Fisher Linear Discriminant Analysis for Text-Image Combination in Multimedia Information Retrieval

Christophe Moulin, Christine Largeron, Christophe Ducottet, Mathias Géry, Cécile Barat

Fisher Linear Discriminant Analysis for Text-Image Combination in Multimedia Information Retrieval

Christophe Moulin[a], Christine Largeron[a], Christophe Ducottet[a,*], Mathias Géry[a], Cécile Barat[a]

[a] *Université de Lyon, F-42023, Saint-Étienne, France ;*
CNRS, UMR 5516, Laboratoire Hubert Curien, F-42023, Saint-Étienne, France ;
Université de Saint-Étienne, Jean-Monnet, F-42023, Saint-Étienne, France.

Abstract

With multimedia information retrieval, combining different modalities - text, image, audio or video - provides additional information and generally improves the overall system performance. For this purpose, the linear combination method is presented as simple, flexible and effective. However, it requires to choose the weight assigned to each modality. This issue is still an open problem and is addressed in this paper.

Our approach, based on Fisher Linear Discriminant Analysis, aims to learn these weights for multimedia documents composed of text and images. Text and images are both represented with the classical bag-of-words model. Our method was tested over the ImageCLEF datasets 2008 and 2009. Results demonstrate that our combination approach not only outperforms the use of the single textual modality but provides a nearly optimal learning of the weights with an efficient computation. Moreover, it is pointed out that the method allows to combine more than two modalities without increasing the

*Corresponding author, Tel: +33 477915787, Fax: +33 477915781
Email address: ducottet@univ-st-etienne.fr (Christophe Ducottet)

complexity and thus the computing time.

Keywords: Multimedia information retrieval, Textual and visual information, Bag-of-words, Parameters learning, Fischer LDA

1. Introduction

The advent of digital cameras, video recorders, smart phones as well as the development of communication networks (e.g. WWW) has led to an explosion of the number of multimedia documents available. Users can easily create, mash and share some documents associating text, image, audio or video. This theoretically infinite amount of data creates a strong desire for efficient multimedia information retrieval systems able to search multimedia documents relevant to an information need. Otherwise, this data is not accessible and thus useless.

Most existing systems consider a single type of information for indexing and searching multimedia documents. Text Based Image Retrieval (TBIR) systems consider only the textual information (e.g. commercial search engines such as Google Images[1], Exalead Images[2] or systems specialized in images retrieval such as Picsearch[3], etc.), while Content Based Image Retrieval (CBIR) systems exploit only the visual content (e.g. [1], [2], QBIC [3], TinEye[4]). Among them, text-based search systems are very popular. They capitalize on the significant progress made in text retrieval. To retrieve documents, including those composed of videos or images, they index the

[1]Google Images: https://www.google.com/imghp

[2]Exalead Images: www.exalead.com/image

[3]Picsearch: http://www.picsearch.com/

[4]TinEye: http://www.tineye.com/

main text of documents plus the metadata like image name, tags, speech transcript, etc. Even though these systems are efficient, their performance is limited since they may ignore the different media content. Other retrieval approaches exploiting media content, called content-based approaches, have been actively studied to develop better image, video or sound/speech search engines as presented in several recent surveys [1, 4, 5, 6]. The question of how to represent the media content is central to these approaches. Most results from state-of-the-art methods in the different media domains are achieved with specific vocabularies and bag-of-Xwords representation, X standing for visual, audio or video [7, 8, 9, 10]. This model which is adopted from textual document representation processes image, video or audio data with textual information retrieval techniques and so, benefits from prior work in this field.

At present, it has been shown that using multimodal approaches yields better results than text-based systems or content-based systems, either in image, video or audio retrieval [11, 12, 13]. A multimodal approach inevitably implies combining diverse modality information, which is most often accomplished at the feature level (early fusion) or at the decision level (late fusion). It is important to determine an optimal fusion strategy to improve overall effectiveness. Many fusion strategies have been presented in the literature. Linear weighted fusion is one of the simplest and most widely used solutions [14, 15, 16]. However, weighting appropriately the different modalities remains an open problem.

In this paper, a new method based on Fisher-Linear Discriminant Analysis is presented, to learn automatically weights in a linear combination model for multimedia information retrieval. Our attention is restricted to multime-

dia documents containing only text and image modalities, that we model using a bag-of-words approach. Performance of our method is validated on multimedia information retrieval tasks in the imageCLEF challenge.

The organization of the rest of paper is as follows. Section 2 discusses some work related to multimedia fusion. Section 3 presents our prior work on multimedia information retrieval and highlights the remaining key issues. Section 4 describes our new approach. Section 5 presents an experimental protocol applied to get the results of Section 6. Finally in section 7 we present our conclusions and prospects for future work.

2. Related work

This paper addresses the problem of combining multiple modalities, especially text and image, to increase the effectiveness of multimodal retrieval system. In this section, we briefly present some background information on multimodal fusion and we introduce some related work to provide the context under which our method was developed. See [13] for a more complete overview of current approaches on multimodal fusion.

Two data fusion strategies are often in opposition: early and late fusion. Early fusion combines the different unimodal features into a single representation (Figure 1). A simple early fusion approach is to normalize and concatenate features into a unique vector. This has been extensively applied for combining texture, color, and shape information in many image applications, including multimodal biometrics [17, 18], face recognition [19], image annotation [20], image classification [21] or retrieval [22]. This simple approach suffers from several limitations, including the *curse of dimensionality*, the

data redundancy which may lead to a decrease in the system performance, incompatibility between feature ranges and types. Dimensionality reduction methods as Principal Component Analysis (PCA) [23] and appropriate normalization are used to solve these problems, as well as more elaborated methods as discussed in [19].

Late fusion processes each modality independently and fusing the results arising from all systems (Figure 1). Late fusion approaches are diverse. They are broadly categorized into two methods: similarity scores methods and rank-based methods. Similarity scores methods exploit the similarity value, e.g. the score, between a query and each document. They assign a final relevance score to a document based on all returned relevance scores from the different retrieval systems. In addition, a normalization step may be required to compare the returned relevance scores.

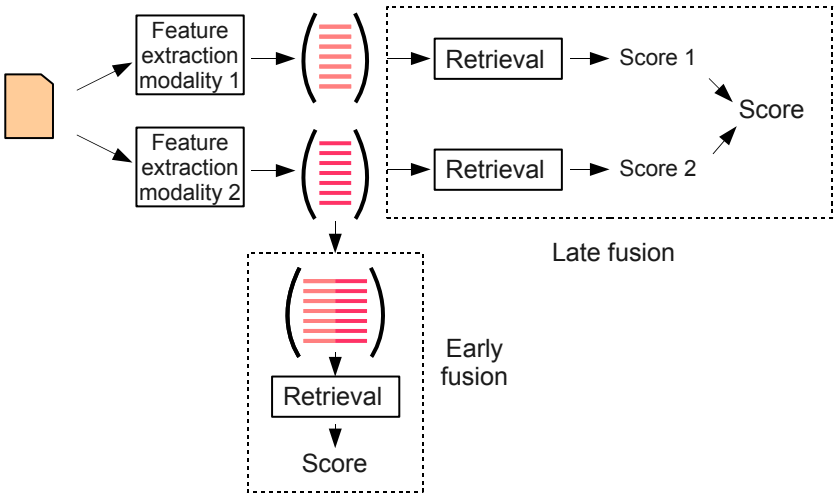

Figure 1: Early and late fusion strategies.

In general, linear combination models are simple and effective methods to

fuse information. In these methods, the final score is obtained as a weighted sum of the scores from each unimodal system. Several formulas have been suggested, including CombSum [15], CombMNZ [15] or standard linear formulations [14, 16, 24]. For instance, the CombSum approach computes the sum of all returned scores, while CombMNZ weighs the CombSum score by the number of systems which have retrieved a document. Since scores are not always available, a second type of late fusion approaches exploit the rank of retrieved documents to improve the retrieval [25]. For example, Borda count and Condorcet methods used by Aslam and Montague in [26, 27] are two well-known solutions.

Many papers evaluate early and late fusion strategies by comparing their performances and investigating their pros and cons [28, 21]. When dealing with mixed feature types, such as text, video or image, late fusion appears to perform better than early fusion. It has the advantage to use finely tuned retrieval models specific to each modality. Hereafter we will focus on this kind of approach and, more specifically, on the linear combination model.

This linear combination model has been widely used in multimedia information retrieval for combining audio/visual features [29], audio/video features [30, 31] or text/visual features [32, 33, 12]. In all these publications, combining features always leads to improve the retrieval system results. However, obtaining a suitable combination is not straightforward. The principal difficulty is to determine the weights of the different modalities. In some papers, the authors consider equal weights [34, 35]. This is evidently the most simple approach, but it does not take into account the strengths and weaknesses of each unimodal retrieval system. Therefore, it would appear

pertinent to attribute more or less weight to a system according to its reliability and performance. To accomplish this, some approaches choose to fix values for the different weights and vary them to study their influence on the system [32, 33]. However, in these latter works, no care is taken to use a specific dataset to learn the parameters and another one to evaluate the system. Consequently, the results can be overestimated.

Other works evaluate the reliability of each system and decide weights accordingly [29, 31]. Moreover, some learning approaches to assign weights are proposed. Then, the combination model, or global retrieval system, is first optimized with training data, and then the optimized model is applied on test data [30, 12]. A standard learning approach is to consider a criterion which measures the performance of the combination model and to numerically optimize this criterion to find the optimal weight values which provide the best outcomes. In [12], to combine text and visual information, some authors apply an exhaustive search of the parameter space in range [0,1] with the training data. Similarly, in [30], to combine audio, visual and synchrony features, other authors perform a grid search in ranges [0,1] to determine the two combination parameters. Such approaches are very time consuming, especially as the number of features increases. From all these publications, the issue of finding the appropriate weights for different modalities remains clearly one of the major drawbacks of the linear combination method and is presented as "an open research issue" in [13]. Our aim is to introduce a solution to learn weights using Fisher Linear Discriminant Analysis [36] for multimedia documents composed of text and images.

3. Textual and visual information retrieval model

In previous work, we developed an Information Retrieval model in order to exploit textual and visual information [12, 37]. This model was based on late fusion and linearly combines textual and visual scores. It is made up of several modules as illustrated in Figure 2. The first stage (1), consists in

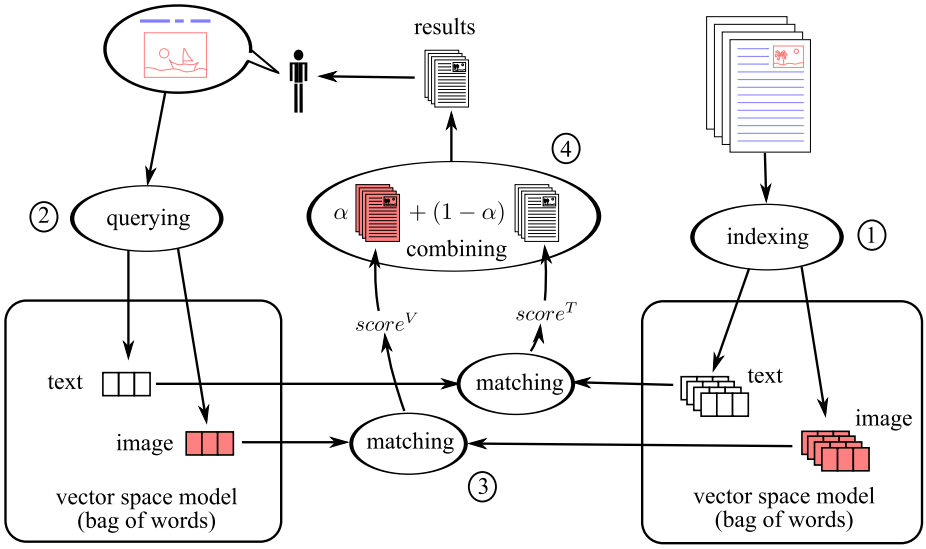

Figure 2: Multimedia information retrieval model

indexing textual and visual information, using a bag-of-words based model (e.g. vector space model). At this step, text and image information are processed independently. During the second stage (2), a textual and visual information query is provided by a user. It is also represented using the same bag-of-words approach. Then, the third stage (3) computes a ranking score between the query and every document of the collection, for each modality (text and image). Finally, the last step (4) combines linearly these textual and visual scores in order to retrieve the most relevant documents

corresponding to the user's need. In this model, the weight corresponding to each kind of information depends on a parameter α.

3.1. Textual representation

Given $\mathcal{D}$, a collection of documents and $T = \{t_1, \ldots, t_j, \ldots, t_{|T|}\}$, an index composed of terms occurring in $\mathcal{D}$, a document d_i is represented as a vector of weights $\vec{d_i} = (w_{i,1}, \ldots, w_{i,j}, \ldots, w_{i,|T|})$ according to the vector space model introduced by Salton et al. [38, 39]. In their model, the importance of a term t_j within the specific document d_i is measured by the term frequency $tf_{i,j}$ while its importance over the corpus is evaluated with the inverse document frequency idf_j. The weight $w_{i,j}$ corresponds to the product of $tf_{i,j}$ by idf_j where $tf_{i,j}$ and idf_j can be computed according to the version of Okapi formula [40] as implemented by the Lemur software [41]:

$$tf_{i,j} = \frac{k_1 \, n_{i,j}}{n_{i,j} + k_1(1 - b + b\frac{|d_i|}{d_{avg}})} \tag{1}$$

where $n_{i,j}$ is the number of occurrences of the term t_j in the document d_i, $|d_i|$ the size of the document d_i, d_{avg} the average size of all documents in the corpus and, k_1 and b are two constants,

$$idf_j = \log \frac{|\mathcal{D}| + 1}{|\mathcal{D}_j| + 0.5} \tag{2}$$

where $|\mathcal{D}|$ is the size of the corpus and $|\mathcal{D}_j|$ the number of documents of $\mathcal{D}$ where the term t_j occurs at least one time.

A query q_k, provided by a user, can also be considered as a short document, and therefore, it can also be represented as a vector of weights. A

score is then computed between the query q_k and a document d_i:

$$score_T(q_k, d_i) = \sum_{t_j \in q_k} tf_{k,j}\, idf_j\; tf_{i,j}\, idf_j \tag{3}$$

3.2. Visual representation

In order to combine the visual information with textual information, we also represent images as weighted vectors. This requires a visual vocabulary $V = \{v_1, \ldots, v_j, \ldots, v_{|V|}\}$ defined, in two steps, using a bag of visual words approach [7]. First, each image is partitioned into a regular grid of 16×16 cells with a minimum of 8×8 pixels for each cell. This uniform partitioning requires low computational complexity. Next, local features are extracted from each cell to describe its texture properties and color. To this end, we choose to compute two different descriptions: $mstd$ and $sift$ (*Scale-Invariant Feature Transform*).

The color description, $mstd$, calculates 6 features equal to the mean and the standard deviation of each normalized color components defined by $R/(R+G+B)$, $G/(R+G+B)$ and $(R+G+B)/(3 \times 255)$ where R, G and B are the red, green and blue components of one pixel. The $sift$ description converts each cell into a 128-dimensional vector which represents the texture information of the image [42, 43].

In a second step and for each description ($mstd$ and $sift$), clustering based on the k-means algorithm is performed over all the cells descriptions to obtain k clusters of features. The center of each cluster corresponds to what we refer to as a visual word. Thus, each description defines a vocabulary (V_{mstd} and V_{sift}), the size of which corresponds to the number of clusters.

By analogy with the bag of textual words, each visual vocabulary can

be used to represent an image, belonging to a document or a query, as a vector of visual terms. This image is decomposed into a 16×16 grid and the local $mstd$ or $sift$ features are computed. The description of each cell is then assigned to the closest visual word using the Euclidean distance and the image is finally represented by a vector of $tf.idf$ weights computed the same way as for text.

Finally, a visual score $score_V(q_k, d_i)$, corresponding respectively to a visual vocabulary V (V_{mstd} or V_{sift}), is calculated between a query q_k and an image document d_i:

$$score_V(q_k, d_i) = \sum_{v_j \in q_k} tf_{k,j} idf_j \; tf_{i,j} idf_j \tag{4}$$

3.3. Textual and visual combination

Using two vocabularies, a textual one T and a visual one V (V_{mstd} or V_{sift}), a global score for a document d_i and a query q_k is defined as a linear combination of the scores corresponding to each modality:

$$score(q_k, d_i) = \alpha \, score_V(q_k, d_i) + (1 - \alpha) \, score_T(q_k, d_i) \tag{5}$$

The parameter α permits to add more or less visual information in the overall score used for the ranking of documents.

In the case where more than one visual vocabulary is considered, the final score computation can be generalized as follows:

$$score(q_k, d_i) = \sum_{j=1,\ldots,|\mathcal{M}|} \alpha_j \, score_j(q_k, d_i) \tag{6}$$

where $\mathcal{M} = \{V_j, j = 1, \ldots, |\mathcal{M}|\}$ denotes a set of multimedia vocabularies containing typically several visual vocabularies and a textual one (e.g. $|\mathcal{M}| =$

3 and $\mathcal{M} = \{V_{mstd}, V_{sift}, T\}$) and α_j corresponds to the fusion parameter associated to vocabulary V_j. In the former two vocabulary case, we had $\alpha_1 = \alpha$ and $\alpha_2 = 1 - \alpha$.

3.4. Learning combination parameters by optimization

It is obvious that the choice of the fusion parameters is very sensitive especially to combine descriptions of different nature (e.g. textual and visual). It seems that the weight assigned to the text and to the visual information should not be the same because the effectiveness of the model based only on a text descriptor is usually better than those based only on a visual descriptor [44, 45]. However, it is not easy to set these parameters, even for an expert.

In order to solve this problem, we presented in a previous work [12, 37], a method learning the values of the parameters using a set of queries and the corresponding list of relevant documents. This set is divided into training and test sets. Given an evaluation criterion, lets take for instance the Mean Average Precision (MAP) [46], the method consists in searching for the combination parameter that optimizes this criterion on the training set of queries. With the obtained value, the effectiveness of the model is then evaluated on the test set of queries.

More precisely, if we consider two descriptions: T based on the text and V based on visual information and if MAP_α denotes the MAP obtained on the training set of queries with the value α for the combination parameter, then the optimal value α^* is given by:

$$\alpha^* = \underset{\alpha \in [0,1]}{\arg\max}\ MAP_\alpha \tag{7}$$

We can note that this method is based on the same principle as those

detailed in [32, 33] except that it avoids the risk of overestimation. Moreover it generalizes the approach described in [34, 35] as it considers not only equal weights but also different weights. For this reason, it can be considered as the state of the art approach to find the best weighting coefficients given a learning dataset. We will further use it in our experiments to evaluate the method introduced in this article over state of art linear combination fusion methods.

3.5. Discussion

Although this MAP optimization method has provided good results [12, 37], it has several drawbacks. Firstly, as the evaluation criterion is not linear in function of the parameter, the optimal parameter can not be calculated analytically. We must therefore use a numerical optimization method such as exhaustive search, gradient descent or Newton's method [47]. Moreover, depending on the optimization method used, the convergence to the global maximum is not necessarily guaranteed especially when the method is applied to a larger number of descriptors. Finally, the main disadvantage of this approach is that its computational complexity is high. More precisely, given $|\mathcal{M}|$ modalities, $|\mathcal{D}|$ documents and $|\mathcal{Q}|$ queries in the training set, we can first pre-compute $|\mathcal{Q}| \times |\mathcal{D}|$ query-document similarities for each modality. Then, for each iteration of the optimization algorithm, we must evaluate the MAP associated to a given set of parameters which requires a complexity of $O(|\mathcal{D}|^2 log(|\mathcal{D}|))$ (sorting and MAP computation). The number of iterations depends on the precision required for the parameters. If n is the number of digits, the number of iterations is polynomial (of order $|\mathcal{M}|$) on n for a grid search.

3.6. Remaining problems

The textual and visual information retrieval model presented in this section can combine a textual modality with one or more visual ones. The optimization method introduced previously to learn the value of the parameters is computational very expensive, especially in the case where several visual modalities are considered. Thus, in the next section, we propose a new approach to learn more efficiently the values of the combination parameters.

4. Learning combination parameters by Fisher Linear Discriminant Analysis

To learn efficiently the combination parameters of the Information Retrieval model, we decided to use the Fisher Linear Discriminant Analysis (Fisher-LDA). For that, we first reformulate the learning of the combination parameters as a dimensionality reduction problem in a binary classification context: find the linear combination which best separate relevant and non-relevant documents for all the queries. This latter problem can be solved using the Fisher-LDA which enables us to derive an analytical method to learn any number of combination parameters for the information retrieval model. Please note that Fisher-LDA is not used to build a classifier but only to find analytically the linear combination, (i.e. the combination parameters) which best separates relevant documents from irrelevant ones.

4.1. Reformulation of the learning problem

In our learning problem, each document may be relevant or not relevant with respect to a query. We can thus define a two class problem where objects

to classify are couples of document-query and the two classes are relevant or non-relevant. Moreover, each object can be described by a vector of variables corresponding to the scores calculated for each considered description.

More formally, considering the set of documents $\mathcal{D}$ and the set of queries $\mathcal{Q}$, the set of objects $\mathcal{X}$ is defined as the set $\mathcal{D} \times \mathcal{Q}$ of all the document-query couples x_ℓ:

$$\mathcal{X} = \{(x_\ell)_{\ell=1,\ldots,|\mathcal{X}|} \ , \ x_\ell = (d_i, q_k)_{i=1,\ldots,|\mathcal{D}|,\, k=1,\ldots,|\mathcal{Q}|}\} \tag{8}$$

Each of the $|\mathcal{X}| = |\mathcal{D}| \times |\mathcal{Q}|$ objects can belong to the set $\mathcal{X}_R$ of relevant objects or to the set $\mathcal{X}_{\overline{R}}$ of non-relevant objects depending on whether the document is relevant or not for the query:

$$\mathcal{X}_R = \{x_\ell \in \mathcal{X} \mid d_i \text{ is relevant for } q_k\} \tag{9}$$

$$\mathcal{X}_{\overline{R}} = \{x_\ell \in \mathcal{X} \mid d_i \text{ is not relevant for } q_k\} \tag{10}$$

In multimedia information retrieval, each object x_ℓ is naturally represented by a vector of variables $\mathbf{x}_\ell$ whose components correspond to the scores, for each vocabulary V_j in $\mathcal{M}$, between document d_i and query q_k:

$$\mathbf{x}_\ell = (x_{\ell,j})_{j=1,\ldots,|\mathcal{M}|} = (score_j(q_k, d_i))_{j=1,\ldots,|\mathcal{M}|} \tag{11}$$

where $\mathcal{M}$ denotes a set of multimedia vocabularies containing typically several visual vocabularies and a textual one, like for instance $|\mathcal{M}| = 3$ and $\mathcal{M} = \{V_{mstd}, V_{sift}, T\}$.

4.2. Resolution by Fisher-LDA

Each object $x \in \mathcal{X}$ can belong to one of the two classes and it is represented by the vector of scores $\mathbf{x} = (x_j)_{j=1,\ldots,|M|}$ corresponding to each vocabulary. Our aim is to determine a linear combination of these scores which best

separate the two classes. This problem is equivalent to finding a factor axis which best separates the two populations, considering the class membership of objects. The canonical discriminant analysis provides a solution to this problem by minimizing the Fisher linear discriminant [48, 49, 50]. Compared to the Principal Component Analysis (PCA) [51], the advantage of Fisher-LDA is that it takes into account the class membership of objects. Note that within the framework of linear discriminant analysis, Fisher's discriminant can also be used to define an optimal Bayesian classifier under assumptions of normally distributed classes and equal covariances [49]. However, these more restrictive hypotheses are not required if a canonical analysis is considered.

Given a score vector $\mathbf{x} = (x_j)_{j=1,\ldots,|\mathcal{M}|}$ and the coefficient vector $\mathbf{z} = (\alpha_j)_{j=1,\ldots,|\mathcal{M}|}$, the discriminant function corresponding to the linear combination is given by:

$$z = {}^t\mathbf{z}\mathbf{x} = \sum_{j=1,\ldots,|\mathcal{M}|} \alpha_j x_j \tag{12}$$

Note that with this formulation, variable z is exactly the score corresponding to a query-document couple given in equation 6 and $\mathbf{z} = (\alpha_j)_{j=1,\ldots,|\mathcal{M}|}$ are the combination coefficients we are looking for.

We can verify that the variance $V(z)$ of variable z is equal to $V(z) = {}^t\mathbf{z}\mathbf{T}\mathbf{z}$ where $\mathbf{T}$ is the covariance matrix associated to the scores. Using Huygens theorem, this matrix can be decomposed into an *within class* covariance matrix $\mathbf{W}$ and a *between class* covariance matrix $\mathbf{B}$. The three $|\mathcal{M}| \times |\mathcal{M}|$

matrices are defined by:

$$\mathbf{T} = \frac{1}{|\mathcal{X}|} \sum_{\ell=1}^{|\mathcal{X}|} (\mathbf{x}_\ell - \mu)^t (\mathbf{x}_\ell - \mu) \tag{13}$$

$$\mathbf{B} = \frac{1}{|\mathcal{X}|} \left(|\mathcal{X}_R| (\mu_{\mathbf{R}} - \mu)^t (\mu_{\mathbf{R}} - \mu) + |\mathcal{X}_{\overline{R}}| (\mu_{\overline{\mathbf{R}}} - \mu)^t (\mu_{\overline{\mathbf{R}}} - \mu) \right) \tag{14}$$

$$\begin{aligned} \mathbf{W} = & \frac{1}{|\mathcal{X}|} \sum_{x_\ell \in \mathcal{X}_R} (\mathbf{x}_\ell - \mu_{\mathbf{R}})^t (\mathbf{x}_\ell - \mu_{\mathbf{R}}) \\ & + \frac{1}{|\mathcal{X}|} \sum_{x_\ell \in \mathcal{X}_{\overline{R}}} (\mathbf{x}_\ell - \mu_{\overline{\mathbf{R}}})^t (\mathbf{x}_\ell - \mu_{\overline{\mathbf{R}}}) \end{aligned} \tag{15}$$

where μ, $\mu_{\mathbf{R}}$ and $\mu_{\overline{\mathbf{R}}}$ denote the mean data vectors computed respectively over all the set $\mathcal{X}$, over the set $\mathcal{X}_R$ of relevant documents or over the set $\mathcal{X}_{\overline{R}}$ of non relevant documents. They are defined as:

$$\mu = \frac{1}{|\mathcal{X}|} \sum_{\ell=1}^{|\mathcal{X}|} \mathbf{x}_\ell \tag{16}$$

$$\mu_{\mathbf{R}} = \frac{1}{|\mathcal{X}_R|} \sum_{x_\ell \in \mathcal{X}_R} \mathbf{x}_\ell \tag{17}$$

$$\mu_{\overline{\mathbf{R}}} = \frac{1}{|\mathcal{X}_{\overline{R}}|} \sum_{x_\ell \in \mathcal{X}_{\overline{R}}} \mathbf{x}_\ell \tag{18}$$

According to Fisher-LDA, the optimal discriminant function z can be obtained by the maximization of Fisher criterion $F(\mathbf{z})$ defined by:

$$F(\mathbf{z}) = \frac{{}^t\mathbf{z}\mathbf{B}\mathbf{z}}{{}^t\mathbf{z}\mathbf{T}\mathbf{z}} \tag{19}$$

It can be shown that the solution is obtained by calculating the first eigenvector of matrix $\mathbf{T}^{-1}\mathbf{B}$ [52].

In the particular case of a two-class problem, the eigenvector (and thus the combination coefficients) are obtained by the following analytical formula:

$$\mathbf{z} = (\alpha_j)_{j=1,\ldots,|\mathcal{M}|} = \mathbf{T}^{-1}(\mu_{\mathbf{R}} - \mu_{\overline{\mathbf{R}}}) \tag{20}$$

4.3. Learning the combination parameters

The learning strategy is similar to that introduced in section 3.4. The combination parameters are first calculated by Fisher-LDA using a training set of queries. Then, the information retrieval model is evaluated with these parameters on a test set of queries. In the training stage, the calculation of the combination parameter is made analytically, firstly by estimating the covariance matrix $\mathbf{T}$ (equation 13) and the mean vectors $\mu_{\mathbf{R}}$ (equation 17) and $\mu_{\overline{\mathbf{R}}}$ (equation 18), secondly by calculating the eigenvector $\mathbf{z}$ (equation 20) which is, by definition, the desired combination parameter vector.

4.4. Utilization of the decision criterion

The aim of the learning step is just to provide an optimal set of combination parameters which are computed according to formula 20. Then, in the test step, these combination parameters are used to process the test queries following the information retrieval model presented in section 3. Firstly, a query provided by a user, is also represented using the bag-of-words model and a document-query score is computed independently for each modality (text and image). Secondly, a global score between the query and each document is computed according to formula 6 using the combination parameters determined in the learning step. Finally the documents are ranked according to this global score and they are returned by the system. The documents with highest scores are considered by the system as the most relevant for the query.

4.5. Discussion

It's interesting to note that the two learning methods (MAP optimization and Fisher-LDA) use the maximization of a specific criterion for a training set of queries. In the first one the criterion is the MAP whereas for the new one it is Fisher criterion. Obviously, the two approaches will lead to diverse combination parameters and thus to different evaluation results. If the evaluation measure of the Information Retrieval (IR) system is also the MAP, the first method is expected to give better results.

However, one great advantage of the new method is its efficiency and generality: the combination parameters are obtained analytically from the mean vectors and covariance matrix of data for any number of descriptions. Its computational cost is low compared to the MAP optimization: it also requires to pre-compute $|\mathcal{M}|\times|\mathcal{Q}|\times|\mathcal{D}|$ query-document similarities but then, the parameters value estimation only requires the computation of the inverse of the covariance matrix (complexity $O(|\mathcal{M}|^6)$, with a small $|\mathcal{M}|$ value) which is independent on the number of documents and on the precision required for the combination parameters.

5. Experiments

In order to evaluate our new Fisher LDA learning method, we used the IR test collection ImageCLEF[5]. Before presenting the results in section 6, we first describe in this section the ImageCLEF dataset, the system settings and the various experiments that have been made.

[5] `http://www.imageclef.org`

5.1. Dataset

The ImageCLEFwiki collection was employed for the competition ImageCLEF 2008 and 2009 [53, 54]. It is one of the few large image retrieval collections with a significant text part. Moreover, the ground truth is available for the two sets of queries proposed in editions 2008 and 2009 of the competition.

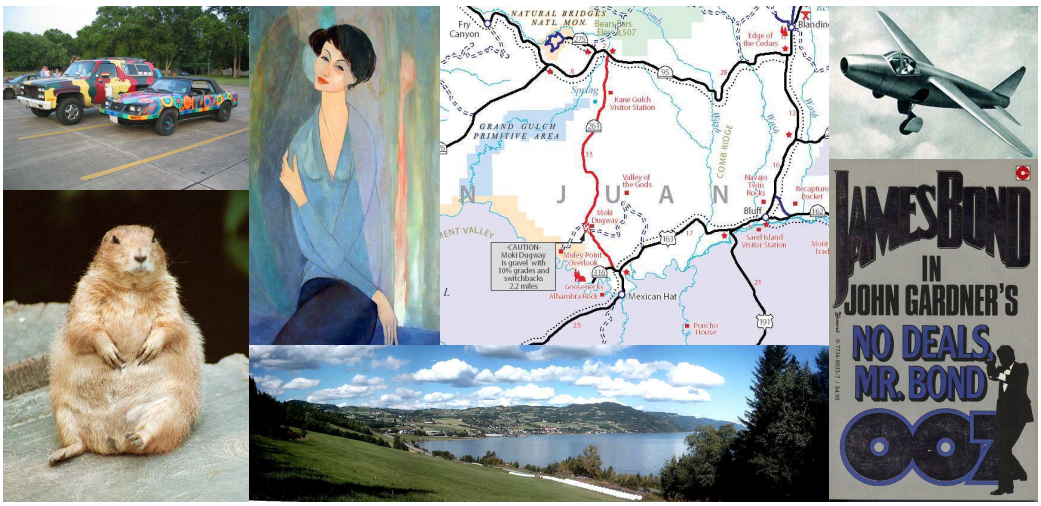

Figure 3: Images excerpted from ImageCLEF collection.

This collection is composed of 151 519 multimedia XML documents extracted from Wikipedia. The documents are made up of an image and a short text. Images have heterogeneous sizes and depict either photos, drawings or screenshots (Figure 3). The textual part of a document is unstructured and consists of a description of the image, information about the Wikipedia user who has uploaded the image, or the copyright of the image. The average number of words per image is about 33.

In 2008, the ImageCLEF collection was provided with 75 queries. All

queries are not provided with a visual part. Thus, in order to have, for each query, the visual information obtained similarly, we have selected as visual component the two first images ranked by a preliminary textual querying step. These 75 queries will correspond to a training collection for learning the combination parameters.

In ImageCLEF 2009, all the 45 given queries were multimedia. The textual part is composed of few words and the visual part corresponds on average to 1.84 images per query. This second set of queries will be used as a testing collection. Information about the ImageCLEF collection is summarized on table 1.

	2008	2009
Number of documents	151 519	
Mean size of documents	33	
Number of queries	75	45
Mean number of image queries	1.97	1.84
Mean size of textual queries	2.64	2.93

Table 1: Collection ImageCLEF 2008 et 2009.

5.2. System settings

The lemur software was used with the default parameters as defined in [41]. The k_1 parameter of BM25 formula is set to 1. As $|d_k|$ and d_{avg} are not defined for a query q_k, b is set to 0 for the $tf_{k,j}$ computation. When $tf_{i,j}$ is estimated for a document d_i and a term t_j, this parameter b is set to 0.5. Moreover, stop-words have not been removed and the Porter stemming have

been applied. The number of visual words, corresponding to the parameter k of the k-means, has been empirically set to 10 000 for both $mstd$ and $sift$ descriptions.

5.3. Experiments

5.3.1. Evaluation criteria and baseline

In order to evaluate the improvements of the combination of different document descriptions, we firstly run experiments considering only one document description, either textual or visual. The best result obtained will correspond to our baseline.

Only queries of the ImageCLEF 2009 collection are considered. We will use two different evaluation measures: R and MAP (cf. Appendix A). R corresponds to the recall and is obtained by dividing the number of relevant retrieved documents by the number of relevant documents to retrieve. The MAP is the mean average precision which is a common criteria used for example to rank participants in ImageCLEF competition [53, 54].

5.3.2. Comparison between the two learning methods

Considering the combination of two descriptions, the goal of this experiment is to compare the new Fisher-LDA method over other linear combination approaches. For that purpose, we use the MAP optimization method introduced in section 3.4 as a reference. Indeed, as explained previously, it can be considered as the state of the art approach to find the best weighting coefficients given a learning dataset.

The two descriptions used are a textual T and a visual (either V_{mstd} or V_{sift}). For both approaches, we learn the parameters on a training set of

queries corresponding to the ImageCLEF 2008 competition. Queries of the ImageCLEF 2009 competition are then used to evaluate the results with the R and MAP measures. The results can be compared to the baseline to verify if the Fisher LDA method is as effective as the MAP optimization approach.

5.3.3. Combining three descriptions with Fisher-LDA

The Fisher LDA method is more efficient than the MAP optimization and can be used to calculate combination parameters for more than two descriptions. Thus, we will then combine T, V_{mstd} and V_{sift} descriptions using the Fisher LDA learning.

6. Results

6.1. Baseline

The first experiments exploit only one textual or visual modality. The results are summarized in table 2. According to the MAP measure, the visual information leads to poor results whatever the description used: for $mstd$, $MAP = 0.0071$ and for $sift$, $MAP = 0.0083$. Moreover, only about 10% of the relevant documents have been retrieved when using only the visual information. As for the results when only the textual information is utilized, the MAP reaches 0.1661, and 73% of the relevant documents are retrieved. The textual results are the best results using only one modality. Thus, we will consider these results as our baseline for comparing the runs combining textual and visual information.

Run	Modalities	MAP	R
Baseline	T	**0.1661**	**0.7336**
	V_{mstd}	0.0071	0.0721
	V_{sift}	0.0083	0.1078

Table 2: Results on the ImageCLEF 2009 collection exploiting textual and visual information separately.

6.2. Comparison between the two learning methods

Tables 3 and 4 illustrate results combining one textual and one visual description using respectively the MAP optimization and the Fisher LDA learning. Whatever the combined modalities, results are improved compared to the textual baseline. Thus, it confirms the benefit of combining different descriptions of multimedia documents.

For the MAP optimization method, Table 3 shows that the optimal α^* parameters, calculated in the training step, differ depending on the description used ($\alpha^*_{V_{mstd}} = 0.034$ and $\alpha^*_{V_{sift}} = 0.077$). Using these parameters in the test step led to a MAP of 0.1791 (respectively 0.1813) when combining the textual information with the $mstd$ (respectively $sift$) description. Compared to the baseline, the MAP is improved while the number of relevant retrieved documents is approximately the same. Thus, it can be concluded that documents are better ranked if a combination of the textual and the visual information is used.

Let's also note that equal weights corresponding to $\alpha = 0.5$ which have been chosen in other studies [34, 35] provide poor results even lower than text only ones. For that reason, they are not presented here.

Run	Modalities	MAP	R
Baseline	T	0.1661	0.7336
MAP optimization	T+V_{mstd} : $\alpha^*_{V_{mstd}}$: 0.034	0.1791	0.7367
	T+V_{sift} : $\alpha^*_{V_{sift}}$: 0.077	**0.1813**	**0.7478**

Table 3: Combination results obtained with the MAP optimization method on ImageCLEF 2009 collection.

For the Fisher LDA learning, Table 4 shows that the MAP is increased by 0.1661 to 0.1801 for the $mstd$ description and to 0.1795 for the $sift$ description. Compared to the MAP optimization, these results are very similar: on the one hand, the MAP is slightly better for the $mstd$ description (from 0.1791 to 0.1801) and on the other hand, it is a little worse for the $sift$ description (from 0.1813 to 0.1795). However, notice that for both descriptions, the Fisher LDA learning always leads to a better recall than the one obtained with the MAP optimization approach. Indeed, the recall is 75.09% (respectively 73.67%) for the $mstd$ description and 75.15% (respectively 74.78%) for the $sift$ description using the Fisher LDA learning (respectively MAP optimization approach).

Statistically, the MAP obtained after Fisher LDA learning is significantly improved with a p-value of 0.002 (respectively 0.008) for the combination of T and V_{mstd} (respectively T and V_{sift}) according to the paired Wilcoxon signed-rank test.

Moreover, if we compute the MAP criteria for α varying from 0 to 0.2 using the ImageCLEF 2009 queries, we can compare the learnt Fisher LDA α parameters to optimals as shown by the curves of Figure 4. These curves

illustrate that both $\alpha_{V_{mstd}}$ and $\alpha_{V_{sift}}$ are closed to the optimal results obtained by combining textual and visual information on ImageCLEF 2009.

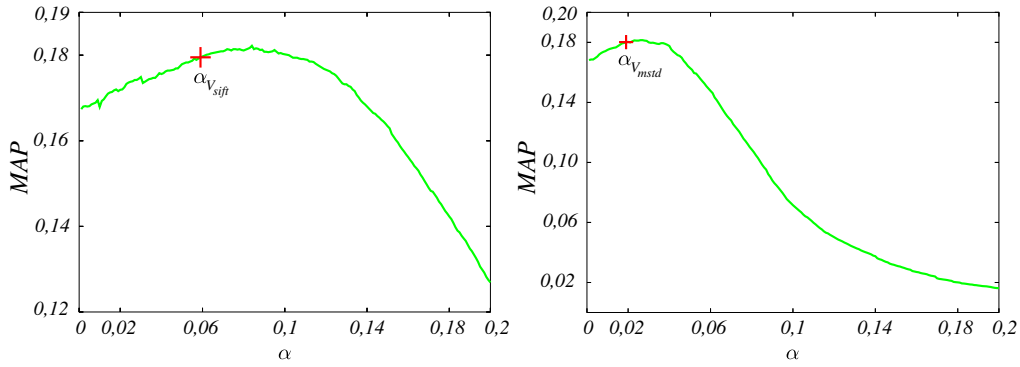

Figure 4: MAP get with ImageCLEF 2009 queries with α varying from 0 to 0.2 for $sift$ (on the left) and $mstd$ (on the right) descriptors.

Run	Modalities	MAP	R
Baseline	T	0.1661	0.7336
Fisher LDA learning	T et V_{mstd} : $\alpha_{V_{mstd}}$: 0.018730	**0.1801**	0.7509
	T et V_{sift} : $\alpha_{V_{sift}}$: 0.059098	0.1795	**0.7515**

Table 4: Combination results obtained with the Fisher LDA learning on the ImageCLEF 2009 collection.

6.3. Combining three descriptions with Fisher-LDA

Table 5 illustrates results combining T, V_{sift} and V_{mstd} modalities using the Fisher LDA learning. Both MAP measure ($MAP = 0.1875$) and number of relevant retrieved documents ($R = 0.7614$) are improved combining three modalities. This MAP improvement is significant with a p-value equals to 0.0003.

Run	Modalities	MAP	R
Baseline	T	0.1661	0.7336
Fisher LDA learning	T, $V_{mstd} : \alpha_{V_{mstd}} : 0.013837$ et $V_{sift} : \alpha_{V_{sift}} : 0.044451$	**0.1875**	**0.7614**

Table 5: Fisher LDA learning results on the ImageCLEF 2009 collection combining T, V_{sift} and V_{mstd} modalities.

On figure 6, we have estimated, for the 45 queries, the relative difference between the MAP obtained combining the three descriptions and the one attained with the text only. For about a quarter of queries, results are degraded by the combination of all descriptions. The worst difference (about -50%) is reached for the query *building site*. However, for most of the queries, the combination leads to an improvement. For half of the queries the difference is higher than 10% and the best improvement is higher than 150% for the query *notes on music sheet*. For this query, the visual information is intuitively important as for the next best difference queries which are: *traffic signs, earth from space* and *red fruit*. Figure 5 presents some examples of relevant images for the queries previously mentioned.

7. Conclusion and future work

In this paper, we addressed the problem of combining textual and visual information for multimedia information retrieval. Our approach was based on the representation of documents as *tf.idf* weighted vectors corresponding to several textual and visual vocabularies. Relying on a linear combination of scores from each vocabulary, two methods for learning the combination

Figure 5: The first fifth results for queries *red fruit* and *notes on music sheet* using textual information and textual and visual information using the Fisher LDA learning.

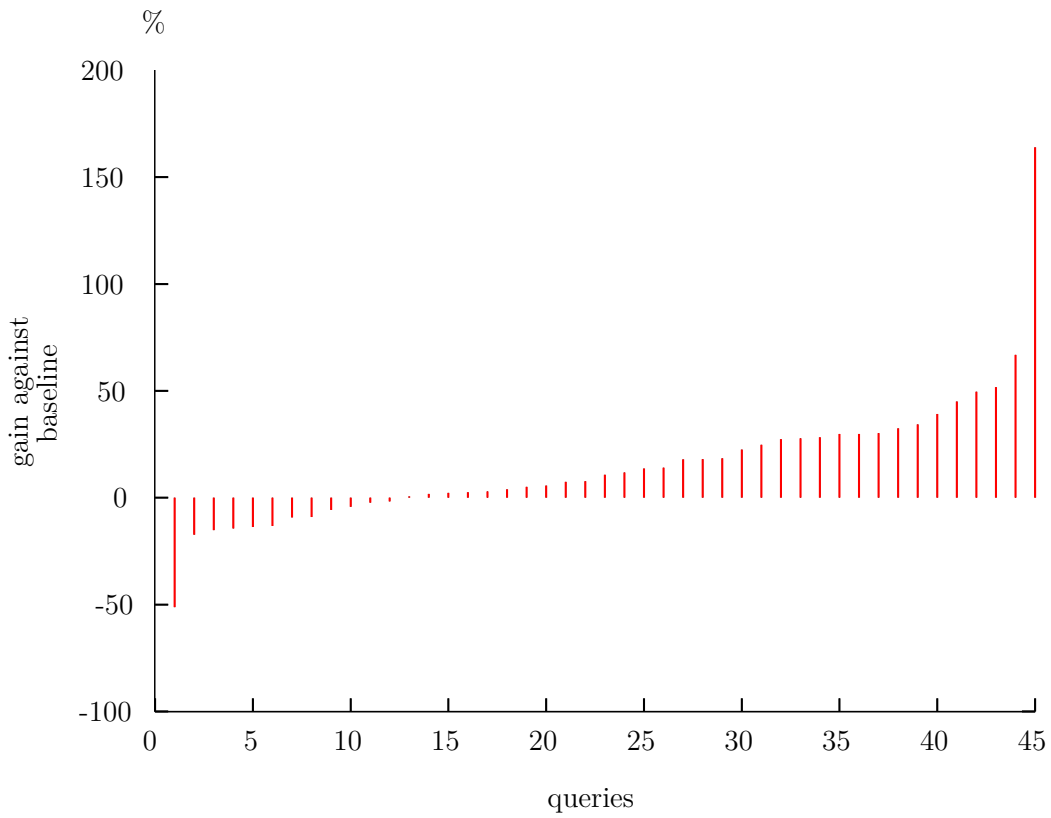

Figure 6: Relative MAP difference between the combination of all document descriptions and the textual results for the 45 ImageCLEF 2009 competition queries.

coefficients were studied.

The first method, already proposed in [12, 37], can be considered as the state of the art linear combination method. It is based on the optimization of the MAP and finds the combination parameters that maximizes the MAP on the training set of queries. The second method, which is the new contribution of this article, uses the Fisher-LDA to calculate the optimal linear combination. As shown in section 4.5, it has one great advantage of providing an analytical solution with a low complexity for any number of modalities (typically a text modality and several visual ones).

We carried out experiments using ImageCLEF collection extracted from Wikipedia. Considering the combination of two vocabularies (one textual and one visual), we showed that the two previous methods can improve the MAP by about 8% compared to the baseline which used the textual vocabulary only. Considering the combination of three vocabularies (one textual and two visuals), only the second method is relevant and provided an increase of 13% over MAP. A more detailed analysis of the results showed that performance varies depending on the requests. For about a quarter of the queries the results were degraded (down to 50%) as they were improved for other queries (up to a 150% improvement).

This work can be extended to other types of multimedia documents and particularly videos. In this context, information from sound, text, image or movement may be used. The textual information may come from different sources as annotation of the video, audio transcription or character recognition in the image part. Given a sampling set of queries for training, Fisher LDA would automatically find the right set of weights corresponding to in-

formation sources. Another interesting perspective would be to adjust the weights depending on the queries. This requires to define classes of queries which have the same weighting parameters and to learn both a query model and the corresponding weighting parameters. Given a test query, after determining its class, it would be possible to apply it a particular set of weighting parameters.

8. Acknowledgement

The authors wish to thank Chris Yukna for his help in proofreading.

Appendix A. Evaluation measures

In order to evaluate the performance of information retrieval systems, there exist different measures based on precision and recall. We consider the recall and the average precision to take into account the ranking of N_k documents returned by the system. For a query q_k, D_k corresponds to the documents of $\mathcal{D}$ which are relevant for q_k. The recall R is obtained by dividing the number of relevant retrieved documents by the number of relevant documents to retrieve:

$$R = \sum_{r=1}^{N_k} rel_k(r))/|\mathcal{D}_k| \tag{A.1}$$

The average precision is obtained by:

$$AP_k = \frac{\sum_{r=1}^{N_k}(P_k(r).\mathrm{rel}_k(r))}{|\mathcal{D}_k|} \tag{A.2}$$

where $\mathrm{rel}_k(r)$ is a binary function equals to 1 if the r^{th} returned documents by the system is relevant or 0 otherwise. The performance of information

retrieval systems are evaluated on a set of queries $\mathcal{Q} = \{q_1, \ldots, q_k, \ldots, d_{|\mathcal{Q}|}\}$ by the mean average precision:

$$MAP = \frac{\sum_{k=1}^{|\mathcal{Q}|} AP_k}{|\mathcal{Q}|} \tag{A.3}$$

References

[1] A. W. Smeulders, M. Worring, S. Santini, A. Gupta, R. Jain, Content-based image retrieval at the end of the early years, IEEE Transactions on Pattern Analysis and Machine Intelligence 22 (2000) 1349–1380.

[2] M. Lew, N. Sebe, C. Djeraba, R. Jain, Content-based multimedia information retrieval: State of the art and challenges, ACM Transactions on Multimedia Computing, Communications, and Applications 2 (2006) 1–19.

[3] M. Flickner, H. Sawhney, J. Ashley, Q. Huang, B. Dom, M. Gorkani, J. Hafner, D. Lee, D. Petkovic, D. Steele, P. Yanker, Query by image and video content: The qbic system, IEEE Computer 28 (1995) 23–32.

[4] S. Antani, R. Kasturi, R. Jain, A survey on the use of pattern recognition methods for abstraction, indexing and retrieval of images and video, Pattern Recognition 35 (4) (2002) 945–965.

[5] Y. Liu, D. Zhang, G. Lu, W.-Y. Ma, A survey of content-based image retrieval with high-level semantics, Pattern Recognition 40 (1) (2007) 262–282.

[6] R. Datta, D. Joshi, J. Li, J. Z. Wang, Image retrieval: Ideas, influences, and trends of the new age, ACM Computing Surveys 40 (2008) 5:1–60.

[7] G. Csurka, C. Dance, L. Fan, J. Willamowski, C. Bray, Visual categorization with bags of keypoints, in: ECCV'04 : 8th European Conference on Computer Vision : workshop on Statistical Learning in Computer Vision, 2004, pp. 59–74.

[8] S. Lazebnik, C. Schmid, J. Ponce, Beyond bags of features: Spatial pyramid matching for recognizing natural scene categories, in: CVPR 2006, IEEE Computer SocietyConference on Computer Vision and Pattern Recognition, Vol. 2, IEEE, 2006, pp. 2169–2178.

[9] J.-J. Aucouturier, B. Defreville, F. Pachet, The bag-of-frames approach to audio pattern recognition: A sufficient model for urban soundscapes but not for polyphonic music, The Journal of the Acoustical Society of America 122 (2007) 881.

[10] C. Schuldt, I. Laptev, B. Caputo, Recognizing human actions: A local svm approach, in: Proceedings of ICPR 2004, 17th International Conference on Pattern Recognition, Vol. 3, IEEE, 2004, pp. 32–36.

[11] R. Yan, A. G. Hauptmann, A review of text and image retrieval approaches for broadcast news video, Information Retrieval 10 (2007) 445–484.

[12] C. Moulin, C. Barat, C. Lemaître, M. Géry, C. Ducottet, C. Largeron, Combining text/image in wikipediamm task 2009, in: CLEF'09 : 10th workshop of the Cross-Language Evaluation Forum, 2009, pp. 164–171.

[13] P. Atrey, M. Hossain, A. El Saddik, M. Kankanhalli, Multimodal fusion

for multimedia analysis: a survey, Multimedia Systems 16(6) (2010) 345–379.

[14] B. T. Bartell, G. W. Cottrell, R. K. Belew, Automatic combination of multiple ranked retrieval systems, in: Proceedings of the 17th annual international ACM SIGIR conference on Research and development in information retrieval, SIGIR '94, Springer-Verlag New York, Inc., New York, NY, USA, 1994, pp. 173–181.

[15] J. A. Shaw, E. A. Fox, Combination of Multiple Searches, in: The Second Text REtrieval Conference (TREC-2, 1993, pp. 243–252.

[16] C. C. Vogt, G. W. Cottrell, Fusion via a linear combination of scores, Information Retrieval 1 (1999) 151–173.

[17] A. Ross, A. Jain, Information fusion in biometrics, Pattern Recognition Letters 24 (2003) 2115–2125.

[18] A. Rattani, D. R. Kisku, M. Bicego, M. Tistarelli, Feature Level Fusion of Face and Fingerprint Biometrics, BTAS 2007. First IEEE International Conference on Biometrics: Theory, Applications, and Systems (2007) 1–6.

[19] Y. Fu, L. Cao, G. Guo, T. S. Huang, Multiple feature fusion by subspace learning, in: Proceedings of the 2008 international conference on Content-based image and video retrieval, CIVR '08, ACM, New York, NY, USA, 2008, pp. 127–134.

[20] K. Barnard, P. Duygulu, D. Forsyth, N. de Freitas, D. M. Blei, M. I.

Jordan, Matching words and pictures, Journal of Machine Learning Research 3 (2003) 1107–1135.

[21] A. Depeursinge, D. Racoceanu, J. Iavindrasana, G. Cohen, A. Platon, P.-A. Poletti, H. Müller, Fusing visual and clinical information for lung tissue classification in high-resolution computed tomography, Artificial Intelligence in Medicine 50 (2010) 13–21.

[22] A. Macedonas, S. Fotopoulos, G. Economou, Improvement of image retrieval by fusing different descriptors, in: Proceedings of the Eight International Workshop on Image Analysis for Multimedia Interactive Services, IEEE Computer Society, Washington, DC, USA, 2007, pp. 75–75.

[23] R. O. Duda, P. E. Hart, D. G. Stork, Pattern Classification (2nd Edition), 2nd Edition, Wiley-Interscience, 2001.

[24] S. Wu, F. Crestani, Data fusion with estimated weights, in: Proceedings of the eleventh international conference on Information and knowledge management, CIKM '02, ACM, New York, NY, USA, 2002, pp. 648–651.

[25] R. Nuray, F. Can, Automatic ranking of information retrieval systems using data fusion, Information Processing and Management 42 (3) (2006) 595–614.

[26] J. A. Aslam, M. Montague, Models for metasearch, in: Proceedings of the 24th annual international ACM SIGIR conference on Research and development in information retrieval, SIGIR '01, ACM, New York, NY, USA, 2001, pp. 276–284.

[27] M. Montague, J. A. Aslam, Condorcet fusion for improved retrieval, in: Proceedings of the eleventh international conference on Information and knowledge management, CIKM '02, ACM, New York, NY, USA, 2002, pp. 538–548.

[28] C. G. M. Snoek, Early versus late fusion in semantic video analysis, in: ACM Multimedia, 2005, pp. 399–402.

[29] P. S. Aleksic, J. J. Williams, Z. Wu, A. K. Katsaggelos, Audio-visual speech recognition using mpeg-4 compliant visual features, EURASIP Journal on Applied Signal Processing 2002 (2002) 1213–1227.

[30] G. Iyengar, H. Nock, C. Neti, Audio-visual synchrony for detection of monologues in video archives, in: Proceedings of the 2003 International Conference on Multimedia and Expo - Volume 2, 2003, pp. 329–332.

[31] M.-C. Cheung, M.-W. Mak, S.-Y. Kung, A two-level fusion approach to multimodal biometric verification, IEEE Int. Conf. Acoust. Speech, Signal Processing (ICASSP) 5 (2005) 485–488.

[32] D. N. F. A. Iskandar, J. Pehcevski, J. A. Thom, S. M. M. Tahaghoghi, Combining image and structured text retrieval., in: INEX'05, 2005, pp. 525–539.

[33] M. Torjmen, K. Pinel-Sauvagnat, M. Boughanem, Methods for combining content-based and textual-based approaches in medical image retrieval, in: CLEF, 2008, pp. 691–695.

[34] L. A. Alexandre, A. C. Campilho, M. Kamel, Combining independent

and unbiased classifiers using weighted average, International Conference on Pattern Recognition 2 (2000) 2495–2498.

[35] Y. Wang, T. Tan, A. K. Jain, Combining face and iris biometrics for identity verification, in: Proceedings of the 4th international conference on Audio- and video-based biometric person authentication, AVBPA'03, Springer-Verlag, Berlin, Heidelberg, 2003, pp. 805–813.

[36] R. Fisher, The use of multiple measurements in taxonomic problems, Annals of Eugenics 7 (2) (1936) 179–188.

[37] C. Moulin, C. Largeron, M. Géry, Impact of visual information on text and content based image retrieval, in: S+SSPR'10 :13th international workshop on Structural, Syntactic, and Statistical Pattern Recognition, 2010, pp. 159–169.

[38] G. Salton, A. Wong, C. Yang, A vector space model for automatic indexing, Communications 18 (1975) 613–620.

[39] G. Salton, M. J. McGill, Introduction to modern Information Retrieval, McGraw-Hill, New York, NY, USA, 1983.

[40] S. Robertson, S. Walker, M. Hancock-Beaulieu, A. Gull, M. Lau, Okapi at trec-3, in: TREC-3 : 3rd Text REtrieval Conference, 1994, pp. 21–30.

[41] C. Zhai, Notes on the lemur tfidf model, Tech. rep., Carnegie Mellon University (2001).

[42] D. Lowe, Object recognition from local scale-invariant features, in:

ICCV'99 : 7th International Conference on Computer Vision, 1999, pp. 1150–1157.

[43] D. Lowe, Distinctive image features from scale-invariant keypoints, International Journal of Computer Vision 60 (2) (2004) 91–110.

[44] S. Tollari, H. Glotin, Web image retrieval on imageval: Evidences on visualness and textualness concept dependency in fusion model, in: CIVR'07 : ACM International Conference on Image and Video Retrieval, 2007, pp. 65–72.

[45] S. Tollari, M. Detyniecki, C. Marsala, A. Fakeri-Tabrizi, M.-R. Amini, P. Gallinari, Exploiting visual concepts to improve text-based image retrieval, in: ECIR'09 : Proceedings of European Conference on Information Retrieval, 2009, pp. 701–705.

[46] C. D. Manning, P. Raghavan, H. Schütze, Introduction to Information Retrieval, Cambridge University Press, 2008.

[47] J. Nocedal, S. Wright, Numerical optimization, Springer verlag, 1999.

[48] R. Fisher, The use of multiple measurements in taxonomic problems, Annals of Human Genetics 7 (2) (1936) 179–188.

[49] W. Klecka, Discriminant analysis, Vol. 19, Sage Publications, Inc, 1980.

[50] S. Mika, G. Ratsch, J. Weston, B. Scholkopf, K. Mullers, Fisher discriminant analysis with kernels, in: Neural Networks for Signal Processing IX, 1999. Proceedings of the 1999 IEEE Signal Processing Society Workshop, IEEE, 1999, pp. 41–48.

[51] I. Jolliffe, Principal Component Analysis, Springer Series in Statistics, 2nd ed, 2002.

[52] G. McLachlan, J. Wiley, Discriminant analysis and statistical pattern recognition, Wiley Online Library, 1992.

[53] T. Tsikrika, J. Kludas, Overview of the wikipediamm task at imageclef 2008, in: CLEF'08 : 9th workshop of the Cross-Language Evaluation Forum, 2008, pp. 539–550.

[54] T. Tsikrika, J. Kludas, Overview of the wikipediamm task at imageclef 2009, in: CLEF'09 : 10th workshop of the Cross-Language Evaluation Forum, 2009, pp. 60–71.