Semantic Combination of Textual and Visual Information in Multimedia Retrieval

Stéphane Clinchant
Xerox Research Centre Europe
6, chemin de Maupertuis
38240 Meylan, France
stephane.clinchant@xrce.xerox.com

Julien Ah-Pine
ERIC Lab, University of Lyon 2
5, avenue Pierre Mendès France
69500 Bron, France
julien.ah-pine@eric.univ-lyon2.fr

Gabriela Csurka
Xerox Research Centre Europe
6, chemin de Maupertuis
38240 Meylan, France
gabriela.csurka@xrce.xerox.com

ABSTRACT

The goal of this paper is to introduce a set of techniques we call *semantic combination* in order to efficiently fuse text and image retrieval systems in the context of multimedia information access. These techniques emerge from the observation that *image and textual queries are expressed at different semantic levels* and that a single image query is often ambiguous. Overall, the semantic combination techniques overcome a *conceptual barrier* rather than a technical one: these methods can be seen as a combination of late fusion and image reranking. Albeit simple, this approach has not been used yet. We assess the proposed techniques against late and cross-media fusion using 4 different ImageCLEF datasets. Compared to late fusion, performances significantly increase on two datasets and remain similar on the two other ones.

Categories and Subject Descriptors

H.3.3 [**Information Storage and Retrieval**]: Information Search and Retrieval—*Retrieval model,Search Process*

General Terms

Algorithms,Theory

Keywords

Information fusion, Multimedia retrieval

1. INTRODUCTION

Nowadays, multimedia data are everywhere: from large digital libraries to the web content, we are surrounded by multimedia information both in the context of our professional or personal activities. For several decades, the multimedia community has been interested in designing multimedia search systems in order to access multimedia collections. Nevertheless, a core challenge that still makes multimedia information retrieval an open problem is the "semantic gap". On the one hand, multimedia data such as images, videos, are stored in machines into a computational representation which consists of low-level features. On the other hand, humans who search digital collections express their information needs by using high-level concepts such as keywords. It is a difficult task to map both representations in order to match the information need of a user and the items of the collection. In particular, it is very challenging to automatically extract the semantic content of an image or a video.

In this paper we are interested in accessing in an efficient way a multimedia collection made of text/image objects or documents. In order to better depict the context of this research work let us take the example of the Wikipedia collection. Traditional systems rely on text based searches. However, the Wikipedia collection is a multimedia one where we encounter texts illustrated with images. Those images can be natural pictures, charts, logos, paintings and so on. In that case, one could be interested in searching for multimedia objects given a multimedia query which would be here a set of keywords along with a set of images. There exist many text based search engines and content based image retrieval (CBIR) systems that respectively address text search and image search tasks. This paper particularly focuses on the problem of combining results provided by both types of systems in the goal of better leveraging the complementarities of the two modalities to enhance multimedia information retrieval. Combining two types of information which are semantically expressed at different levels such as texts and images is an instance of the "semantic gap" problem.

There has been many research works addressing text/image information fusion. The intuition behind the technique we are going to introduce is the following one: since different media are semantically expressed at several levels, one should not combine them independently as most of information fusion techniques employed so far do. On the contrary, one should consider the underlying complementarities that exist between the media when combining them. In the case of text/image data fusion, it is well-known that text based search is more efficient than visual based one since it is more difficult to extract the semantics of an image compared to a text. In the meantime, basic late fusion approaches showed that we can still perform better than text based search by combining them with visual similarities, even with naive approaches. This shows that both media are complementary to each other despite the differences between monomedia performances. The goal of this paper is thus to introduce a set of techniques that better manage the complementarities between text and image search systems when combining them.

The rest of this paper is organized as follows. We first recall prior art on information fusion in section 2. We claim that state-of-the-art models are insufficient to handle the type of *complementarities* that exists between texts and images. As a result, we propose in section 3, a new approach that allows a better combination of text/image information fusion. In particular, we claim that the complementar-

ICMR '11 Trento, Italy

ities between texts and images are *asymmetric* and we propose a simple yet efficient approach to underline this aspect. To validate our proposal we experimented with 4 different digital collections of text/image documents from the ImageCLEF evaluation campaigns. The results we obtained are detailed and analyzed in section 4. Finally, we conclude in section 5.

2. PRIOR ART

This paper deals with data fusion problems in multimedia information access. More particularly, we are interested in multimedia objects that are made of a text and an image. There has been an explosion of such type of digital content for the last years. Likewise, many popular web applications such as Wikipedia, Flickr and so on consist of text/image data. As a consequence, the multimedia community has been interested in developing methodologies and technologies to access text/image content. As an example, for the last 10 years, there has been a large effort to carry out evaluation campaigns in order to boost this research topic. The ImageCLEF sessions are well-known examples of such meetings where research groups participate in challenges to address text/image content retrieval in real world applications [19]. In that context, studies that address the problem of combining textual and visual information in order to bridge the semantic gap are important for our study. Images are represented by low-level features while texts are represented by high-level ones and mapping both types of information is a difficult problem. Ideally, one would like to retrieve images given a text query that describes the content of the latter. This is a strong challenge and most of current tools for searching image collections for example, use a text based image retrieval (TBIR) approach where the texts around an image or even its filename are indexed.

Regarding text/image retrieval, we generally observe better performances for text based image search systems compared to content based image retrieval (CBIR) systems. However, most of research works in text/image information retrieval have shown that combining text and image information even with simple fusion strategies, allow one to increase multimedia retrieval results. It is admitted that as long as the modality dependencies are well exploited, data fusion is beneficial to multimedia retrieval [13]. In the survey paper [16], it is pointed out that one of the major challenges in multimedia information retrieval is *"multi-modal analysis and retrieval algorithms especially towards exploiting the synergy between the various media including text and context information"*. Accordingly, in the literature, there has been many works covering text/image information fusion. Most of the techniques developed in that context fall in three different categories : early fusion, late fusion and transmedia fusion. We attempted to characterize these three families of approaches by distinguishing the inherent steps that they are made of. This is summarized in Figure 1.

The early fusion approach consists in representing the multimedia objects in a multimodal feature space designed via a joint model that attempts to map image based features with text based features. The simplest early fusion method consists in concatenating both image and text feature representations. However, more elaborated joint models such as Canonical Correlation Analysis have been investigated [18, 14, 28, 26].

On the contrary, late fusion and transmedia fusion strategies do not act at the level of the monomedia features representations but rather at the level of the monomedia similarities [7, 4]. In these contexts, we assume that we have monomedia retrieval systems that are efficient and thus that it is better to combine their respective decisions rather than attempting to bridge the semantic gap a priori that is to say at the level of the features.

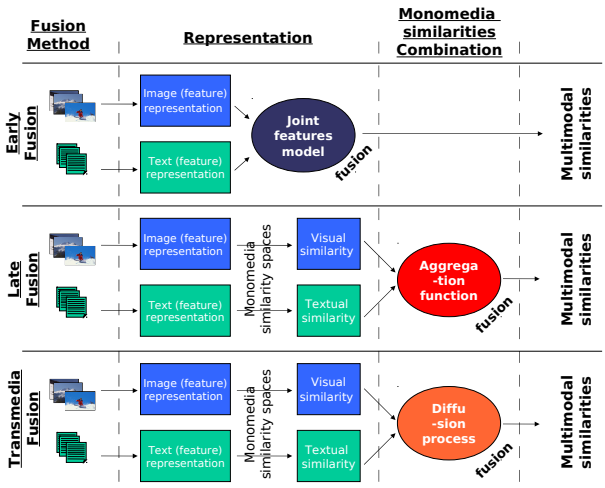

Figure 1: Early, late and transmedia fusion.

Concerning late fusion techniques, they mainly consist in merging the monomedia similarity profiles by means of aggregation functions. In that case, the simplest aggregation technique used is the mean average [10] but more elaborated approaches have been studied (e.g. [5]).

As far as transmedia fusion methods are concerned, diffusion processes that act as a transmedia pseudo-relevance mechanism are used instead of simple aggregation functions as compared to late fusion methods. The main idea is to first use one of the modalities (say image) to gather relevant documents (nearest neighbors from a visual point of view) and then to switch to the other modality (text representations of the visually nearest neighbors) and aggregate their (text) profiles (see for example [12, 17, 1]).

In many experiments reported in the literature [19], it has been shown that either late fusion or transmedia fusion approaches have been performing better than early fusion techniques. Consequently, we are focusing on those techniques and more particularly, we introduce in the sequel the baseline techniques we are going to compare our proposal to.

We assume that we have a multimodal query q and two experts s_t and s_v. s_t represents the text based similarity between the textual representation of q and the text part of the multimedia objects d within the collection we want to search. Likewise, s_v is the similarity scores obtained using a content based image retrieval system given the image(s) (visual part) of q.

• Late Fusion

It consists in running the visual and textual experts independently. Then, each expert returns a list made of its top-K most similar items. After having normalized the two scores' distributions so that they belong to the same scale ($[0, 1]$ in practice), the visual and textual scores are linearly combined as follows:

$$s_{Late}(q, d) = nz(d)^{\gamma} \left(\alpha_t \mathbf{N}(s_t(q, d)) + \alpha_v \mathbf{N}(s_v(q, d))\right) \quad (1)$$

where $\mathbf{N}$ is a normalization operator that transforms a set of similarity scores in order to have values between 0 and 1 ; $\alpha_t = \alpha$, $\alpha_v = 1 - \alpha$ are weights that sum to 1; $nz(d) \in \{0, 1, 2\}$ is the number of experts for which d appears in the top-K list; and $\gamma \in \{0, 1\}$ is a parameter. When $\gamma = 0$ it gives the classic arithmetic mean average (also refered as late fusion in the paper) while $\gamma = 1$ leads to the so-called CombMnz fusion operator. Besides, the weights α_t, α_v allow one to give an asymmetric importance to

Figure 2: Wikipedia Image Query 67: "white house with garden"

the media, as one can set a higher weight to the expert that should be trust the most. Generally, these weights are set manually. As it was pointed out in [9], late fusion is the most used technique in text/image information fusion.

• **Image Reranking**

Another popular choice in text/image fusion is called *image reranking [25, 2]*. It consist of two phases: first a text search is used, then the returned objects according to s_t are reordered but according to the visual similarity s_v. In other words, image reranking constrains the visual system to search *among* the set of objects that was returned by the text search instead of the entire collection. This approach can be formulated as follows :

$$s_{Rerank}(q,d) = I_{\{d \in \mathrm{KNN}_t(q)\}} s_v(q,d) \quad (2)$$

where $\mathrm{KNN}_t(q)$ denotes the set of the K most similar objects to q according to the textual similarities s_t and, $I_{\{A\}} = 1$ if proposition A is true and $I_{\{A\}} = 0$ otherwise.

• **Cross-Media Similarities**

The cross-media similarity proposed in [7] is an approach that is part of the transmedia fusion category. It can be written as follows:

$$s_X(q,d) = \alpha_t s_t(q,d) + \alpha_v \sum_{d' \in \mathrm{KNN}_v(q)} s_v(q,d') s_t(d',d) \quad (3)$$

where $\mathrm{KNN}_v(q)$ denotes the set of the K most similar objects to q using visual similarities s_v, and $s_t(d',d)$ is the textual similarity between the items d and d' of the collection.

3. SEMANTIC COMBINATION

When text is used as query, only a few keywords are usually provided. In contrast, when image is used as query, "all the information it contains" is provided to the system. Indeed, it is generally said that *"a picture is worth a thousand words"* but in the context of information retrieval, which word(s) is meant when an image is used as a query ? CBIR systems attempt to find visually similar images but in many cases we are rather interested in some underlying semantic meanings of an image.

To illustrate this ambiguity, we asked several users to choose words that would help a system to find images "semantically" similar to the one shown in Figure 2. Users chose *"house", "country house stately home", "fancy house", "wealthy house", "house outside", "house, trees, grass and clouds", "manor park", "mansion park stormy weather", "residence mansion park"*. This picture was used as an image query in one of the ImageCLEF tasks. None of the users who were asked, gave the actual text query that was used with this image: *white house with garden*. This simple (and limited) user test actually suggests that images can be interpreted in several ways. Can we expect visual features to correctly represent the semantics of an image when even humans do not agree on the semantic this image conveys ? Moreover, visual features are known to be lower level features than text with respect to a semantic scale: color and texture features do not fully describe a concept such as a word can. Even with so-called "high-level features" such as bags-of-visual words (BOV) or Fisher Vectors [8, 21], we need to train an image classifier per concept with many labeled images to be able to transform these low-level features into higher level concepts.

These "high-level" image features incorporate more relevant information, but they should be exploited accordingly. Let us further illustrate those statements using Figure 3. In the first row, we show one of the ImageCLEF Wikipedia query *"sharks underwater"* with the associated query images. Top retrieved images obtained by the visual expert (second row) are indeed visually similar (blue background with fish like shapes) to the image queries. Yet, they miss an important aspect of the query: none of them actually contains a shark. This information can easily be inferred from the text query. This typical example shows why visual runs get poor performances compared to text runs. For instance, for the ImageCLEF Wikipedia challenge, visual runs obtain 4% of MAP (Mean Average Precision) whereas textual runs reach above 20% of MAP. Visual runs do retrieve similar images but they completely ignore the underlying semantic. The last row of Figure 3 shows the results we obtained with the method we propose.

We can summarize the above observations in the following points:

- There is a semantic mismatch between visual and textual queries: image queries are ambiguous for humans from a semantic viewpoint. Fusion techniques should be asymmetric since text and image are expressed at different semantic levels.

- Visual techniques do work and are effective to retrieve visually similar images but they usually fail in multimedia retrieval due to the semantic mismatch.

Therefore, this difference in semantic levels should result in an *asymmetry* when considering the two different media. However, state-of-the-art multimedia fusion systems rely on symmetric schemas such as late fusion or CombMnz operator. These fusion operators can only express a "weak asymmetry" by different choices of α_t and α_v (e.g. in in Eq. (1)). In what follows, we propose a semantic filtering method which, when associated to the late fusion approach, leads to better results even without guessing or learning the different weights for each modality.

Previous analysis and several reports on ImageCLEF challenges showed that:

1. Late fusion is able to outperform text based systems

2. Image reranking tends to have lower performances than textual based systems as several ImageCLEF participants noticed[1].

Note that, an interesting way to start addressing the semantic mismatch is to use the image reranking technique. In fact, by enforcing the visual system to search *among* the set of retrieved objects by the text expert, we somehow impose that images visually similar to the query images share a common semantic (given by the textual query). However, while in this way image reranking does alleviate the semantic mismatch, it does not take into account the textual scores explicitly. This observation led us to defend the

[1]E.g. in [25]: *"Multimodal runs involved a k-NN inspired visual reranking of textual results and actually degraded the final quality of results."*.

Figure 3: Multimodal query and results on the ImageCLEF Wikipedia Dataset. Top row displays the text query and its two associated images. Middle row shows the most similar objects according to visual similarities. Bottom row shows the results of our semantic combination of text and image.

following argument: *image reranking is not really a fusion of textual and visual information. It is a visual similarity corrected by a semantic filter and we could do better.*

Consequently, if image reranking is "rather" a visual method, the textual similarity should also be taken into account. Therefore, what we propose is to *combine image reranking with late fusion* in order to overcome their respective weaknesses. The strength of image reranking is to realign the visual system to search in a relevant subset with respect to the semantic viewpoint, while the strength of late fusion relies on a well performing text expert. While our proposal seems to be simple, we argue that the combination of image reranking and late fusion has not been tried before, to our knowledge, due to a conceptual gap. Indeed, in the community, image reranking is seen as a combination of text and image. So generally, it is assumed that adding the text scores would lead to no extra gain since the text was already used. For example, [2] explains that: *"Thus, the textual module works as a filter for the visual module, and the work of the visual module is to re-order the textual results list. In this way, there has not been used an explicit fusion algorithm to merge the textual result list and the visual result list."*

We will show, through our experiments, that on the contrary, combining textual scores with image scores that went through a semantic filter beforehand is a winning strategy since it significantly outperforms both the late fusion and the image reranking results.

To present our *semantic combination* we first introduce the *semantic filtering* of image scores according to the text expert:

$$SF(q,d) = \begin{cases} 1 & \text{if } d \in \mathrm{KNN}_t(q) \\ 0 & \text{otherwise} \end{cases} \quad (4)$$

where $\mathrm{KNN}_t(q)$ denotes the set of the K most similar objects to q according to the textual similarities. After normalization, the semantically filtered image scores are combined with the text ones:

$$s_{LSC}(q,d) = \alpha_t \mathbf{N}(s_t(q,d)) + \alpha_v(\mathbf{N}(SF(q,d)s_v(q,d))) \quad (5)$$

where $\alpha_t = \alpha$ and $\alpha_v = 1 - \alpha$ are positive weights that sum to 1. We call this method *Late Semantic Combination* (LSC).

Eq. (5) can be simply interpreted as a late fusion between the text retrieval and the image reranking method. Another combination method is the non-parametric CombProd operator, where the scores are multiplied by each other. This is rank equivalent to a geometric mean of the scores. We call such a combination *Product Semantic Combination* (PSC) , which amount to rank documents with:

$$s_{PSC}(q,d) = \mathbf{N}(s_t(q,d)) \times \mathbf{N}(SF(q,d)s_v(q,d)) \quad (6)$$

Note that for both aggregation methods, the idea is that the system considers only image scores for documents that were retrieved by the textual expert among the top-K. Hence, the selection of these K multimodal documents can be seen as a semantic filtering of the corresponding images before the fusion. While this function seems to be quite simple, it has several advantages. On one hand, we will show in the experiments that this simple strategy allows us to obtain significant improvements over late fusion and image reranking. On the other hand, we also gain in terms of computational cost as we only need to compute visual similarities between the query and the filtered objects. Indeed, instead of letting the two experts search and rank independently, we first run the textual expert and provide the top-K list to the visual expert. In that way, the system becomes scalable even for very large datasets[2].

4. EXPERIMENTS

We tested and compared the proposed approach to state-of-the-art fusion techniques using 4 different ImageCLEF datasets:

- **IAPR.** The IAPR TC-12 photographic collection consists of 60 topics and 20,000 still natural images taken from locations all around the world and including an assorted cross-section of still natural images [11]. This includes pictures of different sports and actions, photographs of people, animals, cities, landscapes and many other aspects of contemporary life. Image captions include the title of the image, the location from which the photograph was taken, and a semantic description of the content of the image (as given by the photographer).
- **BELGA.** The Belga News Collection contains 498,920 images from Belga News Agency, which is an image search engine for news photographs. Each photograph will be up to a maximum of 512 pixels in either width or height, accompanied with a caption composed of English text up to a few sentences in length. Each caption can contain the date and the place where the image was captured. From this collection we used only the subset of 73240 images for which relevance judgements were provided and topics without cluster information (topic 26 to 50).

[2]Current retrieval systems such as Google, Yahoo, Bing are able to handle several millions even billions of documents and retrieve relevant documents based on textual queries in few seconds. On the contrary even with one of the best state-of-the-art CBIR systems (e.g. [22]), the content based image retrieval performances are far lower than text based ones.

- **WIKI.** The Wikipedia collection consists of 70 topics and 237,434 images and associated user-supplied annotations in English, German and/or French. In addition, the collection contains the original Wikipedia pages in wikitext format from where the images were extracted.

- **MED.** The medical image collection consists of 16 topics and 77,477 medical images of different modalities, such as CT, MR, X-Ray, PET microscopic images but also graphical plots and photos. In the ad-hoc retrieval task [20], the participants were given a set of 16 textual queries with 2-3 sample images for each query. The queries were classified into textual, mixed and semantic queries, based on the methods that are expected to yield the best results. In our experiments we did not consider this explicit query classification, but handled all queries in the same way.

4.1 Text Representation

Standard preprocessing techniques were first applied to the textual part of the documents. After stop-word removal, words were lemmatized and the collection of documents indexed with Lemur[3]. We varied the model for text retrieval including state-of-the-art methods such as standard language models and information models [30, 6]. The idea of language models is to represent queries and documents by multinomial distributions [29, 24]. Those distributions are estimated by maximizing their likelihood. Then, document distributions are smoothed with a Dirichlet Prior and the Cross-Entropy[4] measure was used to rank documents:

$$s_t(q,d) = CE(q|d) = \sum_w p(w|q) \log p(w|d) \tag{7}$$

Information models can be introduced as follows: the more a word deviates in a document from its average behavior in the collection, the more likely it is "significant" for this particular document. This can be easily captured in terms of information. For example, if a word has a low probability of occurrence in a document, according to the distribution collection, then the amount of information it conveys is more important if it appears. Term frequencies are first normalized and a log-logistic distribution is chosen to model the frequencies Tf_w of a given word in a collection. Documents are ranked by using a mean information over each query term: x_{wd} notes the discrete term frequency of w and t_{wd} a normalized frequency of w in d.

$$s_t(q,d) = \sum_w -x_{wd} \log P(Tf_w > t_{wd}) \tag{8}$$

As there are small differences between the performances of different families of text models, we used in our experiments language models for IAPR and MED and information models for BELGA, WIKI datasets. Additional experiments when the text model changes are not included here. Instead, we focus on varying the image representation in order to cover a wider range of performances.

4.2 Image Representation

As for image representations, we experimented with two popular approaches, the BOV [27, 8], where an image is described by a histogram of quantized local features and the Fisher Vector, proposed in [21]. Both of them are based on an intermediate representation, the visual vocabulary, built on the low-level feature space. We mainly used two types of low-level features, the SIFT-like Orientation Histograms (ORH) and the local RGB statistics (COL) and built an independent visual vocabulary for each of them.

The visual vocabulary was modeled by a Gaussian mixture model (GMM) $p(x|\lambda) = \sum_{j=1}^{N} w_i \mathcal{N}(x|\mu_i, \Sigma_i)$, where each Gaussian corresponds to a visual word. In the BOV representation, the low-level descriptors are transformed into the high-level N-dimensional descriptor (where N is the number of Gaussians) by cumulating over all low-level descriptors x_t for each Gaussian the probabilities of generating a descriptor:

$$\gamma(I) = [\sum_{t=1}^{T} \gamma_1(x_t), \sum_{t=1}^{T} \gamma_2(x_t), \ldots, \sum_{t=1}^{T} \gamma_N(x_t)] \tag{9}$$

where

$$\gamma_i(x_t) = \frac{w_i \mathcal{N}(x_t|\mu_i, \Sigma_i)}{\sum_{j=1}^{N} w_j \mathcal{N}(x_t|\mu_j, \Sigma_j)}. \tag{10}$$

The Fisher Vector [21] extends the BOV by going beyond counting (0-order statistics) and by encoding statistics (up to the second order) about the distribution of local descriptors assigned to each visual word. It characterizes a sample $X = \{x_t, t = 1 \ldots T\}$ by its deviation from the GMM distribution:

$$G_\lambda(I) = \frac{1}{T} \sum_{t=1}^{T} \nabla_\lambda \log \left\{ \sum_{j=1}^{N} w_j \mathcal{N}(x_t|\mu_j, \Sigma_j) \right\}. \tag{11}$$

To compare two images I and J, a natural kernel on these gradients is the Fisher Kernel [21].

$$K(I,J) = G_\lambda(I)^\top F_\lambda^{-1} G_\lambda(J), \tag{12}$$

where F_λ is the Fisher Information Matrix. As F_λ^{-1} is symmetric and positive definite, it has a Cholesky decomposition denoted by $L_\lambda^\top L_\lambda$. Therefore $K(I,J)$ can be rewritten as a dot-product between normalized vectors Γ_λ with: $\Gamma_\lambda(I) = L_\lambda G_\lambda(I)$ which we refer to as the *Fisher Vector* (FV) of the image I. Given the visual parts of a query q and a document d, we thus use $K(q,d) = s_v(q,d)$ as the visual similarity measure.

As suggested in [23], we further used a square-rooted and L2-normalized versions of the BOV and FV and also built a spatial pyramid [15]. Regarding the pyramid, we repeatedly subdivide the image into 1, 3 and 4 regions: we consider the FV of the whole image (1x1); the concatenation of 3 FV extracted for the top, middle and bottom regions (1x3) and finally; the concatenation of four FV one for each quadrants (2x2). We used the dot product (linear kernel) to compute the similarity between the concatenation[5] of all FV for ORH and COL.

4.3 Experimental results

Table 1 shows the Mean Average Precision (MAP) for the 4 datasets when the Spatial Pyramid of ORH and COL Fisher Kernel model of images is used. Late fusion results are given when the weights α are optimized on all queries of one dataset. Similarly, the cross-media parameters are optimized to provide the best performances. In that case, the optimal number of nearest neighbors K in Eq. (3) is 4 for IAPR, 1 for BELGA, 42 for WIKI and 2 for MED. We also provide the results obtained with cross-media when the best number of nearest neighbors is unknown and manually set to 3.

Table 1 allows us to compare the best late fusion, the best cross-media, the best image reranking and the proposed LSC (Eq. (5))

[3] http://www.lemurproject.org/

[4] Actually, the Cross-Entropy multiplied by -1.

[5] Note that we do not need to explicitly concatenate all these vectors as $\langle [u,v],[u',v'] \rangle = \langle u,u' \rangle + \langle v,v' \rangle$.

Model Collection	IAPR	BELGA	WIKI	MED
s_t	26.3	56.2	20.5	31.4
s_v	22.1	3.3	5.5	0.9
Best Late Fusion	34.0	56.2	21.9	31.4
Image Reranking	27.6	42.4	19.4	8.3
CombMNZ	33.5	56.0	23.7	27.8
Best Cross-Media	**42.1**	**57.0**	21.6	31.4
Cross-Media with K=3	40.5	56.6	20.6	31.4
Best PSC	34.8	56.2	26.5	33.8
Best LSC	35.4	56.3	**26.6**	**36.9**

Table 1: Mean Average Precision (%) results using Fisher Kernel representation for images (LSC, PSC: K=1000 for SF).

and PSC (Eq. (6)) models. Furthermore, in Figure 4 we show the variation of the MAP for late fusion, CombMnz, Cross-Media and LSC when we vary α ($\alpha \in [0, 0.1, 0.2, ..., 1]$). From Table 1 and the plots in Figure 4 we can have the following conclusions:

- LSC outperforms PSC even without tuning the α parameter.
- LSC generally achieves better results then late fusion, image reranking and CombMnz on all datasets but the BELGA one, where none of these techniques is able to significantly outperform the performance obtained with the text based system ($\alpha = 1$). Moreover, performances are significantly increased by more than 17% on WIKI and by more than 23% on MED compared to the best late fusion.
- LSC is more robust than late fusion with respect to the choice of α. Even with $\alpha = 0.5$ (equal weighting) the LSC fusion results are better or very close to the best performance of late fusion on all datasets.
- The cross-media method outperforms all methods on IAPR and BELGA datasets. However, it is outperformed by late fusion, LSC, PSC on the two other sets (WIKI and MED). This shows that the performances of the cross-media method highly depend on the dataset and on the relation between texts and images in the collection as well. Indeed, while the text is rather aligned to the image in IAPR and BELGA (the texts actually depicts the content of an image), the captions of WIKI and MED are more complementary. Thereby, note that we considered here the optimal α and the optimal K in Eq. (3) and these values are not known a priori in general.
- LSC which relies on semantic filters can improve pure text results even when the cross-media fails to do so.
- The implementation of textual, visual systems and cross - media are competitive as they reach state-of-the-art performances. For example, best published results were 32% of MAP for IAPR in 2007, 35.6% of MAP for MED in 2010 and 27.7% of MAP for WIKI in 2010. Hence, the proposed methods are comparable or outperform the best published results of several ImageCLEF campaigns.

To sum up, the ranking based on LSC is much more robust than the ranking based on the cross-media similarities. We can also see that LSC seems to have less variation of performances when we vary α. Moreover, $\alpha = 0.5$ (equal weighting) gives good results on all datasets. This is very interesting for several applications as in many cases there is no training data available to estimate the best α (and the best K for the cross-media). Hence, the only parameter our method relies on is K in Eq. (4), which is the number of top documents we consider regarding the semantic filter. This parameter does impact the overall performances as Table 2 shows. However if we consider K not too small (e.g greater than 1000) the improvement becomes less important.

K	200	500	800	1000	1200	1500
IAPR	30.6	34.7	35.2	35.4	35.4	35.4
WIKI	24.5	25.6	26.2	26.6	27.2	27.7
MED	32.3	36.1	36.8	36.9	36.9	36.9

Table 2: Best performances of LSC for different K (for SF) on different datasets.

Finally, Figure 5 shows the best performances of late fusion and the LSC method (α optimized) when the image representation is changed. Image representations include several BOV with different numbers of Gaussians, different sizes of Fisher Vectors based on only ORH or both ORH and COL and with or without using a Spatial Pyramid of FV. For each image representation, the best performance with late fusion and the LSC method is shown on the corresponding datasets. We recall that text baselines reach 26% of MAP on IAPR and 20.5% of MAP on Wikipedia. These figures show that despite the very poor performances of some of the visual experts, both fusion methods are able to take advantage of the visual information to improve the text based system, showing that visual and textual are indeed complementary as claimed previously. However, the gain for the proposed method, LSC, is always much higher than late fusion, showing that it better exploits the "asymmetric" complementarity between texts and images and that it also handles the potential noises better.

As for illustration, we show in Figure 4.3 the top seven results found by late fusion against the LSC method for another Wikipedia query. In that example, we can see that the LSC fusion finds more relevant documents at early precision than the late fusion.

5. CONCLUSION

Different media, such as texts and images, are expressed at different semantic levels. As a result, one modality usually outperforms the other one when accessing a multimedia collection by means of monomedia search systems. Despite this observation, media are in fact complementary and their aggregation can improve the retrieval performance. In this paper we discussed common pitfalls in multimedia information retrieval: the underlying semantic asymmetry between text and visual information and the potential ambiguity of visual queries. These observations suggest that visual expert should be used to rank documents only in a subset of the collection: the subset found by the text expert. Then, we proposed a way to correct state-of-the-art fusion methods such as late fusion with semantic filters in order to go beyond image reranking techniques. Finally, we validated our new fusion models with experiments on 4 benchmark datasets. We showed that the proposed semantic combination models for textual and visual information fusion can significantly improve multimedia retrieval performances while being more robust with respect to the choice of the mixing parameter.

6. REFERENCES

[1] J. Ah-Pine, M. Bressan, S. Clinchant, G. Csurka, Y. Hoppenot, and J. Renders. Crossing textual and visual content in different application scenarios. *Multimedia Tools and Applications*, 42(1):31–56, 2009.

[2] J. Benavent, X. Benavent, E. de Ves, R. Granados, and A. Garcia-Serrano. Experiences at ImageCLEF 2010 using

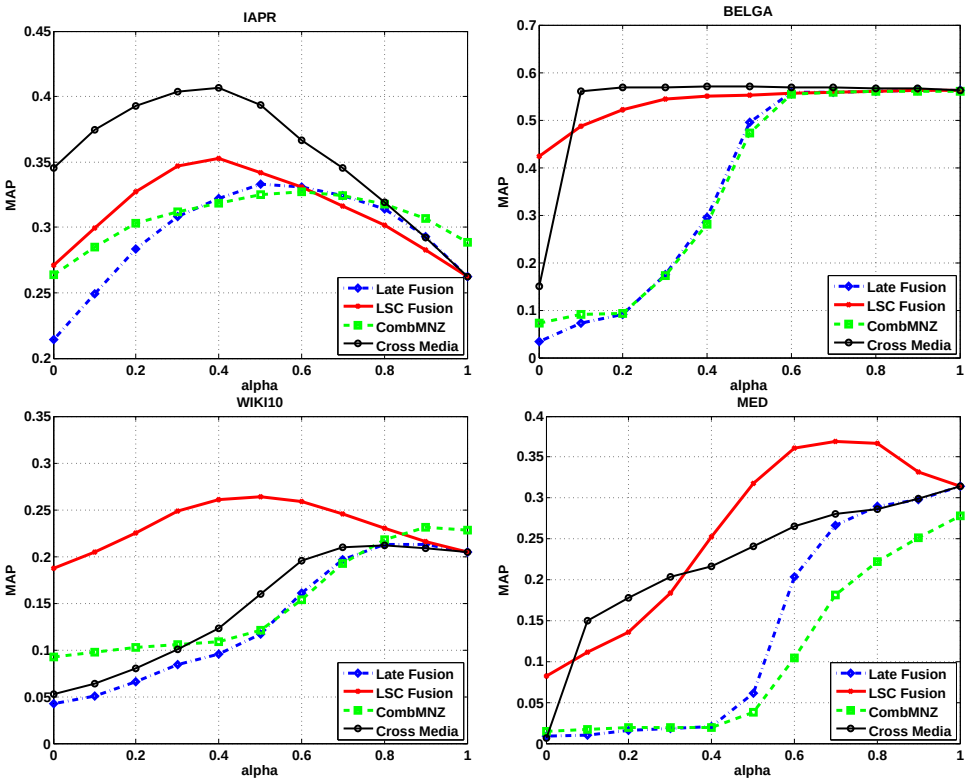

Figure 4: LSC against late fusion, CombMnz, cross-Media with $K = 3$ on 4 ImageCLEF datasets. Figures show Mean Average Precision vs α.

CBIR and TBIR Mixing Information Approaches. In Braschler et al. [3].

[3] M. Braschler, D. Harman, and E. Pianta, editors. *CLEF 2010 LABs and Workshops, Notebook Papers, 22-23 September 2010, Padua, Italy*, 2010.

[4] E. Bruno, N. Moënne-Loccoz, and S. Marchand-Maillet. Design of multimodal dissimilarity spaces for retrieval of video documents. *PAMI*, 30(9):1520–1533, 2008.

[5] J. C. Caicedo, J. G. Moreno, E. A. Niño, and F. A. González. Combining visual features and text data for medical image retrieval using latent semantic kernels. In *Multimedia Information Retrieval*, 2010.

[6] S. Clinchant and E. Gaussier. Information-based models for ad hoc IR. In *SIGIR*. ACM, 2010.

[7] S. Clinchant, J. Renders, and G. Csurka. XRCE's participation to ImageCLEF . In *CLEF Working Notes*, 2007.

[8] G. Csurka, C. Dance, L. Fan, J. Willamowski, and C. Bray. Visual categorization with bags of keypoints. In *ECCV Workshop on Statistical Learning for Computer Vision*, 2004.

[9] A. Depeursinge and H. Müller. *Fusion Techniques for Combining Textual and Visual Information Retrieval*, chapter 6. Volume INRE of Müller et al. [19], 2010.

[10] H. J. Escalante, C. A. Hernández, L. E. Sucar, and M. M. y Gómez. Late fusion of heterogeneous methods for multimedia image retrieval. In *MIR*, 2008.

[11] M. Grubinger, P. Clough, A. Hanbury, and H. Müller. Overview of the ImageCLEFphoto 2007 photographic retrieval task. In *CLEF Working Notes*, 2007.

[12] J. Jeon, V. Lavrenko, and R. Manmatha. Automatic image annotation and retrieval using cross–media relevance models. In *Proceedings of the ACM SIGIR conference*, pages 119–126. ACM press, 2003.

[13] J. Kludas, E. Bruno, and S. Marchand-Maillet. Information fusion in multimedia information retrieval. In *AMR Int. Workshop on Retrieval, User and Semantics*, 2007.

[14] V. Lavrenko, R. Manmatha, and J. Jeon. A model for learning the semantics of pictures. In *NIPS*, 2003.

[15] S. Lazebnik, C. Schmid, and J. Ponce. Beyond bags of features: Spatial pyramid matching for recognizing natural scene categories. In *CVPR*, 2006.

[16] M. S. Lew, N. Sebe, C. Djeraba, and R. Jain. Content-based multimedia information retrieval: State of the art and challenges. *TOMCCAP*, 2(1):1–19, 2006.

[17] N. Maillot, J.-P. Chevallet, V. Valea, and J. H. Lim. IPAL Inter–Media Pseudo–Relevance Feedback Approach to ImageCLEF 2006 photo retrieval. In *CLEF Working Notes*, 2006.

[18] Y. Mori, H. Takahashi, and R. Oka. Image–to–word transformation based on dividing and vector quantizing images with words. 1999.

[19] H. Müller, P. Clough, T. Deselaers, and B. Caputo, editors. *ImageCLEF- Experimental Evaluation in Visual Information Retrieval*, volume INRE. Springer, 2010.

[20] H. Müller, J. Kalpathy-Cramer, I. Eggel, S. Bedrick, C. E. K. Jr., and W. Hersh. Overview of the clef 2010 medical image retrieval track. In Braschler et al. [3].

[21] F. Perronnin and C. Dance. Fisher Kernels on visual vocabularies for image categorization. In *CVPR*. IEEE, 2007.

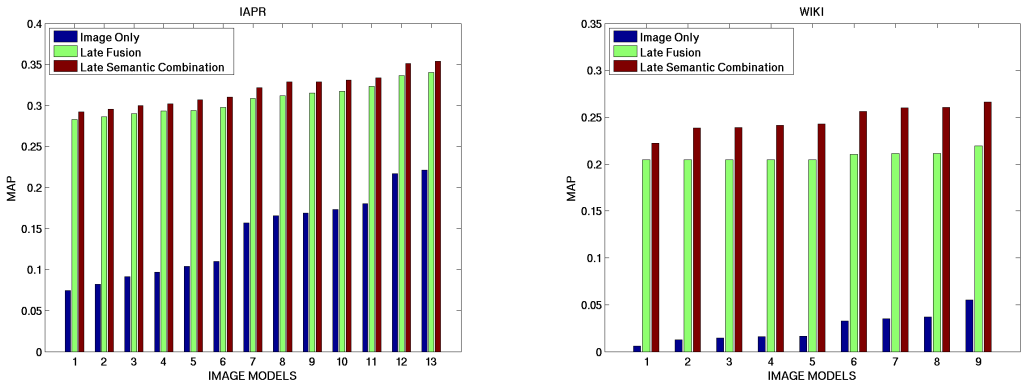

Figure 5: Best late fusion and LSC fusion performances against CBIR performances for different image representations on IAPR and WIKI datasets.

Figure 6: Multimodal query on ImageCLEF Wikipedia Dataset. Top row displays the text query and its two associated images. Middle Row shows the most similar objects according to Late Fusion. Bottom row shows the results with our LSC method.

[22] F. Perronnin, Y. Liu, J. Sánchez, and H. Poirier. Large-scale image retrieval with compressed fisher vectors. In *CVPR*, 2010.

[23] F. Perronnin, J. Sánchez, and T. Mensink. Improving the Fisher Kernelfor large-scale image classification. In *ECCV*, 2010.

[24] J. M. Ponte and W. B. Croft. A language modeling approach to information retrieval. In *SIGIR*. ACM, 1998.

[25] A. Popescu. Télécom bretagne at imageclef wikipediamm 2010. In Braschler et al. [3].

[26] N. Rasiwasia, J. C. Pereira, E. Coviello, G. Doyle, G. R. G. Lanckriet, R. Levy, and N. Vasconcelos. A new approach to cross-modal multimedia retrieval. In *ACM Multimedia*, 2010.

[27] J. S. Sivic and A. Zisserman. Video Google: A text retrieval approach to object matching in videos. In *ICCV*, 2003.

[28] A. Vinokourov, D. R. Hardoon, and J. Shawe-Taylor. Learning the semantics of multimedia content with application to web image retrieval and classification. 2003.

[29] C. Zhai and J. Lafferty. A study of smoothing methods for language models applied to information retrieval. *ACM Trans. Inf. Syst.*, 22(2):179–214, 2004.

[30] C. Zhai and J. D. Lafferty. Model–based feedback in the language modeling approach to information retrieval. In *CIKM*, 2001.