above definitions, denoted respectively as (8) SUM, (9) CONC and (10) HIER, become clearly equivalent when only one modality is considered.

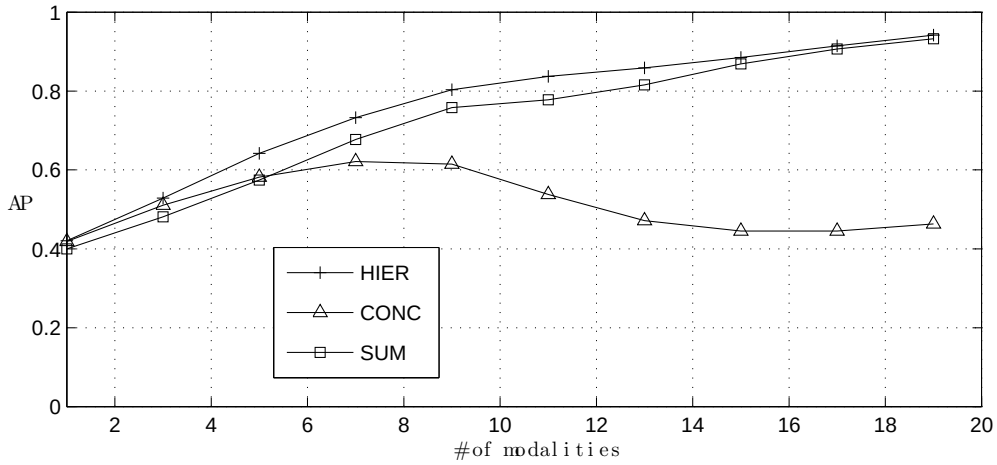


Fig. 10. Artificial data: Average precision of multimodal fusion when the number of modalities is growing.

We first compare the efficiency of the algorithms to fuse more and more channels. Each monomodal space is generated with different class configurations. This is achieved by randomly setting $\sigma_{\mathcal{P}}$ from a gaussian distribution $N(120, 10)$ so that spaces are more and less informative relatively to the positive class. Figure 10 gives the AP performance. The HIER and SUM strategies are effectively able to gain from the addition of information sources, whereas the learning within the CONC space suffers from the increase of the dimensionality induced by the concatenation of new modalities.

The second test consists in judging the ability to discriminate positive elements when classes overlap. For that purpose, the width $\sigma_{\mathcal{P}}$ is drawn from sliding probability distribution $N(\bar{\sigma}_{\mathcal{P}}, 10)$, with the average width $\bar{\sigma}_{\mathcal{P}}$ going from 90 to 160. The number M of modalities involved is set to 10. The HIER strategy exhibits the best discriminative behavior when the data becomes hard to separate. These results are followed by the SUM approach which finally gives an acceptable performance compared to the CONC space. The reasons of the CONC's under-performance could be again attributed to the problem of the curse of dimensionality, which is even more sensitive for the SVM classification in case of low signal-to-noise ratio [14].

The last test concerns the robustness of the fusion algorithms to corrupted modalities. This problem is simulated by replacing more and more modalities by totally uninformative feature spaces (where the data samples are drawn from the same uniform distribution). For the remaining non-corrupted spaces, the classes are set to be separable (small $\sigma_{\mathcal{P}}$). As expected, only the HIER

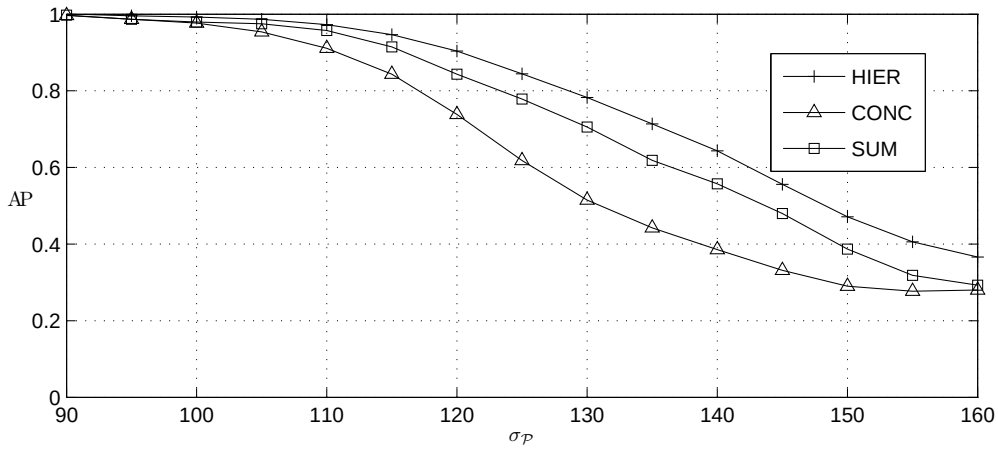


Fig. 11. Artificial data: Discriminating power with data less and less separable.

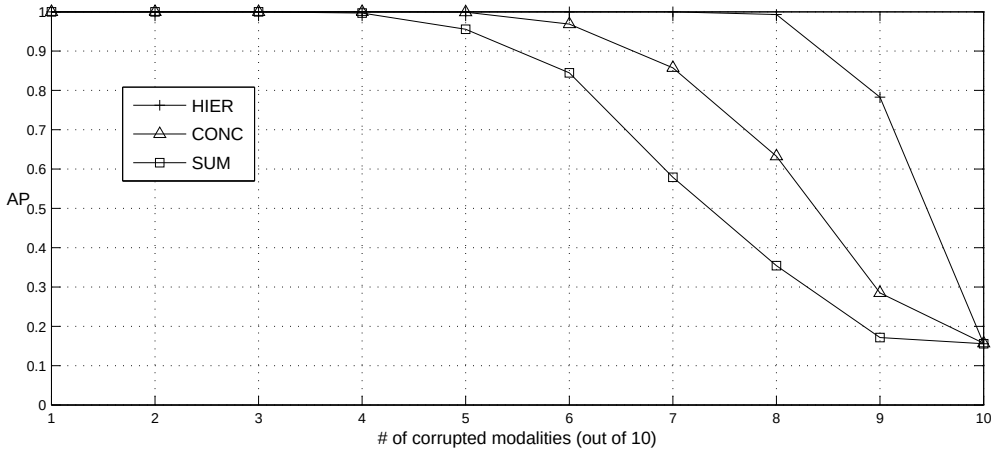


Fig. 12. Artificial data: Effect of corrupted modalities on the retrieval performance.

algorithm is robust to corrupted channels, whereas the performance of the others drop quickly as the number of uninformative spaces become the majority of the available sources (Figure 12). As pointed out in section IV-A, this result confirms that the SUM space is particularly sensitive to unreliable information.

2) Real data experiments: We now report experiments conducted on the TRECVID videos. For the completeness of the report, both high level (semantic) feature extraction[1] and topic

[1]though the high-level feature extraction task as defined in TRECVID guideline is strictly speaking considered as a classification task and not a retrieval one.

search tasks are considered. Performance comparison with approaches proposed by TRECVID-05 participants may be found in [24].

We consider to retrieve the 39 concepts (defined in [21]) used to annotate video shots of development set (see the list on table I). The queries are composed of twenty positive and twenty negative examples. For a given concept, the AP is averaged over 10 randomly selected query instances. The MAP is then computed by taking the mean AP over all concepts. The fusion strategies are compared together and against monomodal searches. Results are presented in table (I).

Comparing first the monomodal search performance, we observe that the visual features (color, faces and salient features) are more informative than the others. This confirms that the TRECVID concepts are mainly correlated to visual content. The face feature performs surprisingly well but it is essentially due to some concepts strongly correlated to the presence of people in the scene (”Face”, ”Person”, and ”Studio”) . On contrary, the color information is in general more reliable. Finally, we can note that none of the features goes below the random guess baseline which in this case may indicate a counterproductive effect on the overall system.

The MAP estimated over the 39 concepts indicates clearly that the HIER strategy is the best way to fuse modalities. Its performance is globally far better than HIERLin, CONC and SUM. Moreover, its behavior is quite stable since it outperforms others fusion techniques for 28 concepts, and best monomodal search for 31 concepts (out of 39). We observe also that CONC and SUM fusion strategies have globally equivalent performance and outperforms monomodal searches. Careful inspection of these results reveal that the fusion processes are in general not reliable since both CONC and SUM underperform the best monomodal searches for respectively 27 and 21 concepts (out of 39). As for HIERLin (hierarchical fusion with linear monomodal classifiers), the performance largely remains low compared to other strategies and moreover it never benefits from multimodality, with a notable exception for the ”Face” concept. This result indicates that in a realistic configuration, the $(1 + x)$ assumption underlain by the linear learning does not hold in general. When the assumption holds however (as for the ”Face” class characterized with a low visual intra-class variance), the linear implementation performs effectively.

The next evaluation is conducted on the TRECVID-05 Search Task. We follow the protocol detailed in section V-B.2. In particular, documents are incrementally retrieved through a Rele-

TABLE I

SEMANTIC FEATURE RETRIEVAL. FOR EVERY CONCEPT, THE TRAINING SET IS COMPOSED OF 20 POSITIVE AND 20 NEGATIVES EXAMPLES.

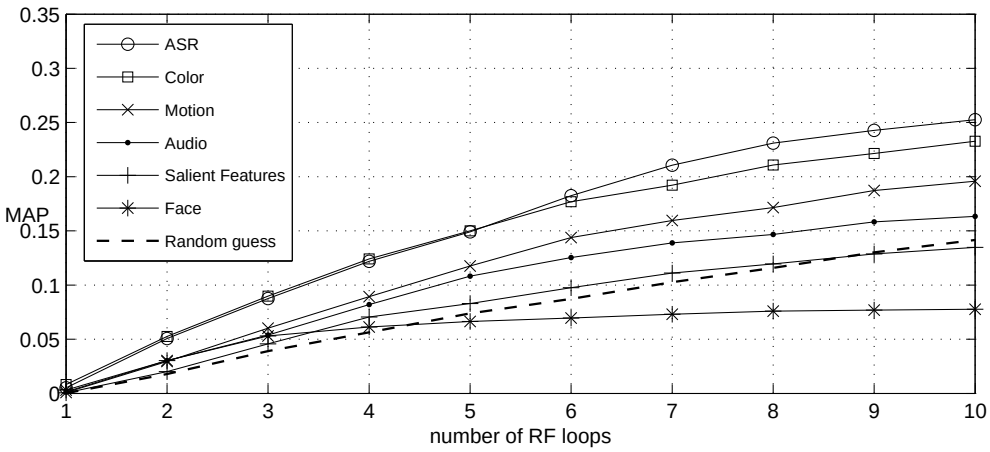
	HIER	HIERLin	CONC	SUM	Text	Color	Motion	Audio	Sal. Ft.	Face	Random
Airplane	0.037	0.000	0.002	0.020	0.001	**0.053**	0.001	0.000	0.004	0.001	0.0003
Animal	**0.060**	0.001	0.047	0.060	0.001	0.023	0.012	0.012	0.009	0.001	0.0004
Boat, Ship	0.015	0.002	0.008	**0.023**	0.001	0.020	0.002	0.002	0.003	0.000	0.0003
Building	0.025	0.009	0.020	0.025	0.012	**0.045**	0.023	0.017	0.024	0.022	0.0154
Bus	0.003	0.000	0.011	**0.034**	0.000	0.002	0.002	0.000	0.000	0.000	0.0002
Car	**0.016**	0.001	0.008	0.012	0.001	0.015	0.008	0.007	0.004	0.003	0.0034
Charts	**0.013**	0.002	0.008	0.011	0.005	0.005	0.002	0.002	0.002	0.000	0.0002
Computer, screen	**0.133**	0.040	0.042	0.114	0.017	0.064	0.025	0.022	0.007	0.092	0.0031
Corporate-Leader	0.010	0.008	0.007	**0.018**	0.008	0.007	0.003	0.003	0.001	0.004	0.0005
Court	**0.009**	0.002	0.002	0.001	0.004	0.007	0.001	0.001	0.000	0.006	0.0002
Crowd	**0.191**	0.037	0.084	0.054	0.011	0.087	0.106	0.056	0.109	0.030	0.0278
Desert	**0.010**	0.000	0.003	0.004	0.001	0.004	0.001	0.001	0.006	0.001	0.0004
Entertainment	**0.177**	0.039	0.113	0.002	0.002	0.086	0.050	0.140	0.028	0.033	0.0085
Explosion, Fire	**0.019**	0.001	0.006	0.002	0.001	0.002	0.001	0.001	0.013	0.001	0.0005
Face	0.816	**0.83**	0.714	0.712	0.574	0.367	0.346	0.335	0.475	0.711	0.2301
Flag-US	**0.023**	0.004	0.005	0.011	0.009	0.005	0.004	0.002	0.002	0.010	0.0004
Gvt-Leader	**0.067**	0.028	0.036	0.013	0.029	0.019	0.012	0.019	0.015	0.060	0.0083
Maps	**0.148**	0.010	0.072	0.106	0.008	0.035	0.007	0.010	0.003	0.047	0.0005
Meeting	**0.035**	0.007	0.032	0.018	0.003	0.030	0.013	0.011	0.020	0.012	0.0039
Military	**0.027**	0.000	0.013	0.006	0.002	0.013	0.012	0.007	0.008	0.000	0.0028
Mountain	0.008	0.000	0.002	**0.023**	0.000	0.005	0.001	0.002	0.011	0.000	0.0004
Natural-Disaster	0.008	0.000	0.001	**0.012**	0.000	0.001	0.000	0.000	0.003	0.001	0.0004
Office	**0.009**	0.001	0.007	0.005	0.001	0.004	0.001	0.004	0.004	0.001	0.0008
Outdoor	**0.534**	0.033	0.338	0.315	0.062	0.360	0.324	0.193	0.345	0.188	0.1319
People-Marching	**0.038**	0.000	0.016	0.008	0.002	0.005	0.010	0.004	0.011	0.001	0.0009
Person	0.979	0.947	0.965	**0.983**	0.781	0.767	0.620	0.602	0.798	0.965	0.5237
Police, Security	0.003	0.000	0.002	**0.005**	0.000	0.002	0.002	0.001	0.001	0.000	0.0004
Prisoner	0.018	0.001	0.003	**0.026**	0.000	0.006	0.003	0.002	0.001	0.020	0.0012
Road	0.016	0.001	0.011	0.010	0.002	**0.021**	0.008	0.007	0.009	0.005	0.0046
Sky	**0.189**	0.006	0.052	0.029	0.008	0.147	0.020	0.023	0.091	0.006	0.0202
Snow	0.014	0.000	**0.026**	0.023	0.001	0.007	0.001	0.005	0.007	0.000	0.0001
Sports	**0.039**	0.001	0.024	0.006	0.002	0.039	0.006	0.006	0.003	0.002	0.0007
Studio	**0.735**	0.494	0.637	0.667	0.314	0.473	0.178	0.234	0.060	0.615	0.0179
Truck	**0.004**	0.000	0.002	0.002	0.000	0.002	0.001	0.001	0.002	0.000	0.0003
Urban	0.021	0.002	0.016	0.022	0.004	**0.024**	0.015	0.010	0.014	0.006	0.0068
Vegetation	0.068	0.013	0.026	0.019	0.005	**0.104**	0.011	0.015	0.017	0.003	0.0110
Walking, Running	**0.088**	0.006	0.054	0.007	0.002	0.050	0.071	0.017	0.011	0.003	0.0107
Waterscape	**0.047**	0.005	0.011	0.020	0.000	0.032	0.002	0.003	0.010	0.000	0.0005
Weather	**0.254**	0.056	0.159	0.157	0.116	0.069	0.012	0.022	0.009	0.021	0.0003
MAP	**0.1258**	0.0663	0.0919	0.0920	0.0510	0.0772	0.0492	0.0461	0.0549	0.0736	0.0269

vance Feedback loop. The MAP is measured at each RF loop and monomodal searches (Figure 13.a) are compared with fusion strategies to measure the benefit raised from the combination of all modalities (Figure 13.b).

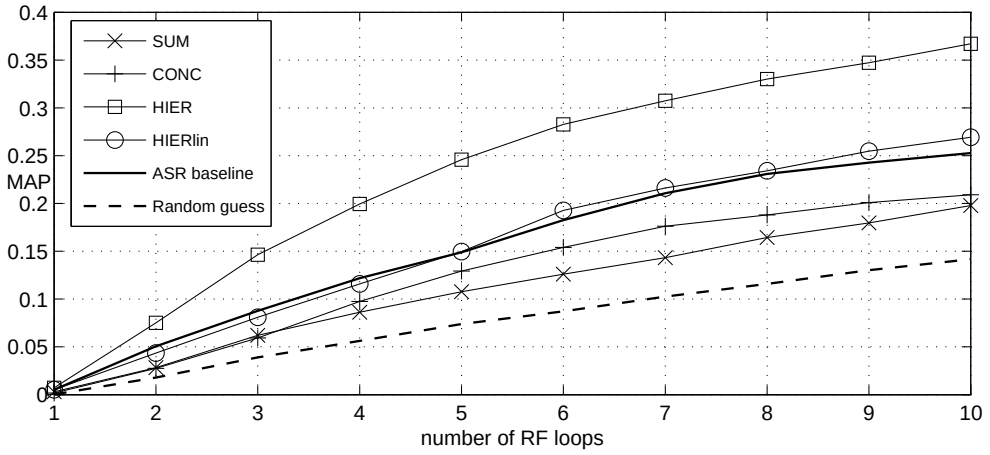
Examining monomodal results first, the figure is radically different from the previous experiment. We note now that text (ASR) is the most reliable information source. We also note that low-level audiovisual information (color, motion and audio) permits also to retrieve a significant fraction of relevant documents. On the other hand, high level visual features, *eg* face similarities and visual saliency, provide contrasted results: Though they perform comparably to motion or audio features for the first iterations, their global performance falls under the random guess precision as the number of examples grows. This behavior indicates that the retrieval model converges to a specific class of content (*eg* faces and objects) that excludes a significant part of relevant documents from the hit-list. This data remains unseen from the SVM during all RF loops, thus preventing retrieval model to generalize to all visual aspects of the sought topics.

We now look at fusing modalities. Figure 13.b displays the results obtained with the 3 strategies. Additionally, we also report random guess curve as a low baseline and the ASR curve as a high baseline. Fusion has an interest whenever it outperforms the best monomodal results (ASR in our case).

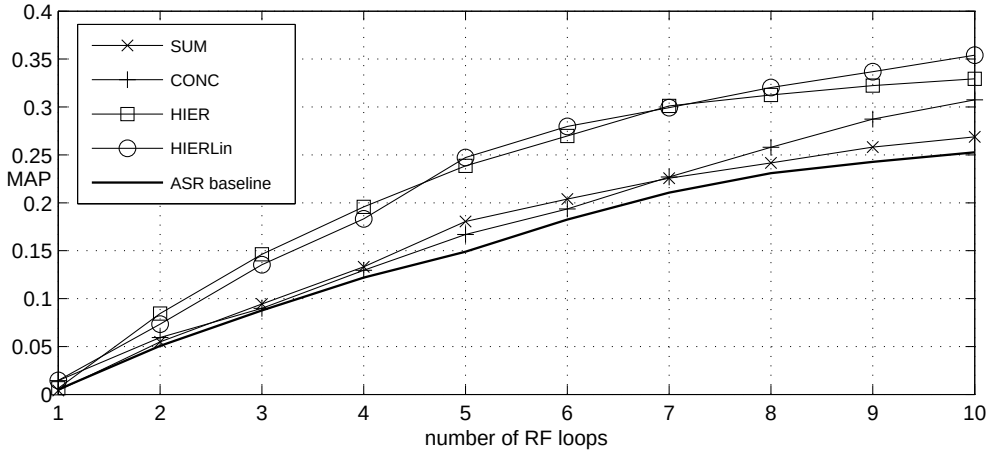
The results obtained are particularly clear. Compared to the ASR curve, only the HIER strategy provides significant improvement. Slight improvement is measured using HIERLin, while CONC and SUM remain below the "ASR only" MAP. This experiment seems to indicate that CONC and SUM are not robust to underperforming modalities, contrary to HIER, and to a lesser extent HIERLin. To definitively confirm this fact, the same experiment is conducted by removing the two corrupted modalities and fusing only reliable information (Figure 13.c). We now note that all strategies really benefit from the fusion, though HIER and HIERLin still outperform the SUM and CONC strategies. It worth noting also that surprisingly HIERLin outperforms HIER after iteration 5, indicating that the problem is linearly solvable as the dimensionality of the dissimilarity spaces increases (the growth is induced by the training set expansion). Finally, the most important observation is made by comparing the HIER performance in Figure 13.b and 13.c. Fusing the 6 modalities (with two corrupted) is more effective than using only the reliable channels. This shows that HIER is able to add only positive contributions from each modality, even when this contribution is very moderate.



(a)



(b)



(c)

Fig. 13. TRECVID 05 Search Task: a) Relevance feedback simulation for every modality, b) when fusing 6 modalities and c) when fusing only the 4 best modalities (ASR, Color, Motion and Audio). At each iteration, up to 10 positive and negative examples are added to the query. SVM with automatic scale setting is used (where appropriate).

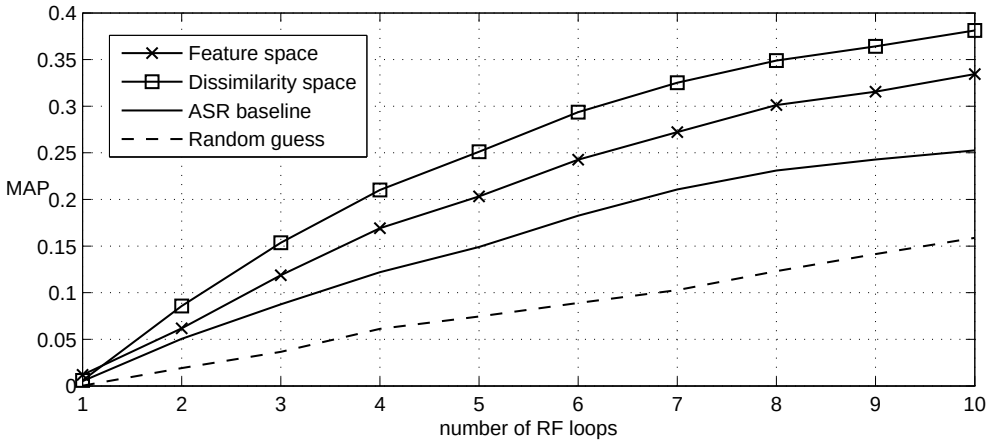


Fig. 14. TRECVID 05 Search Task: Comparison of HIER fusion strategy operated in feature spaces and in corresponding dissimilarity spaces. The six modalities are used.

Finally, we compare the HIER fusion strategy when applied directly in feature spaces and in the corresponding dissimilarity spaces (let us note that this hierarchical approach for fusing feature spaces has been already proposed in [35] and is considered as a baseline technique). For the six feature spaces, RBF-kernels of base classifiers are adapted to their respective distance measures, $k_d(x, y) = e^{-\frac{d(x,y)}{2\sigma^2}}$, as suggested in [5]. Similarly to dissimilarity space, the scale parameter is estimated for each query by equation (7). The results shown in Figure 14 indicate that the HIER strategy is also performing well in the original feature spaces (it actually largely outperforms the ASR search). However the MAP curve stays below the one obtained in dissimilarity space. In fact, this result generalizes to the multimodal case what we observed for monomodal learning (section V-B.2, Figure 6): For an asymmetric setup and with a similar learning scheme, it is actually more effective to work in dissimilarity space than in feature space.

D. Real time constraint

We finally provide the computation load of each algorithm (Table II) for various training sets and for 3 and 10 modalities. These times have been obtained on a standard PC PIV 2GHz (Matlab implementation). 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.

TABLE II

COMPUTATION LOAD (IN SECOND)

	Training set		Algorithms		
	$\lvert\mathcal{P}\rvert$	$\lvert\mathcal{N}\rvert$	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

VI. DISCUSSION

In this paper, we have presented a novel dissimilarity-based approach providing an original well-founded solution to the problem of multimodal fusion in interactive content-based retrieval setup. Data are projected in query-dependent dissimilarity spaces. We have shown that this representation transforms the asymmetric learning problem into a binary one, which in turn enhances the performance. By construction, those spaces are inherently of low dimension, thus simplifying further the complexity of the data representation as well as the processing of the queries. The processing of the queries consists in incrementally training a kernelized SVM from examples provided by users through relevance feedback loops. In addition, we have also proposed and validated a way of automatically setting the kernel scale to adapt the learning to each RF loop.

On the basis of the proposed dissimilarity space, we have designed three fusion strategies allowing to combine dissimilarities coming from various feature spaces/modalities. Exhaustive evaluations on both artificial and real video data allow us to better understand the behavior and to rank the efficiency of the three approaches for retrieval tasks. Tests on artificial data have demonstrated the superiority of the HIER strategies in terms of class discrimination and robustness to corrupted modalities. Experiments carried out on a large scale video retrieval benchmark further confirm this superiority. The HIER algorithm appears to be the only fusion scheme able to benefit from multimodality whenever some channels appear to be corrupted or to be uninformative with respect to the sought topics. As a consequence, this strategy is definitively to be considered for implementing multimodal retrieval, even if the computation overhead is not

negligible compared to CONC and SUM.

However, further improvements should be added to overcome various shortcomings such as specific learning algorithm for imbalance data, strategies for active learning or fast access to the dissimilarity data. As stated in the introduction, state-of-the art indexing and machine learning techniques are already available to solve these problems. On the basis of the proposed dissimilarity space, these techniques could easily be incorporated at different stages of the algorithm (indexing structure, machine learning algorithms, RF paradigm) so as to offer a scalable and effective multimedia retrieval system.

REFERENCES

[1] V. Athitsos, J. Alon, S. Sclaroff, and G. Kollios. BoostMap: A method for efficient approximate similarity rankings. In *CVPR (2)*, pages 268–275, 2004.

[2] L. Boldareva and D. 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.

[3] Eric Bruno, Nicolas Moenne-Loccoz, and Stphane Marchand Maillet. Learning user queries in multimodal dissimilarity spaces. In *Proceedings of the 3rd International Workshop on Adaptive Multimedia Retrieval, AMR'05*, Glasgow, UK, July 2005.

[4] Eric Bruno, Nicolas Moenne-Loccoz, and Stéphane Marchand-Maillet. Unsupervised event discrimination based on nonlinear temporal modelling of activity. *Pattern Analysis and Application*, 7(4):402–410, December 2004.

[5] A. B. Chan and N. Vasconcelos. Classifying video with kernel dynamic textures. In *IEEE Conference on Computer Vision and Pattern Recognition (to appear)*, 2007.

[6] 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.

[7] E. Chávez, G. Navarro, R. Baeza-Yates, and J.L. Marroquin. Searching in metric spaces. *ACM Computing Surveys*, 33(3):273–321, September 2001.

[8] Y. Chen, X.S. Zhou, and T.S. Huang. One-class svm for learning in image retrieval. In *IEEE International Conference on Image Processing*, 2001.

[9] T.F. Cox and M.A.A. Cox. *Multidimensional scaling*. Chapman & Hall, London, 1995.

[10] Nello Cristianini, John Shawe-Taylor, Andre Elisseeff, and Jaz Kandola. On kernel-target alignment. In *Advances In Neural Information Processing Systems, Nips*, 2001.

[11] 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.

[12] C. Faloutsos and K. Lin. FastMap: A fast algorithm for indexing, data-mining and visualization of traditional and multimedia datasets. In *Proceedings of the 1995 ACM SIGMOD International Conference on Management of Data*, pages 163–174, San Jose, California, 22–25 1995.

[13] J. Gu, L. Lu, H.J Zhang, and J. Yang. Dominant feature vectors based audio similarity measure. In *PCM*, number 2, pages 890–897, 2004.

[14] T. Hastie, R. Tibshirani, and J. Friedman. *The elements of statistical learning*. Springer Series in Statistics. Springer-Verlag, New York, 2001.

[15] D Heesch and S Rueger. NNk networks for content-based image retrieval. In *26th European Conference on Information Retrieval*, Sunderland, UK, 2004.

[16] Winston H. Hsu and Shih-Fu Chang. Generative, discriminative, and ensemble learning on multi-modal perceptual fusion toward news video story segmentation. In *ICME*, Taipei, Taiwan, June 2004.

[17] A.K. Jain, A. Vailaya, and X. Wei. Query by video clip. *Multimedia Syst.*, 7(5):369–384, 1999.

[18] J. Kittler, M. Hatef, R.P.W. Duin, and J. Matas. On combining classifiers. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 20(3):226–239, 1998.

[19] D.G Lowe. Object recognition fron local scale invariant features. In *Proceedings of the International Conference in Computer Vision, ICCV'99*, pages 1150–1157, Corfu, 1999.

[20] Nicolas Moënne-Loccoz, Eric Bruno, and Stéphane Marchand-Maillet. Interactive partial matching of video sequences in large collections. In *IEEE International Conference on Image Processing*, Genova, Italy, 11-14 September 2005.

[21] M. R. Naphade, L. Kennedy, J. R. Kender, S.-F. Chang, J. R. Smith, P. Over, , and A. Hauptmann. A light scale concept ontology for multimedia understanding for trecvid 2005. Technical report, IBM Research, 2005.

[22] G. P. Nguyen, M. Worring, and A. W. M. Smeulders. Similarity learning via dissimilarity space in CBIR. In *MIR '06: Proceedings of the 8th ACM international workshop on Multimedia information retrieval*, pages 107–116, New York, NY, USA, 2006. ACM Press.

[23] C.S. Ong, A.J. Smola, and R.C. Williamson. Hyperkernels. In *Advances in Neural Information Processing Systems, NIPS*, number 15, 2003.

[24] P. Over, T. Ianeva, W. Kraaij, and A. F. Smeaton. Trecvid 2005 an overview. In *In Proceedings of TRECVID 2005, 2005. NIST, USA*, 2005.

[25] N.C. Oza, R. Polikar, J. Kittler, and F. Roli. Multiple classifier systems. In *Series: Lecture Notes in Computer Science*, volume 3541. Springer, 2005.

[26] E. Pekalska and R.P.W. Duin. The use of dissimilarities for object recognition. In *Proceedings of EOS Conference on Industrial Image and Machine Vision*, pages 50–53, Hannover, Germany, 2005. EOS European Optical Society.

[27] 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.

[28] 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.

[29] 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.

[30] K. Tieu and P. Viola. Boosting image retrieval. In *International Conference of Computer Vision, ICCV'01*, pages 228–235, 2001.

[31] S. Tong and E. Chang. Support vector machine active learning for image retrieval. In *ACM International Conference on Mutlimedia*, pages 107–118, 2001.

[32] V. Vapnik. *The Nature of Statistical Learning Theory*. Springer Verlag, New York, 1995.

[33] P. Viola and M. Jones. Robust real-time face detection. *International Journal of Computer Vision (IJCV)*, 57(2):137–154, 2004.

[34] J. T. Wang, X. Wang, D. Shasha, and K. Zhang. MetricMap: An embedding technique for processing distance-based queries in metric spaces. *IEEE Transactions on Systems, Man, and Cybernetics, Part B*, 35(5):973–987, 2005.

[35] 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.

[36] Huilin Xiong, M.N.S. Swamy, and M.O. Ahmad. Optimizing the kernel in the empirical feature space. *IEEE Transactions on Neural Networks*, 16(2):460–474, March 2005.

[37] 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.

[38] J. Yang and A.G. Hauptmann. Multi-modality analysis for person type classification in news video. In *Electronic Imaging'05 - Conference on Storage and Retrieval Methods and Applications for Multimedia*, San Jose, USA, Jan 2005.

[39] P. Zezula, P. Savino, G. Amato, and F. Rabitti. Approximate similarity retrieval with m-trees. *VLDB Journal*, 7(4):275–293, 1998.

[40] 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.

[41] 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.