Combining Multimodal Preferences for Multimedia Information Retrieval

Bruno, Eric; Kludas, Jana; Marchand-Maillet, Stéphane

How to cite

BRUNO, Eric, KLUDAS, Jana, MARCHAND-MAILLET, Stéphane. Combining Multimodal Preferences for Multimedia Information Retrieval. In: Proceedings of the 9th ACM SIGMM International Workshop on Multimedia Information Retrieval, MIR 2007. Augsburg (Germany). [s.l.] : [s.n.], 2007. p. 71–78. doi: 10.1145/1290082.1290095

This publication URL: https://archive-ouverte.unige.ch//unige:47725

Publication DOI: 10.1145/1290082.1290095

Combining Multimodal Preferences for Multimedia Information Retrieval

Eric Bruno, Jana Kludas, Stéphane Marchand Maillet
Viper group - Computer Vision and Multimedia Lab
University of Geneva, Switzerland
name.surname@cui.unige.ch

ABSTRACT

Representing and fusing multimedia information is a key issue to discover semantics in multimedia. In this paper we address more specifically the problem of multimedia content retrieval by first defining a novel preference-based representation particularly adapted to the fusion problem, and then, by investigating the RankBoost algorithm to combine those preferences and a learn multimodal retrieval model. The approach has been tested on annotated images and on the complete TRECVID 2005 corpus and compared with SVM-based fusion strategies. The results show that our approach equals SVM performance but, contrary to SVM, is parameter free and faster.

Categories and Subject Descriptors

H.5.1.f [**Image/video retrieval**]; I.2.6.g [**Machine learning**]; I.2.6.b [**Concept learning**]

General Terms

Theory

Keywords

multimodal fusion, multimedia indexing and retrieval, RankBoost

1. 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 us 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 where 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 [5], imbalance classification algorithms [28], automatic kernel setting [27] or automatic labelling of training data [25]. 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 [12, 6]; Wu *et al* propose the super-kernel fusion to determine optimal combination of features for video retrieval [24]. In [10], 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 [16].

All these studies have in common the fact that they consider feature spaces to represent knowledge from 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 heterogeneous feature space components. This problem is even more sensible when similarity measures are complex functions or procedures, such as prediction functions for temporal distances [3] or graph exploration for semantic similarities [19]. The diversity of the features involved is also a difficulty when dealing with fusion and learning. The multimedia descriptors may indeed 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 [22][26].

In [2], we considered an alternative representation based on dissimilarity spaces [17]. This representation is a first step to obtain a unified representation of multimodal content but still leaves open the problem of how to properly scale the similarity values to make them homogeneous. In this paper, we propose to consider only ordering information from dissimilarity spaces. The result is a *preference space* where every item is indexed with ranking positions. The

MIR'07, September 28–29, 2007, Augsburg, Bavaria, Germany.

scaling issue is thus completely alleviated and we effectively obtain a unified representation of multimodal content, but at the price of losing an important amount of the initial information (section 3). Retrieving items from the preference space is then a ranking problem (section 4) that can be addressed using the RankBoost algorithm (section 5). Experiments on artificial and real data (images and videos) show that the preference space associated to RankBoost competes with SVM learning in feature and dissimilarity spaces and is therefore a valid approach for multimodal retrieval (section 6).

2. PROBLEM DEFINITION

Let us consider a collection $\mathcal{X}$ containing l multimedia documents x that we are interested in ranking.

The *query by example* search paradigm consists in gathering user's judgements indicating, for some objects, whether they are relevant or irrelevant to the user request. This set, denoted $\mathcal{Q}$, is called the *query* and is composed of positive and negative subsets, respectively

$$\mathcal{P} = \{x_i^+\}_{i=1}^p \text{ and } \mathcal{N} = \{x_i^-\}_{i=1}^n,$$

with $\mathcal{Q} = \mathcal{P} \cup \mathcal{N}$ and $q = p+n$. The query $\mathcal{Q}$ is then used to train a machine that will produce a decision function ranking documents according to their relevance to the query.

This paradigm might be embedded in the *Relevance Feedback* (RF) strategy, where these two steps (user judgement and ranking estimation) are iterated until the search converges to a satisfactory result.

3. MULTIMODAL CONTENT REPRESENTATIONS

Expressing multimodal content involves first to extract various descriptors from the multimedia objects. Ideally, each descriptor depicts an appropriate aspect of the multimodal features of the documents. Assuming such descriptors are available, we discuss in the following how efficient representations may be derived to store descriptors and to facilitate their fusion.

3.1 Feature-based representation

Assuming m distinct descriptors are designed (and extraction procedures implemented), the multimodal represention of an object x is the set of m feature vectors $\{\mathbf{x}^k\}_{k=1}^m$ living respectively in feature spaces $\{\mathcal{F}^k\}_{k=1}^m$. The dimension of each feature space intrinsically depends of the descriptor they express. The feature-based representation is rather straightforward, but not really convenient since it mixes heterogeneous vectors of various dimensions and scales. Fusion and ranking algorithms need to manage the diversity of the representation, thus making them more dependent on complex parameter setting procedures and less flexible to handle new descriptors.

To avoid this situation, modality-independent representations are desirable. For that purpose, (dis)similarity-based representations have been recently proposed [1, 2, 9, 15]. As pointed out by these authors, similarities are convenient to manipulate multimodal information since they form a homogeneous representation of the content. Moreover, similarity representations are generally made such as their dimensionality remain much lower than their feature counterparts.

3.2 Dissimilarity-based representation

In [4], we proposed a *Query-based Dissimilarity Space* (QDS), derived from the dissimilarity spaces introduced by Pekalska *et al* [17]. For a given feature space $\mathcal{F}^k$, the corresponding QDS, denoted $\mathcal{D}_\mathcal{P}^k$, is defined relatively to the positive set $\mathcal{P}$ by the mapping $\mathbf{d}^k(x,\mathcal{P}) \in \mathbb{R}^p$

$$\mathbf{d}^k(x,\mathcal{P}) = [d^k(x,x_1^+), d^k(x,x_2^+), \ldots \, d^k(x,x_p^+)]^T, \qquad (1)$$

where $d^k(x,x_i^+) \in \mathbb{R}^+$ is the dissimilarity from any object $x \in \mathcal{X}$ to the prototype x_i^+ when the measure is done in $\mathcal{F}^k$. Using QDS, an object x is thus represented with a set of m dissimilarity vectors $\{\mathbf{d}^k\}_{k=1}^m$ living in p-dimensional dissimilarity spaces $\{\mathcal{D}_\mathcal{P}^k\}_{k=1}^m$,

$$\mathcal{D}_\mathcal{P}^k = \begin{pmatrix} d^k(x_1,x_1^+) & d^k(x_2,x_1^+) & \ldots & d^k(x_l,x_1^+) \\ d^k(x_1,x_2^+) & d^k(x_2,x_2^+) & \ldots & d^k(x_l,x_2^+) \\ & & \vdots & \\ d^k(x_1,x_p^+) & d^k(x_2,x_p^+) & \ldots & d^k(x_l,x_p^+) \end{pmatrix}. \quad (2)$$

The QDS approach provides a unified representation of multimodal information channels. Moreover it is particularly adapted to the class asymmetry typically exhibited by the positive and negative classes [4]. However, the issue of how properly scaling dissimilarity spaces so that modalities become easily comparable still remains . This problem might be left out to the fusion and ranking algorithms [2], but a more elegant solution would be to end up with a fully homogeneous multimodal representation.

3.3 Preference-based representation

We propose to simplify the QDS representation by replacing the dissimilarity components $d^k(x,x_i^+)$ with the ranking position $\pi_i^k(x) \in \mathbb{N}$ of an object x with respect to the prototype x_i^+ according to the dissimilarity d^k and the collection $\mathcal{X}$,

$$\pi_i^k(x) = \sum_{x_j \in \mathcal{X}} [\![d^k(x_j,x_i^+) \leq d^k(x,x_i^+)]\!]. \qquad (3)$$

The notation $[\![\kappa]\!]$ is defined to be 1 if predicate κ holds and 0 otherwise. For the sake of readability, the double indices notation π_i^k is simplified to π_j, $j = 1, \ldots, pm$, with j iterating over all objects $x_i^+ \in \mathcal{P}$ and for the m modalities.

The multimodal representation of an object x now becomes a unique vector of *preferences* $\boldsymbol{\pi}(x) = [\pi_1(x), \ldots, \pi_{pm}(x)]^T$. The multimodal *preference space* embedding all objects $x \in \mathcal{X}$ is

$$\Pi_\mathcal{P} = \begin{pmatrix} \pi_1(x_1) & \pi_1(x_2) & \ldots & \pi_1(x_l) \\ \pi_2(x_1) & \pi_2(x_2) & \ldots & \pi_2(x_l) \\ & & \vdots & \\ \pi_{pm}(x_1) & \pi_{pm}(x_2) & \ldots & \pi_{pm}(x_l) \end{pmatrix}. \quad (4)$$

It consists in a unique pm-dimensional space providing a fully homogeneous representation of multimodal information. Similarly to the QDS approach, $\Pi_\mathcal{P}$ represents the two classes $\mathcal{P}$ and $\mathcal{N}$ asymmetrically since every element is evaluated relatively to the positive instances only. We will see later how this asymmetry is decisive in learning accurate rankings with only linear functions.

It is worth noting however that we obtain this modality-independent representation at the price of losing most information about the initial feature distributions; only ordering information is actually preserved. Our objective now is to define a machine learning effectively able to learn from preferences as efficiently as learning directly in feature spaces or in dissimilarity spaces.

4. THE RANKING PROBLEM

The ranking problem could be formulated as follows: For each item $x \in \mathcal{X}$, it exists ranking features $\pi_1, \ldots, \pi_{pm}$ ordering x from most preferred to least preferred. In our formulation, $\pi_i \in \mathbb{N}$ and $\pi_i(x_1) < \pi_i(x_0)$ means x_1 preferred to x_0.

Additionally to the ranking features, there exists a *feedback* function $\Phi : \mathcal{X} \times \mathcal{X}$ which provides to the learner the desired form of the final ranking. Formally $\Phi(x_1, x_0) > 0$ means that x_1 should be ranked above x_0 while $\Phi(x_1, x_0) < 0$ means the opposite. $\Phi(x_1, x_0) = 0$ means no preferences between x_0 and x_1 and the magnitude of $|\Phi(x_1, x_0)|$ indicates how important is to rank x_1 above or below x_0. The *bipartite* feedback function is special but common case in document retrieval: the function is said bipartite if there exists two disjoint set $\mathcal{X}_1$ and $\mathcal{X}_0$ such that Φ ranks all instances x_1 of $\mathcal{X}_1$ above instances x_0 of $\mathcal{X}_0$. These subsets are respectively the positive and negative subsets $\mathcal{P}$ and $\mathcal{N}$ we defined in section 2.

Learning such a feedback function implies estimating a ranking $H : \mathcal{X} \to \mathbb{R}$ through the optimization of a ranking loss function penalizing every mis-ordered pair of items. We consider the loss proposed in [7]

$$\sum_{\substack{x^- \in \mathcal{N} \\ x^+ \in \mathcal{P}}} \Phi(x^+, x^-) \left[H(x^+) - H(x^-) \right]. \qquad (5)$$

The function $H(x)$ is a ranking of items x stating that x^+ is ranked higher than x^- whenever $H(x^+) > H(x^-)$. Interestingly, in case of bipartite feedback, the problem becomes separable and the ranking loss simplifies to [7]

$$\sum_{x \in \mathcal{Q}} w(x) s(x) H(x), \qquad (6)$$

where the user feedback is carried by both

$$s(x) = \begin{cases} +1 & \text{if } x \in \mathcal{P} \\ -1 & \text{if } x \in \mathcal{N} \end{cases}, \qquad (7)$$

and $w(x)$ a weight giving the importance of the rank of the item x.

5. RANKBOOST

Following the boosting principle, the final ranking H results from a weighted sum of weak rankings $h_t : \mathcal{X} \to \mathbb{R}$

$$H(x) = \sum_{t=1}^{T} \alpha_t h_t(x), \qquad (8)$$

which is estimated through an Adaboost-like algorithm, namely RankBoost [7] (see Figure 1). This greedy coordinate-wise search algorithm aims at iteratively minimizing the normalization factor Z_t by choosing at each round an appropriate pair $\{\alpha_t,\ h_t\}$. For a given weak hypothesis $h_t \in [-1, 1]$, it has been shown [20] that Z_t is minimized for

$$\alpha_t = \frac{1}{2} \ln \frac{1 + r_t}{1 - r_t}, \qquad (9)$$

```
Given two disjoint subsets $\mathcal{P}$ and $\mathcal{N}$ and labels $s(x)$
over $\mathcal{P} \cup \mathcal{N}$ as defined in (7)

Initialize
    $w_1(x) = \begin{cases} 1/p & \text{if } x \in \mathcal{P} \\ 1/n & \text{if } x \in \mathcal{N} \end{cases}$

For $t = 1, \ldots, T$
    – Train weak learner using $w_t$
    – Get weak ranking $h_t : \mathcal{X} \to \mathbb{R}$
    – Compute $r = \sum_x w_t(x) s(x) h_t(x)$
    – Choose $\alpha_t \in \mathbb{R}$
    – Update
      $w_{t+1}(x) = \frac{1}{Z_t} w_t(x) e^{-\alpha_t s(x) h_t(x)}$
      where $Z_t$ is a normalization factor

Output the final ranking $H(x) = \sum_{t=1}^{T} \alpha_t h_t(x)$
```

Figure 1: The RankBoost algorithm for bipartite feedback

where r is the weighted classification rate

$$r_t = w_t(x) s(x) h_t(x). \qquad (10)$$

The algorithm is run over a number T of iterations which is predefined or may depend on the training error. In our implementation, the loop is stopped whenever the training error is equal to 0, with a maximum of $2pm$ iterations.

5.1 Weak ranking

The weak ranking h_t is produced through a *weak learner*. It has to provide a new ranking from ranking features π_i conforming the best to the bipartite feedback. For example, the weak learner proposed in [7] selects at each iteration the ranking feature π_i minimizing the training error. The output preserves only relative-ordering information so as to be independent of specific preference values,

$$h(x) = \begin{cases} +1 & \text{if } \pi_i(x) < \theta \\ -1 & \text{if } \pi_i(x) \geq \theta \end{cases}. \qquad (11)$$

As illustrated in Figure 2, this weak learner consists in fitting a step function to the user feedbacks $\{s(x_j)\}_{j=1}^{q}$ sorted by increasing order of $\pi_i(x_j)$. The best weak ranking is the one maximizing equation (10) over the q candidate weak rankings for the pm preferences π_i. The evaluation of all candidates is done in $O(qpm)$.

As defined in (11), the function $h(x)$ provides at each iteration a binary ranking. The final ranking H (eq. (8)) is thus an injection on $\mathcal{X}$ whose image has as at most cardinality 2^T, *ie* $H(x) : \mathcal{X} \to \{v^1, \ldots, v^{2^T}\}$. Typically when the training set is small or when the ranking problem is simple, RankBoost converges in a few iterations (T small) and consequently provides a coarse ranking partitioning the collection $\mathcal{X}$ in few blocks. To get a finer ranking, we propose to use the a soft ranking function,

$$h(x) = 2e^{-\gamma \pi_i^2(x)} - 1. \qquad (12)$$

Learning this weak ranking consists of choosing the pair

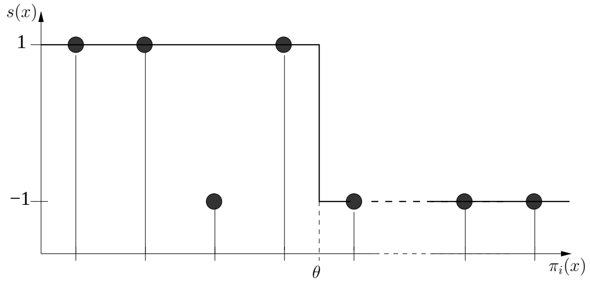


Figure 2: Binary weak ranking. The $\pi_i(x)$'s are ordered in increasing order.

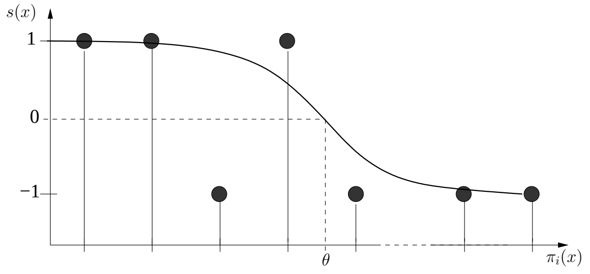


Figure 3: Soft weak ranking. The $\pi_i(x)$'s are ordered increasing order.

(π_i, γ) that maximize the classification rate r_t (10). Given a ranking feature π_i, a grid search on γ is achieved rather than a time-consuming non-linear regression. The grid vertices are positioned at the middle of the q ranking intervals (see Figure 3). With this approximation, the weak learner complexity remains $O(qpm)$.

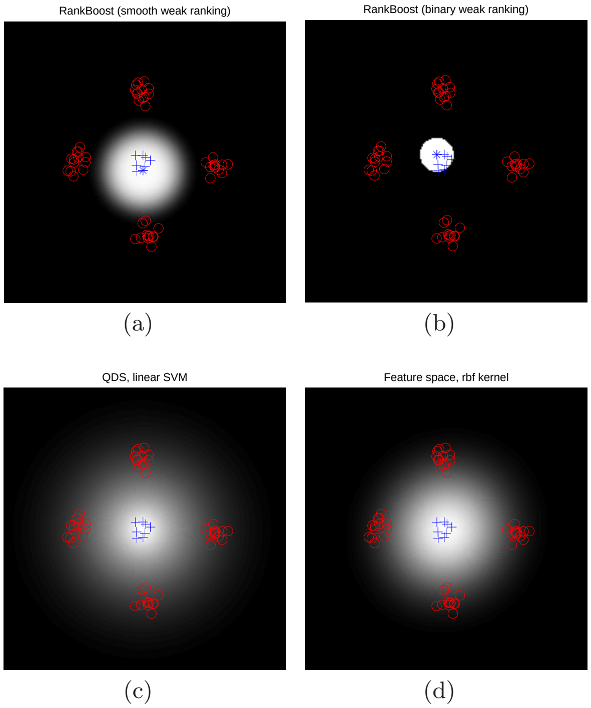
6. EXPERIMENTS

The behavior and performance of ranking data in the three representation spaces (feature, dissimilarity and preference) are studied here. As stated before, RankBoost (soft and binary weak ranking) will be used to learn preferences. As far as feature and dissimilarity spaces are concerned, ranking are produced with the SVM algorithm. Depending of experiment, linear or non-linear (*eg* using RBF kernel) SVM is used.

6.1 Toy examples

Artificial data allows us to concretely illustrate how rankings are learned in the various representation spaces. The following toy examples are made so as to be representative of the class asymmetry we generally meet in real applications. For every learning technique, the learned ranking is superimposed to the items; white areas correspond to top ranks and black areaus to the last rank. Moreover, prototypes selected by Rankboost are indicated with the $*$ marker.

The first example (Figure 4) corresponds to an ideal separable case, where all the positive instances (cross marker) belong to the same cluster, while the negative samples are distributed around (circle marker). The corresponding dissimilarity space is built using pairwise Euclidean distances while the preference space is derived by ordering dissimilarities. Linear SVM is used to learn in QDS while a RBF-SVM with an appropriate scale parameter operates in feature space.


RankBoost (smooth weak ranking)

RankBoost (binary weak ranking)

(a)

(b)

QDS, linear SVM

Feature space, rbf kernel

(c)

(d)

Figure 4: Cross toy example

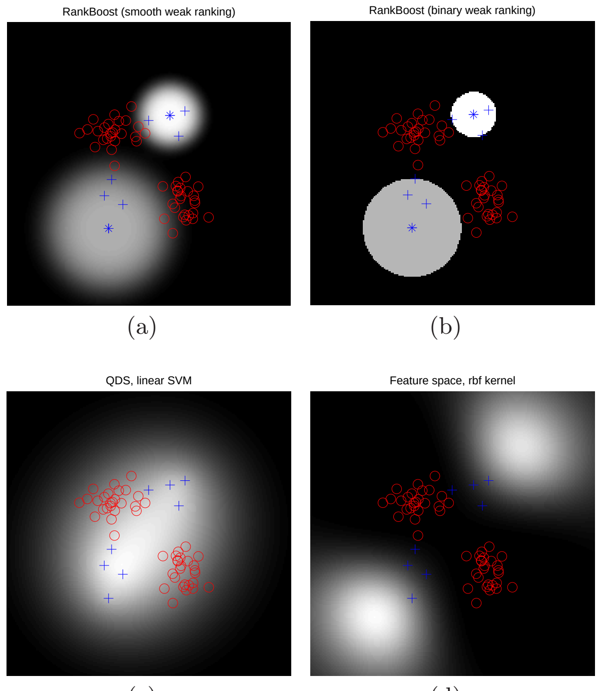
In each case (preferences, dissimilarities and features, respectively in Figure 4.a, b, c and d), a perfect ranking has been estimated. As the class setup is simple, only one weak ranking (indicated by the selected prototype) is necessary for RankBoost (Figure 4.a and b). It implies that the final ranking is binary when using the binary weak ranking function, while learning with the soft weak ranking provides us with a more convenient continuous ranking. Notice that a linear function is also enough to catch the positive class within dissimilarity space (see [4] for justifications), while a non-linear RBF-based ranking function is needed in feature space.

The second example (Figure 5) depicts a less obvious problem, the XOR configuration. The classes are no longer linearly separable neither in preference space nor in dissimilarity space. In that case, the linear SVM used in QDS failed in estimating the ranking. On contrary, RankBoost succeeds in finding the two positive clusters and selects one prototype per cluster.

6.2 Real data

6.2.1 Corel image collection

The studied image collection is a subset of the Corel collection. It contains 1159 images annotated with 1 to 10 keywords per image (including some non-sense descriptions). The images are categorized into 49 classes. Textual and visual features are considered for fusing experiments: The vector space model $\mathcal{F}^{\text{text}}$ containing tf-idf weights is built from keywords (2035 terms). The color space $\mathcal{F}^{\text{color}}$ contains 166 bins HSV histograms and the texture space $\mathcal{F}^{\text{texture}}$ is made



(a) (b)

(c) (d)

Figure 5: XOR toy example

of Gabor filter bank outputs (120 dimensions). Cosine distance is considered for textual features, while Euclidean is used in visual feature spaces.

Fusion is operated in feature space, dissimilarity space and preference space. In feature and dissimilarity space we have considered a state-of-the-art hierarchical fusion scheme [2, 24]. At the first level, base classifiers are trained in each monomodal space. At the second level, a super classifier is used to fuse soft-outputs of all base classifiers. Base classifiers and super classifier are RBF SVM. Optimal classifier parameters have been determined through a leave-one-out cross validation.

Retrieval performance is given in terms of Mean Average Precision (MAP). Average Precision (AP) 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. The MAP is simply the AP averaged over several classes. Additionally to the algorithm performance, a baseline consisting in retrieving randomly documents is always provided. All results are displayed in Figure 6.

Multimodal retrieval (Figure 6.b) and text-only search (Figure 6.a) are studied. In both cases we observe that for RankBoost, soft ranking outperforms largely binary ranking. Moreover, the soft ranking performs similarly to the SVM approaches whereas it uses only a degraded version of the original features. The second observation we can make is that the multimodal retrieval outperforms only very slightly the keyword-only search, whatever the approach considered. This result seems to indicate that keywords bring much of the category information, and that color or texture low-level information are of little help in that case. This observation is confirmed by analyzing the ranking features selected by RankBoost to build retrieval models: among all retrieval instances, text information is used for 93% of them, while color and texture features are only used for respectively 36% and 17% of the cases.

6.2.2 *TRECVID video corpus*

We now consider the TRECVID 2005 benchmark. In our setup, videos are segmented into around 89'500 segments using the common shot reference [18]. 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) are provided as positive examples. Further details about the Search Task may be found in [21]. During the experiments, we only considered video shots as positive examples. The positive examples are completed with ten negative examples randomly selected 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 search depth of 1000 is given by TRECVID). The process is repeated ten times. Following the TRECVID evaluation protocol, the performance was measured at each iteration by MAP at 1000. Additionally to the algorithm performance, a baseline consisting of retrieving randomly documents is always provided.

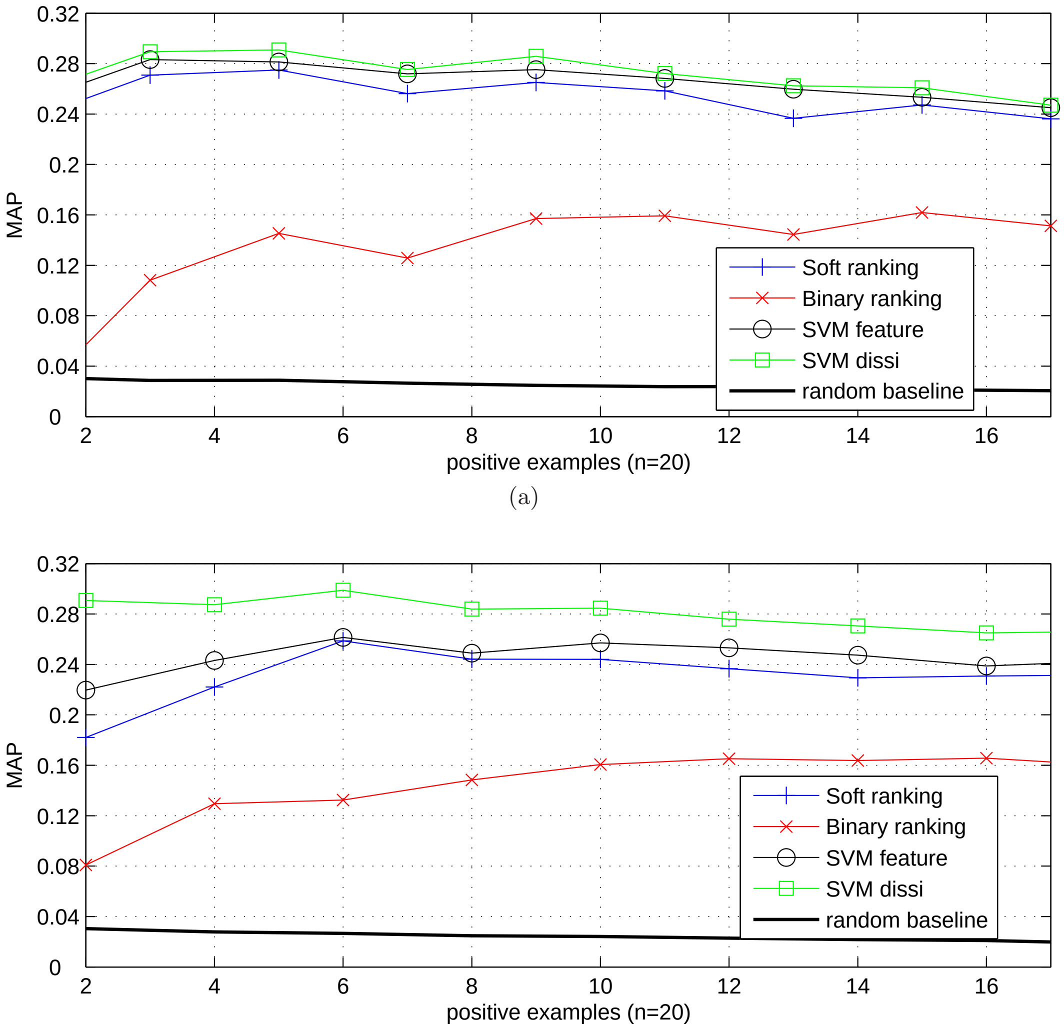
The multimodal features are derived from the six following text and audiovisual descriptors:

- Color histogram, $4 \times 4 \times 4$ bins in YCbCr space
- Motion vector histogram, 66 bins quantization of the MPEG block motion vectors [11]
- Local features, SIFT descriptors extracted around the Lowe salient points [13],
- Face detection [23],
- Word occurrence histogram (vector space model) computed from ASR,
- Dominant audio features [8] extracted from the audio stream.

The distance measures used are Euclidean for color and motion histograms. An approximation of the minimal matching distance is applied on local features to determine partial similarities [14]. Euclidean distance in the 30-dimensional eigenface space gives the similarity between the detected faces. Cosine distance is used for the vector space model and finally the audio similarity measure proposed in [8] is used for audio features.

The fusion strategies remain the hierarchical RBF SVM approach in feature spaces and dissimilarity spaces. For feature space however, we adapt the RBF-kernel to the distances used, $k_d(x, y) = e^{-\frac{d(x,y)}{2\sigma^2}}$ (it is worth noting that k_d is strictly a RBF-kernel when d is an Euclidean distance). Optimal classifier parameters have been cross-validated using the TRECVID development set.

[1] the speech transcripts extracted by Automatic Speech Recognition (ASR) are also available.



(a)

(b)

Figure 6: Image retrieval results with a) multimodal fusion and b) keywords-only search

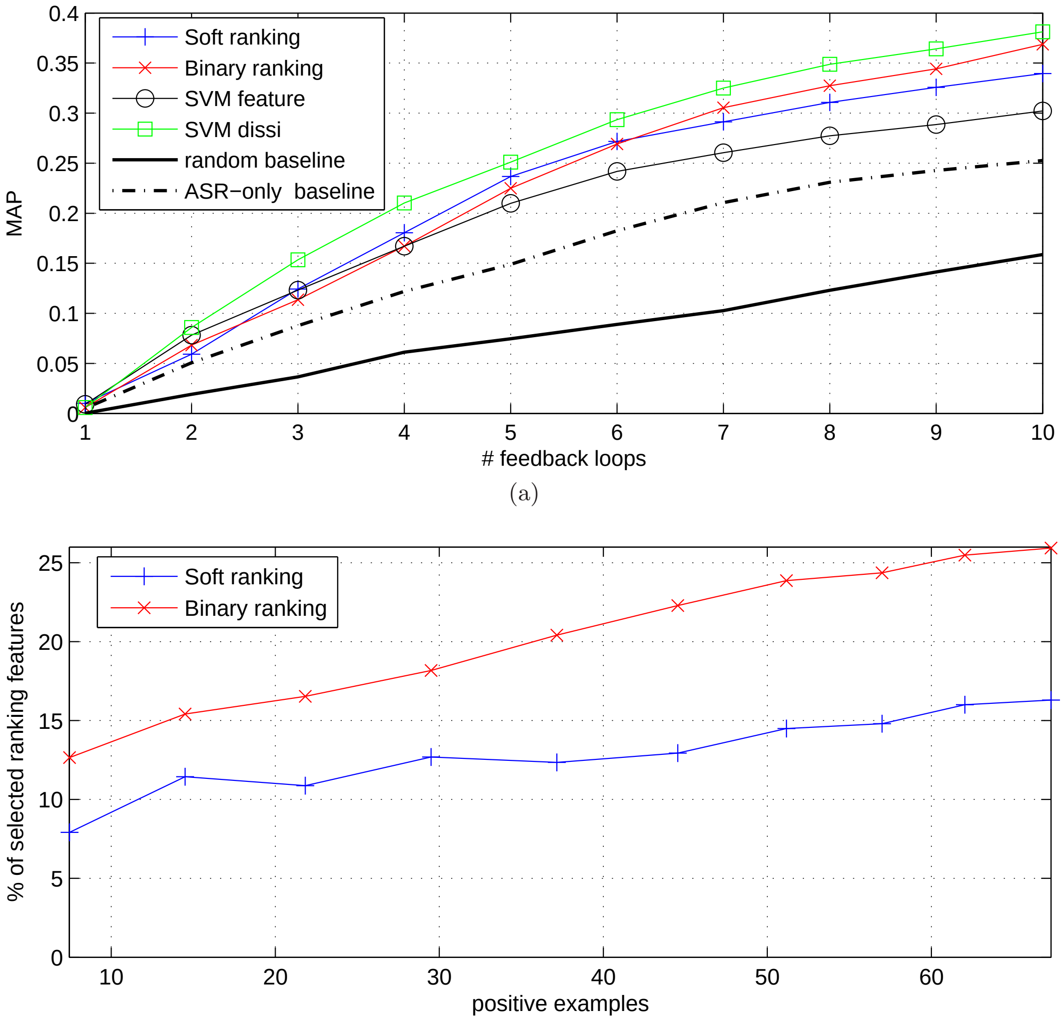
MAP results are given in Figure 7.a. We compare multimodal retrieval techniques with the best monomodal search (hierarchical SVM in $\mathcal{D}_{\mathcal{P}}^{\text{ASR}}$). In that case, the retrieval accuracy benefits from multimodal fusion. Whatever the fusion strategy we consider it definitely outperforms the ASR-only search. The overall RankBoost performance remains very close to the best retrieval result provided by the hierarchical SVM in dissimilarity space. Soft ranking and binary ranking have now similar performance and the latter is even slightly better when the training set becomes large. However, soft ranking systematically selects less features than binary ranking to produce the final ranking (Figure 7.b) and thus converges faster and provides simpler retrieval models. This is confirmed by the computational time (Table 1) as we observe that soft ranking is slightly faster than binary ranking. It is also interesting to note that RankBoost is around 20 times faster than the hierarchical SVM approaches.

Table 1: Computational time (in second, Intel Xeon 2.80GHz)

p+n	SVM in $\mathcal{F}$	SVM in $\mathcal{D}_{\mathcal{P}}$	binary rkg	soft rkg
20	0.46	0.36	0.009	0.01
30	0.80	0.68	0.028	0.024
60	2.95	2.75	0.17	0.12
100	9.43	9.23	0.56	0.52

7. CONCLUSION

The preference space we introduced in this paper is a very lightweight representation of the original feature space where all information relative to multimedia content is stored. Additionally to the above argument, preferences have the strong advantage to completely abstract multimodal content from dimensionality and scaling issues, and thus to facilitate fusion of heterogeneous descriptors. The challenge is then how to implement retrieval algorithms in preference space that



(a)

(b)

Figure 7: Multimodal video retrieval using a Relevance Feedback strategy with a) Mean Average Precision, and b) percentage of ranking features selected by RankBoost.

are as effective as techniques based on more traditional representations (*eg* feature space). The RankBoost algorithm offers us a very convenient solution, especially when considering the soft ranking function as a weak ranking. The performance is very close to state of the art SVM-based fusion algorithm operating in feature or dissimilarity spaces. The algorithm is parameter free and thus avoid any lengthy and hazardous parameters estimation. Finally, RankBoost is really fast compared to SVM-based approaches which is a crucial argument for online retrieval systems.

8. ACKNOWLEDGMENTS

This work is funded by the swiss Interactive Multimodal Information Management (IM2) project.

9. REFERENCES

[1] 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.

[2] Eric Bruno, Nicolas Moenne-Loccoz, and Stéphane 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.

[3] 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.

[4] Eric Bruno, Nicolas Moenne-Loccoz, and Stéphane Marchand-Maillet. Asymmetric learning and dissimilarity spaces for content-based retrieval. In *Proc. of International Conference on Image and Video Retrieval (CIVR)*, pages 330–339, Tempe, AZ, July 2006.

[5] 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.

[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] Y. Freund, Ra. Iyer, R. E. Schapire, and Y. Singer. An efficient boosting algorithm for combining preferences. *Journal of Machine Learning Research*, 4:933–969, November 2003.

[8] 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.

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

[10] 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.

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

[12] 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.

[13] 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.

[14] 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.

[15] 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.

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

[17] 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.

[18] C. Petersohn. Fraunhofer HHI at TRECVID 2004: Shot boundary detection system. In *TREC Video Retrieval Evaluation Online Proceedings*, 2004.

[19] 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.

[20] Robert E. Schapire and Yoram Singer. Improved boosting algorithms using confidence-rated predictions. *Journal of Machine Learning*, 37(3):297–336, 1999.

[21] A. F. Smeaton, P. Over, and W. Kraaij. Evaluation campaigns and trecvid. In *MIR '06: Proceedings of the 8th ACM international workshop on Multimedia information retrieval*, pages 321–330, New York, NY, USA, 2006. ACM Press.

[22] 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.

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

[24] 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.

[25] 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.

[26] 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.

[27] 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.

[28] 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.