# Leveraging image, text and cross-media similarities for diversity-focused multimedia retrieval

Julien Ah-Pine, Stephane Clinchant, Gabriela Csurka, Florent Perronnin, Jean-Michel Renders

# Contents

# Chapter 1
# Leveraging image, text and cross-media similarities for diversity-focused multimedia retrieval

Julien Ah-Pine, Stephane Clinchant, Gabriela Csurka, Florent Perronnin and Jean-Michel Renders

**Abstract** section abstract

## 1.1 Introduction

Information, especially digital information, is no longer monomodal: web pages can have text, images, animations, sound and video; we have audiobooks, photoblogs and videocasts; valuable content within a photo sharing site can be found in tags and comments as much as in the actual visual content. Nowadays, it is difficult to visit a page within a popular social network without finding a large variety of content modes surrounded by a rich structure of social information such profiles, interest groups, consumer behaviour or simple conversations. This major shift in the way we access content and the type of content we access is largely due to the connected, easily accessible, global nature of the internet. The democratization of the tools of production and delivery have also strongly contributed to this phenomen, one of which is low cost camera-phones combined with accessible publishing tools. Such a scenario poses a strong need for tools that enable interaction with multimodal information.

The scientific challenge is to understand the nature of the interaction between these modalities, and in particular between text and images. How can text be associated with an image (and reciprocally an illustrative image with a text)? How can we organize and access text and image repositories in a better way than naive late fusion techniques? The main difficulty is the fact that visual and textual features are expressed at different semantic levels.

Naive techniques combine the scores from a text retrieval system and from an image retrieval system into a single relevance score: this is the late fusion approach.

Xerox Research Centre Europe
6 ch. de Maupertuis, 38240 Meylan, France
e-mail: FirstName.LastName@xrce.xerox.com

Departing from the classical late fusion strategy, recent approaches have considered fusion at the feature level (early fusion), estimating correspondences or joint distributions between components across the image and text modes from training data.

One of the first approaches in this family is the co-occurrence model by Mori et al (1999) where keywords are assigned to patches based on the co-occurence of clustered image features and textual keywords in a labeled training data-set. Another similar approach proposes to find correlations between images and attached texts using the Kernel Canonical Correlation Analysis (Vinokourov et al, 2003). With the developement of image representation with visual vocabularies (Sivic and Zisserman, 2003; Csurka et al, 2004), new techniques appeared such as Probabilistic Semantic Analysis (Barnard et al, 2003; Monay and Gatica-Perez, 2004) or Latent Dirichlet Allocation (Blei et al, 2003). They propose to extract latent semantics from images. Machine translation models insprired Duygulu et al (2002); Iyengar et al (2005), where these models were generalized to images, where the translation is done between words and image regions. Another group of works used graph models to represent the structure of an image through a graph. Carbonetto et al (2004); Li and Wang (2003) build a Markov network to represent interactions between blobs (Carbonetto et al, 2004; Li and Wang, 2003), while (Pan et al, 2004) a concept graph (Pan et al, 2004).

The use of pseudo-relevance feedback or any related query expansion mechanisms has been used widely in information retrieval. Several works inspired by cross-lingual retrieval systems were proposed in this direction. In cross-lingual systems, a user generates her query in one language (e.g. English) and the system retrieves documents in another language (e.g. French). The analogy here is to consider the visual feature space as a language constitued of blobs or patches, simply called *visual words*.

Hence, based on query expansions models, Jeon et al (2003) proposed to extend the cross-lingual relevance models to cross-media relevance models. These models were further generalized to continuous features by Lavrenko et al (2003) with non-parametric kernels, while Feng et al (2004) modeled the distribution of words with Bernoulli distributions.

The trans-media relevance model we describe in this chapter (see section 1.4) can also be seen as a cross-media relevance model. The basic idea is to first use one of the media types to gather relevant multimedia information and then, in a second step, use the dual type to perform the final task (retrievel, annotation, etc). These approaches can be seen as an "intermediate level" fusion since the media fusion takes place after a first mono-media retrieval step based on mono-modal similarities (see sections 1.2 and 1.3).

This book chapter is structured in four sections: *visual methods, textual methods, cross-media technique and diversity-focused retrieval*. For each of these sections, we discussed briefly the main algorithms and show a few experimental results. Then, we draw partial conclusions on these methods before moving on to the next family of techniques. The thread of the presentation goes along with the performance of the presented technology: visual methods have generally lower performances than textual ones. Similarly, textual methods are outperformed by cross-media techniques.

| Notation | Description |
|---|---|
| $N$ | Number of documents in the collection |
| $d$ | A document of the collection |
| $d^T, d^V$ | The textual and visual part (image) of $d$ |
| $S$ | A matrix of similarities between documents |
| $S^T, S^V$ | A matrix of text-based or image-based similarities |
| $S^{VT}, S^{TV}$ | A matrix of cross-modal image-text or text-image similarities |
| $q$ | A query |
| $s_q$ | A similarity vector between the query and the documents |
| $s_q^T, s_q^V$ | A text-based and image-based similarity vector |
| $s_q^{VT}, s_q^{TV}$ | A cross-modal image-text and text-image similarity vector |

**Table 1.1** Notations.

Finally, methods adressing diversification of the top results, to offer a better user experience, are built upon the cross-media ones.

For a better following of different sections, in Table 1.1 we summarized our main notations.

## 1.2 Content Based Image Retrieval

Content-based image retrieval (CBIR), also known as query by image content (QBIC) is the application of computer vision to the image retrieval problem, that is, the problem of searching for digital images in large databases based on visual retrieval as opposed to the text or tag based retrieval of images. The term *content-based* means that the search will analyze the visual contents of the image, where content in this context might refer to colors, shapes, textures, or any other piece of information that can be derived from the image itself. The process involves computing a feature vector for the unique characteristics of the image. While in early CBIR systems mainly global features or rather low-level features were used, recent systems tend to extract these features more locally and transfrorm them to some higher level representations. One of the most succesful approach is to transform low level image descriptors to "higher" level descriptors are the popular bag-of-visual word (BOV) representation of the images (Sivic and Zisserman, 2003; Csurka et al, 2004) based on a visual vocabulary built in the low level feature space. When the visual vocabulary is represented by a probability density, the Fisher kernel framework proposed by (Jaakkola and Haussler, 1999) is applicable and the image can be represented by Fisher Vectors as proposed by Perronnin and Dance (2007).

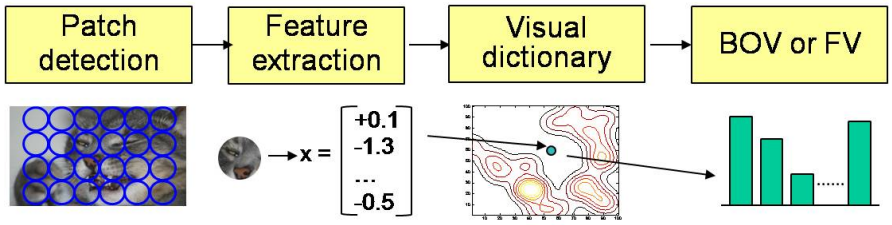


**Fig. 1.1** The main steps to obtain BOV or Fisher Vector representation of images.

### *1.2.1 Fisher Vector Representation of Images*

The Fisher Vector (FV) proposed by Perronnin and Dance (2007), can be seen as an extension of the popular bag-of-visual word (BOV) representation proposed by Sivic and Zisserman (2003); Csurka et al (2004). Both are based on an intermediate representation, *the visual vocabulary* built in the feature space. If a probability density function (in our case a Gaussian Mixture Model) is used to model the visual vocabulary, we can represent an image by the gradient of the log likelihood with respect to the parameters of the model. The Fisher Vector is the concatenation of these partial derivatives and describes in which direction the parameters of the model should be modified to best fit the data (extracted image features). While both (BOV and FV) representations were heavily used for image categorization, they are class independent high level image representations and hence suitable for image retrieval too.

The mains steps to obtain such representations are illustrated in Figure 1.1. First local patches are either detected using interest point detectors, low level image segmentation, or simply regular sampling. Then low-levvel features are computed on those patches such as color, texture, SIFT, shape, etc. In our experiments we sampled patches on regular grids at multiple scales and computed histograms of oriented gradients (HOG) and local color statistics (RGB means and standard deviations). The *Visual Vocabulary* can be built on a set of patches extracted from a randomly selected set of images using e.g. Kmeans, Mean Shift, GMMs or Random Forest. The high-level image signature is computed by accumulating word occurrences (BOV) or by building the Fisher Vectors as described below.

In our case the visual vocabulary is a GMM with parameters $\Phi = \{\omega_m, \mu_m, \sigma_m, i = 1...M\}$ [1] trained on a set of features extracted from images to estimate their distribution in the low-level features space:

$$p(x|\Phi) = \sum_{m=1}^{M} \omega_m p_m(x|\Phi). \tag{1.1}$$

[1] We consider diagonal covariance matrices and we denote by $\sigma_m^2$ the variance vector.

Here each Gaussian component $\mathcal{N}(\mu_m, \sigma_m)$ can be seen as the representation of a visual word and given a new low-level feature $x_l$, the probability that it was generated by the Gaussian $m$ is:

$$\gamma_m(x_l) = \frac{\omega_m p_m(x_l|\Phi)}{\sum_{m=1}^{M} \omega_m p_m(x_l|\Phi)}. \tag{1.2}$$

In the BOV representation of the image, the low-level descriptor $x_l$ is then transformed into a high-level $M$-dimensional descriptor as follows:

$$\gamma(x_l) = [\gamma_1(x_l), \gamma_2(x_l), \ldots, \gamma_M(x_l)] \tag{1.3}$$

To get a global signature (BOV) for an image or more generally the visual part of a document represented by a set of extracted low level image features $d^V = \{x_l, l = 1 \ldots L\}$, we simply average $\gamma(x_l)$ over l.

The Fisher Vector is an alternative to this BOV image representation based on the Fisher kernel framework proposed by Jaakkola and Haussler (1999). The main idea is to consider the gradient vector of the log likelihood according to the parameters of $\Phi$. Assuming that the $x_l$'s were generated independently by $\Phi$, we can write this log likelihood as follows:

$$\log p(d^V|\Phi) = \frac{1}{L} \sum_{l=1}^{L} \nabla_\Phi \log p(x_l|\Phi). \tag{1.4}$$

We consider the gradients of $\log p(x_l|\Phi)$ with respect to the mean and standard deviation parameters (the gradient with respect to the weight parameters brings little additional information) and as suggested by Perronnin and Dance (2007), we further normalize them by the Fisher Information matrix (having a whitening effect on different dimensions):

$$F_\Phi = E_{d^V}\left[ (\nabla_\Phi \log p(d^V|\Phi)) (\nabla_\Phi \log p(d^V|\Phi))^T \right] .$$

In the case of diagonal covariance matrices and an approximation of the Fisher Information matrix we obtain the following closed form formulas (see details in (Perronnin and Dance, 2007)):

$$f_{\mu_m^r}(x_l) = \frac{\sigma_m^r}{\sqrt{\omega_m^r}} \frac{\partial \log p(x_l|\Phi)}{\partial \mu_m^r} = \gamma_m(x_l) \frac{x_l^r - \mu_m^r}{\sigma_m^r \sqrt{\omega_m^r}} , \tag{1.5}$$

$$f_{\sigma_m^r}(x_l) = \frac{\sigma_m^r}{\sqrt{2\omega_m^r}} \frac{\partial \log p(x_l|\Phi)}{\partial \sigma_m^r} = \gamma_m(x_l) \frac{(x_l^r - \mu_m^r)^2 - (\sigma_m^r)^2}{(\sigma_m^r)^2 \sqrt{2\omega_m^r}} . \tag{1.6}$$

where the superscript $r, r = 1 \ldots R$ denotes the $r$-th dimension of a vector and $R$ is the dimensionality of the feature space. The Fisher Vector $f_\Phi(x_l)$ of the observation $x_l$ is the concatenation of all these partial derivatives leading to a $2 * M * R$ dimensional vector. Finally, to obtain the image representation $f_\Phi(d^V)$ we take the average over the Fisher Vectors from all the extracted patches $x_l$, $l = 1..L$.

We define the visual similarity between two visual documents $d_1^V$ and $d_2^V$ using the L1 distance between the L1 normalized Fisher Vectors:

$$S^V(d_1^V, d_2^V) = -\|\tilde{f}_\Phi(d_1^V) - \tilde{f}_\Phi(d_2^V)\|_1 \quad (1.7)$$

where $\tilde{f}_\Phi(d_i^V)$ is $f_\Phi(d_i^V)$ after normalized it to L1-norm equal to 1.

### *1.2.2 Image Retrieval at ImageCLEF Photo Retrieval Tasks*

We used the above described Fisher Vector based image retrieval in our ImageCLEF Photo Retrieval experiments. Actually, as we used two type of low level features, we built two independent visual vocabularies, one for color features (local RGB statistics) and one for texture (orientation histograms). Therefore, before computing the similarity between two images using Equation (1.7) we first concatenated the two Fisher Vectors (texture and color one).

One specificity of the ImageClef Photo Retrieval Challenge, compared to the classical query image based retrieval is that for each topic there is not one, but several query images $q_i^V$, ($i = 1,..M$, where $M$ is generally 3). Therefore, the main question we can ask is how to combine the information from different images to get a better retrieval performance. We investigated three different strategies:

- **I1** : We considered the mean of the $M$ Fisher Vectors (this can be seen as the concatenation of the $M$ set of patches $q_i^V$ into single $q^V$ one) and used this mean Fisher Vector to query the database.
- **I2** : The database images were ranked according to each image independently and the $M$ ranked list was combined using round-robin type selection (i.e. intermixing the $M$ lists) and eliminating the repetitions.
- **I3** : We combine the three similarity scores (with respect to each image of the query) by averaging the scores after Student normalization.

Table 1.2 compares these three strategies on the IAPR TC12 Benchmark Photo Repository used in ImageClef Photo Retrieval 2007 and 2008 sessions. It shows results on the 39 topics of the Year 2008. We can see that the early fusion (mean Fisher Vector) performs worse than late score level fusions. The reason might be that the different query images contained complementary information, and searching for images that are similar to all of them was not the best option. Indeed, late combination of scores allowed for better performances, where the best performance was obtained by the score averaging strategy **I3**.

**Table 1.2** Performances (MAP and P@20) of different strategies for image retrieval.

| Run description | MAP | P@20 |
|---|---|---|
| **I1** | 0.119 | 0.255 |
| **I2** | 0.130 | 0.301 |
| **I3** | **0.151** | **0.328** |

While our method was the best performing CBIR system in both sessions (2007 and 2008), the overall performances are quite poor. These data sets beeing multimodal (containing images with texts), and therefore the main idea was to investigate both text based retrieval (still most commercial image retrieval systems work based on textual retrieval of images) and especially the combination of the two modalities.

## 1.3 Text Representation and Retrieval

As is shown later, our cross media technique relies on a text retrieval system in order to compute a text "similarity" measure. In this section, we briefly summerize the techniques we used during our participations in ImageClef-Photo. Overall, we used state of the art information retrieval methods: language models[2]. The next section will detail the standard language modelling approach to textual information retrieval. We also explored successfully query expansion techniques that are described in Section 1.3.2.

### *1.3.1 Language Models*

First the text is pre-processed including tokenization, lemmatization, word decompounding and standard stop-word removal. Then starting from a traditional bag-of-word representation (assuming independence between words), we adopt the language modeling approach to information retrieval. The core idea is to model a document $d^T$ by using a multinomial distribution over the words denoted by the parameter vector $\theta_d^T$. A simple language model (LM) could be obtained by considering the frequency of words in $d^T$ (corresponding to the maximum likelihood estimator):

$$P_{ML}(w|d^T) = \frac{\#(w,d^T)}{|d^T|} \ .$$

where $\#(w,d^T)$ is the number of occurences of word $w$ in $d^T$ and $|d^T|$ is the lenght of d in tokens. The probabilities should be further smoothed by the corpus language model:

$$P_{ML}(w|C) = \frac{\sum_d \#(w,d^T)}{|C|}$$

using the Jelinek-Mercer interpolation :

$$\theta_{d^T,w} = \lambda \ P_{ML}(w|d^T) + (1-\lambda) \ P_{ML}(w|C) \ . \tag{1.8}$$

[2] Other models to represent the texts such as BM25 and DFR models could also be used in principle without altering significantly our results.

Using this language model, we can define the similarity between two documents using the cross-entropy function:

$$S^T(d_1^t, d_2^t) = \sum_w P_{ML}(w|d_1^t) \log(\theta_{d_2^t,w})) \quad (1.9)$$

### *1.3.2 Text Enrichment at ImageCLEF Photo Retrieval Tasks*

In this section, the different text enrichment techniques, we used during the different sessions, are introduced. In fact, there are several incentives to enrich text associated to images:

- The relative sparsity of the textual representation of the photos. Textual representations of photos are usually short text. At best, they consist of a single paragraph and at worst, images have simply very few tags. Overall, textual annotations of images are shorter than standards documents used in text collections, such as Web document or news articles.
- The gap between the lexical fields of these descriptions and the queries : queries may be expressed in a more abstract way than the factual description of the photos.
- Textual queries are short, often shorter than what is considered *short* for classical information retrieval benchmarks. An image and a short text can be considered as the equivalent of long queries for classical text information retrieval. Thus, queries may need some expansion to exploit associated concepts or words relevant to the queries in order to get a better recall.

In the following, the different text-enrichment mechanisms, used in 2007, 2008 and 2009, are described. In short, *Flickr Related Tags* served to enrich documents in 2007. Then, we experiment document enrichment with the Open Office Thesaurus and visual concepts. Lastly, co-occurence measures between words were used to expand textual queries in 2009.

**Year 2007: Enriching Text with Flickr.** Motivated by the fact that, this year, the textual content of the documents was very poor (text annotations were limited to the <TITLE> fields of documents), we decided to enrich the corpus thanks to the Flickr database [3], at least for texts in English. Flickr API provide a function to get tags related to a given tag [4]. According to Flickr documentation, this function returns *a list of tags "related" to the given tag, based on clustered usage analysis*. It appears that queries, on one hand, and photographic annotations on the other hand, adopt a different level of description. Queries are often more abstract and more general than annotations. As a consequence, it is easier and more relevant to enrich the annotations than the queries : related tags are often at the same level or at the upper (more general) semantic level. Table 1.3 show some example of enrichment terms, related to the annotation corpus. We can observe that the related terms do

[3] http://www.flickr.com/services/api/

[4] http://www.flickr.com/services/api/flickr.tags.getRelated.html

encode a kind of semantic similarity, often towards a more abstract direction, but also contain also noise or ambiguities.

**Table 1.3** Corpus Terms and their related terms from Flickr.

| Corpus Term | Top 5 related Terms |
|---|---|
| Jesus | christ, church, cross, religion, god |
| classroom | school, class,students, teacher, children |
| hotel | lasvegas, building, architecture, night |
| Riviera | france, nice, sea, beach, french |
| Ecuador | galapagos, quito, southamerica, germany, worldcup |

Below, is an example of an enriched document where each original term has been expanded with its top 20 related terms:

DOCNO: annotations/00/116.eng
ORIGINAL TEXT: Termas de Papallacta Papallacta Ecuador
ADDED TERMS: chillan colina sur caracalla cajon piscina snow roma italy maipo thermal nieve volcan argentina mendoza water italia montaa araucania santiago quito southamerica germany worldcup soccer football bird andes wm church fifa volcano iguana cotopaxi travel mountain mountains cathedral sealion market

Enriching the text corpus partially solved the term mismatch but it also introduced a lot of noise in a document. Hence, most of the probabilitic mass of the language model is devoted to the the original text of a document.

**Year 2008: Enriching Text with Visual Concept and Open Office Thesaurus.** In 2008, we investigated the use of external resource in order to enrich text. Another issue that we wanted to address was the use of the visual concepts provided by the organizers as extra "textual words", refining the original textual representation of the photo by higher-level visual information.

The first variant we developed consisted in exploiting the English Open Office thesaurus[5] to enrich the textual description of the photos and/or the queries. Several strategies can be chosen. We chose the following ones:

- Document enrichment: we added all synonyms and broader terms to the terms of the original description, when they are covered by some thesaurus entry. To give more weight to the original terms, they were artificially replicated 15 times.
- Query enrichment: we added all the synonyms and narrower terms to the terms of the original description, when available. To give more weight to the original terms, they were artificially replicated 5 times.

Note that we also simultaneously enriched both the queries and the documents, but this resulted in performance deterioration (too much noise introduced).

As pseudo-relevance feedback (PRF) is another way to do query expansion, we systematically ran experiments with and without pseudo-relevance feedback for each setting (baseline, document enrichment, query enrichment). The top ten terms

[5] Available on *http://wiki.services.openoffice.org/wiki/Dictionaries*

of the top ten documents were used to expand the initial query language model by convex linear combination (coefficient =0.6 for the feedback model). Query model updating was based on the mixture model method (**?**). The performances (Mean Average Precision and Precision@20) are given in Table 1.4.

**Table 1.4** Performances (MAP and P@20) of different text enrichment strategies.

| | Without PRF | | With PRF | |
|---|---|---|---|---|
| Run Description | MAP | P@20 | MAP | P@20 |
| Baseline | 0.215 | 0.259 | 0.239 | 0.293 |
| Document Enrichment | 0.231 | 0.268 | 0.260 | 0.308 |
| Query Enrichment | 0.218 | 0.264 | 0.257 | 0.282 |

It clearly appears that combining document enrichment by thesaurus and query expansion by PRF (using the thesaurus-enriched documents in the first feedback phase) gives the best results. Doing query semantic enrichment followed by PRF (using the thesaurus-enriched query in the first feedback phase) gives slightly worse results. In any case, the use of this external resource is beneficial with respect to a standard PRF query expansion.

The second variant we developed aimed at assessing the benefits of introducing automatically detected visual concepts. These concepts were generated by the two best image categorization systems in the ImageClef Visual Concept Detection Task **?**, from XRCE and RWTH and provided by organizers for the Visual Photo Retrieval Task. Note that the XRCE method used the Fisher Vector image representation as described in Section 1.2.1.

The approach to combine these visual concepts with the text was as follows: we enriched both the documents and the queries with the visual concepts (e.g. indoor, outdoor, building, sky, night, animal, etc.) automatically associated with the images and built language models with the enriched texts. Then we applied our retrieval model as described above. This can be considered as a simplistic way of doing multi-media retrieval. The obtained performances (Mean Average Precision and Precision@20) are given in Table 1.5. We can see clearly, that using the visual concepts increases the retrieval performance. However as shown later, this performance is far below the results we can obtain with cross-media similarity measures (MAP=0.44, P@20=0.57).

**Table 1.5** Performances (MAP and P@20) of the combinations with automatically detected visual concepts.

| | Without PRF | | With PRF | |
|---|---|---|---|---|
| Run description | MAP | P@20 | MAP | P@20 |
| Baseline | 0.215 | 0.259 | 0.239 | 0.293 |
| XRCE Visual Concepts | 0.241 | 0.297 | 0.269 | 0.334 |
| RWTH Visual Concepts | 0.232 | 0.271 | 0.258 | 0.308 |

**Year 2009: Enriching Text by means of Lexical Entailment / Term Similarity.** In 2009, textual queries were very short with typical length of 1 or 2 words. In general, single keyword queries can be ambiguous. Query expansion techniques could help in finding several meanings or different contexts of the query word. As one of the goals was to promote diversity for the Photo Retrieval Task, query expansion methods could help in finding new clusters. In fact, if a term has several meanings or different contexts, the most similar words to this term should partially reflect the diversity of related topics associated to it. The Chi-Square statistics was used to measure the similarity between two words **?**, although any other term similarity measure or lexical entailment measure could be used.

Hence, for each query word $q_w$, we computed the Chi-Square statistics of the latter with all other words (including $q_w$). We kept only the top ten words and divided the scores by the maximum value (given by the inner statistic of $q_w$ with itself). Table 1.6 displays, for some query terms, the most similar terms with the renormalized Chi-Square statistics. To illustrate that co-occurrence measures can handle diversity of word senses, one can look at the most similar terms of the *euro* term. The most similar term bear the notion of lottery, currency or football event, which were all relevant and richer than the themes indicated by the topic images (currency and euro stadium).

**Table 1.6** Query Terms and their most similar terms.

| obama 1 | strike 1 | euro 1 |
|---|---|---|
| barack 0.98 | hunger 0.04 | million 0.05 |
| springfield 0.16 | protest 0.02 | billion 0.05 |
| illinois 0.16 | worker 0.01 | currency 0.03 |
| senator 0.09 | caracas 0.01 | 2004 0.03 |
| freezing 0.08 | led 0.01 | coin 0.02 |
| formally 0.08 | venezuela 0.01 | devil 0.02 |
| ames 0.07 | chavez 0.001 | qualify 0.02 |
| democrat 0.06 | nationwide 0.001 | qualification 0.02 |
| paperwork 0.04 | retaliatory 0.001 | profit 0.01 |

To sum up our models representing texts, we used standard language models to compute what we refer to as *textual similarities*. Over the years, we have also tried to compensate for the relative sparsity of texts, whether documents or queries, with the help of external resources or co-occurence techniques. These enrichment techniques all improved the performance of the monomodal textual system. However, when the image queries are also taken into account, their impact is moderate and depends heavily on the task and the collection.

## 1.4 Text-Image Information Fusion

Understanding the nature of the interaction between text and images is a real scientific challenge that was studied a lot in the last few years. The main difficulty is to overcome the *semantic gap* and, especially, the fact that visual and textual features are expressed at different semantic levels. Here we describe our cross-media similarity measure we developed and succesfully applied in the context of the ImageClef Photo Evaluation forum to the multi-modal Photo Retrieval Task.

The main idea was to go beyond naive techniques that combine the scores from a text retrieval system and from an image retrieval system into a single relevance score called as late fusion approach. We also wanted to avoid the early fusion models that exploiting the correlations between the different modalities are generally quite complex and have shown rather poor performances in the past due to to variations in their level of semantic meaning (words vs. low level image features), and in dimensionality.

Our method was inspired by the *trans-media pseudo feedback* proposed in (**?**Maillot et al, 2006; Clinchant et al, 2007) which is an extention of pseudo-relevance feedback, where the first retrieval step is done in one modality (e.g. textual), then the media type is switched to the other modality (e.g. visual), and the new query process is done in this new modality with a query built with the top retrieved documents in the first step. These models have shown significant improvement on retrieval performance in multi-modal databases (**?**Clinchant et al, 2007; Ah-Pine et al, 2009c).

Cross-media similarities draw their inspiration from the trans-media relevance feedback method. However, instead of extracting words (i.e. features) with a pseudo feedback method to build a new query, cross-media similarities directly combine the mono-modal similarities. They can be understood as a diffusion process of similarities, or as a particular kernel combination. These cross-media similarities described in the next section were at the heart of our runs which we submitted in 2007, 2008 and 2009 and have proven their effectiveness (Clinchant et al, 2007; Ah-Pine et al, 2008, 2009b).

### *1.4.1 Cross-Media Similarities*

This method introduced by Clinchant et al (2007) assumes that two similarity matrices $S^T$ and $S^V$ over the same set of multimedia objects denoted $d_i = (d_i^T, d_i^V); i = 1,\ldots,N$ was precomputed on the database. The former matrix $S^T$ is related to textual based similarities whereas the latter matrix $S^V$ is based on visual similarities and they are both *NxN* matrices. Typically, we used the Equation (1.9) to compute $S^T$ and Equation (1.7) to compute $S^V$, however any other textual or visual similarity can be used. Both matrices were normalized such that the proximity measures distribution of each row varies between 0 and 1.

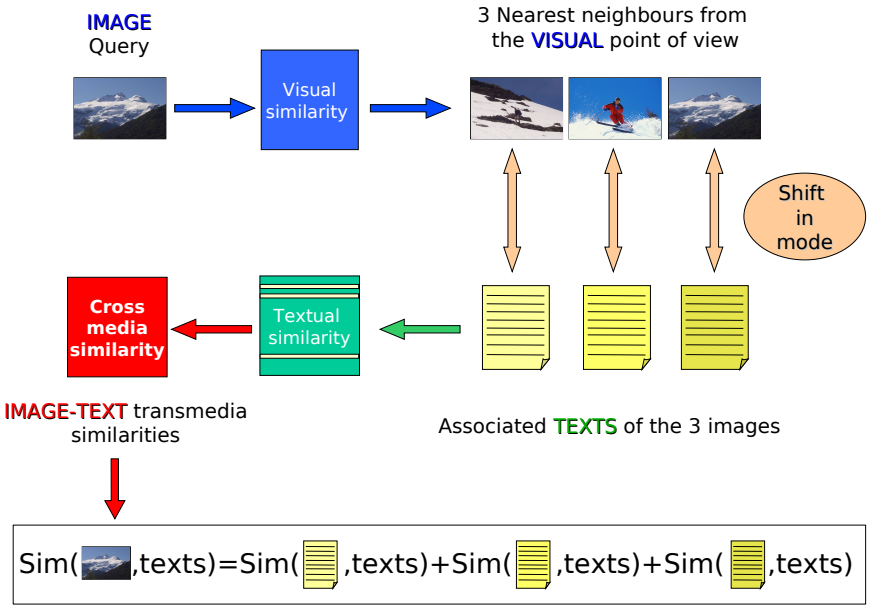


**Fig. 1.2** Illustration of the trans-media pseudo feedback mechanism.

Let us denote by $\kappa(S,k)$ the thresholding function that, for all rows of $S$, puts to zero all values that are lower than the $k^{th}$ highest value and keeps all other components to their initial value.

Accordingly, we define the cross-media similarity matrices that combine two mono-media similarity matrices as follows:

$$S^{VT} = \kappa(S^V, k^V).S^T \tag{1.10}$$

$$S^{TV} = \kappa(S^T, k^T).S^V \tag{1.11}$$

where the . symbol designates the standard matrix product. Note that the number ($k^T$ and $k^V$) of the top highest values according to the textual, respectively visual similarities can be different. This intermediate fusion method can be seen as a graph similarity mixture through a two-step diffusion process, the first step being performed in one mode and the second step being performed in the other one (see (Ah-Pine et al, 2008; **?**) for further details). This method is depicted in Figure 1.2.

Let us precise that in the more specific case of information retrieval, we are given a multimedia query $q$ ($q^T$ denoting the text part and $q^V$ the image part of $q$). In that case, as far as the notations are concerned, we rather have the following cross-media score definition:

$$s_q^{VT} = \kappa(s_q^V, k^V).S^T \tag{1.12}$$

$$s_q^{TV} = \kappa(s_q^T, k^T).S^V \tag{1.13}$$

where $s_q^T$ is the $N$ dimensional similarity row vector of the textual part of the query $q^T$ with a set of multimedia objects (their textual part $d_i^T$) and respectively $s^V$ is the similarity row vector of the visual part of the query $q^V$ with the the same set of multimedia objects (but their image part $d_i^V$).

#### 1.4.1.1 Fusing all Similarities

Cross-media similarities that we have recalled in the previous subsection, attempt to better fill in the semantic gap between images and texts. They allow to reinforce the mono-media similarities. Therefore, the final similarity we used is a late fusion of mono-media and cross-media similarities. This late combination have proved to provide better results according to the results we obtained for the Photo Retrieval Tasks.

The final pairwise similarity matrix that evaluates the proximity relationships between multimedia items of a set of elements is given by:

$$S = \alpha^T S^T + \alpha^V S^V + \alpha^{VT} S^{VT} + \alpha^{TV} S^{TV} \tag{1.14}$$

where $\alpha^T, \alpha^V, \alpha^{VT}, \alpha^{TV}$ are four weights that sum to 1.

Similarly, when we are given a multimedia query, the final relevance score is computed as follows:

$$s_q = \alpha^T s_q^T + \alpha_V s_q^V + \alpha^{VT} s_q^{VT} + \alpha^{TV} s_q^{TV} \tag{1.15}$$

### *1.4.2 Cross-Media Retrieval at ImageCLEF Photo Retrieval Tasks*

#### 1.4.2.1 Year 2007/2008

The two main aspects we want to analyze in this section are on one hand the effect of the number of the selected documents for the trans-media feedback and on the other hand the performance of the cross-modal retrieval compared to mono-modal retrieval. We can notice that the Equation (1.15) is quite general and just by varying the weighting parameters we can easily deduce mono-modal similarities, late fusion or different cross-modal similarities (see Table 1.7). Here the goal being to compare different configurations of the Equation (1.15) given the same $S^T$ and $S^V$ as input, we didn't used exactly the same configuration[6] as in the challenge and hence the

[6] While the same IAPR TC12 Benchmark Photo Repository data was used in year 2007 and 2008, in the session 2007 the image descriptions were not used. In the experiments reported here they

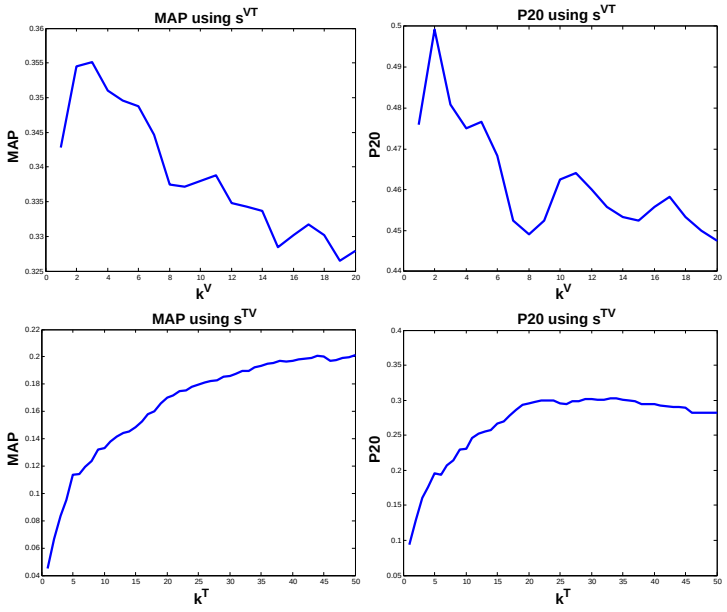


**Fig. 1.3** Performances (MAP and P@20) of $(s^{VT})$ and $(s^{TV})$ with variable $k^V$ and $k^T$.

results are not directly comparable with those reported in (Clinchant et al, 2007, 2008; Ah-Pine et al, 2008, 2009c). Nevertheless they are about the same magnitude and of similar behavior leading to same conclusions.

Before a comparative analyzes of methods, let first analyze the effect of the number of top elements in the cross media similarity measures given by Equations (1.12 and (1.13). Figure 1.3 shows the ranking performances of $(s^{VT})$ and $(s^{TV})$ for variable $k^V$ respectively $k^T$. We can see that while using the top 2 or 3 visually similar images make a big difference using more images decreases the performance. The main reason might be that non relevant top images in $(s^{VT})$ introduces too much textual noise in the pseudo relevance feedback. Concerning the $(s^{VT})$, while the performance vary more smoothly with $k^T$, it is globally lower performing than $(s^{TV})$.

Table 1.7 shows a comparison of the ranking performances using (1.15) with different weighting parameters. The results here are averages over the 60 query topics. In the case of cross modalities we used $k^V = 2$ and $k^T = 25$. Analyzing the table we can see that combining images with text helps both using late fusion approach as in computing cross-media similarity. The only exception was $(s^{TV})$ where it does not seem to help the visual pseudo relevance feedback, probably due to the noise we introduce. The best results were hence obtained when we combine the cross-modal similarity $s^{VT}$ with the late fusion $s^T + s^V$ (showed in the last row of Table 1.7). In

were however used. On the other hand, as the 39 topics of 2008 was a subset of the 2007 topics, here we perform and show average performances over all 60 topics.

**Table 1.7** Comparison of the retrieval performances obtained by the Equation (1.15) with different weighting parameters on the IAPR data set.

| Run description | $\alpha^T$ | $\alpha^V$ | $\alpha^{VT}$ | $\alpha^{TV}$ | MAP | P@20 |
|---|---|---|---|---|---|---|
| Textual $(s^T)$ | 1 | 0 | 0 | 0 | 0.263 | 0.308 |
| Visual $(s^V)$ | 0 | 1 | 0 | 0 | 0.18 | 0.326 |
| Late Fusion | 0.5 | 0.5 | 0 | 0 | 0.348 | 0.45 |
| Cross-$s^{VT}$ | 0 | 0 | 1 | 0 | 0.33 | 0.47 |
| Cross-$s^{TV}$ | 0 | 0 | 0 | 1 | 0.178 | 0.296 |
| Cross-all | 0.25 | 0.25 | 0.25 | 0.25 | 0.387 | 0.457 |
| Cross-$s^T, s^{VT}$ | 0.5 | 0 | 0.5 | 0 | 0.411 | 0.522 |
| Cross-$s^T, s^V, s^{VT}$ | 0.25 | 0.25 | 0.5 | 0 | **0.441** | **0.573** |

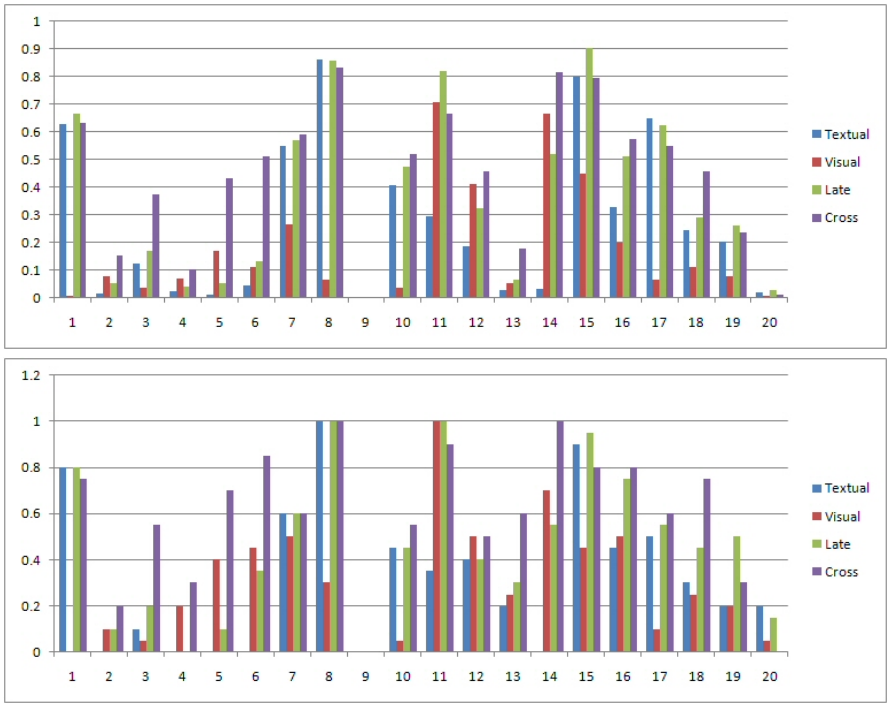


**Fig. 1.4** Performances (MAP and P@20) for the first 20 topics.

Figure 1.4 we show the retrieval performances for different queries. Here for better visibility we illustrate the results only for the first 20 topics.

#### 1.4.2.2 Year 2009

One of the main novelty in 2009 Photo Retrieval Task compared to the previous years was that the new data set contained half million images from Belga News. This was 25 times more than the IAPR TC12 Benchmark Photo Repository, and hence we had to answer some new questions. One of the main one was how the proposed method scales with this new data. In the case of such data, the mono-modal similarity matrices becomes huge (500 000 x 500 000), which requires both high computational cost and storage capacity. Recently, Perronnin (2010) proposed a method to handle large scale retrieval with Fisher Vectors, which unfortunately we didn't had time to test. However, we could test directly the Equation (1.15) on a sub-set of the Belga News data set, namely the 73240 judged images for which relevance scores were provided by the organizers after the challenge.

**Table 1.8** Comparison of the retrieval performances obtained by the Equation (1.15) with different weighting parameters on the Belga News Images data set.

| Run description | $\alpha^T$ | $\alpha^V$ | $\alpha^{VT}$ | $\alpha^{TV}$ | MAP | P@10 |
|---|---|---|---|---|---|---|
| Textual ($s^T$) | 1 | 0 | 0 | 0 | 0.372 | 0.69 |
| Visual ($s^V$) | 0 | 1 | 0 | 0 | 0.012 | 0.146 |
| Late Fusion | 0.5 | 0.5 | 0 | 0 | 0.25 | 0.61 |
| Cross-$s^{VT}$ | 0 | 0 | 1 | 0 | 0.19 | 0.64 |
| Cross-$s^{TV}$ | 0 | 0 | 0 | 1 | 0.012 | 0.152 |

As Table 1.8 shows the performance on this data set being rather poor, neither the late fusion nor the cross-media similarity managed to extract new information from the image to improve the text based retrieval. We have to mention that the image retrieval task from Belga News images is quite different from most CBIR experiments in the literature. The main difference is that the visual similarity between images is in most cases unrelated to the semantic similarity we are seeking. Indeed a large amount of query topic in session 2009 was related to well known personalities. While the image representations (BOV, FV) as described in section 1.2 have shown very good performance when retrieving visual class object, similar scenes, touristic locations they are not suitable to recognize personalities in different circumstances[7]. Indeed, with this unique global image representation two group of people, or two different tennis player in the field will be visually more similar than the photo of the same persons e.g. being interviewed. Hence, the visual similarity alone has real difficulties to correctly retrieve images for most topics in the session 2009 and explains why even the best performing systems in the challenge leaded to so poor performance (MAP= 0.014 with P10=0.15), while pure textual retrieval methods reached a MAP of 0.5 with P10 around 0.8.

[7] Here we mean the BOV or FV constructed on patches extracted on the whole image and not mean the BOV built on specific facial locations of detected faces cannot be used for face recognition as in (Everingham et al, 2006)

As in most cases even the nearest neighbors of images were generally not semantically similar given the topic (except for near duplicates) the fusion methods leaded also to poor performance. However, while the original textual ranking was significantly decreased by the visual pseudo relevance feedback due to added noise, the poor image re-ranking was significantly improved by the textual relevance feedback. Nevertheless, it perform worse than the pure textual ranking.

All these said, we cannot say that the cross-modal similarities were not successfully used in the 2009 session. The main idea to avoid the poor performances shown in 1.8 was to use the text to filter out most non-relevant images. This had the further advantage to reduce significantly the computational and storage cost as instead of computing the entire $S^V$ respectively $S^T$, we only computed small sub-parts of it. Actually, for each topic (or even subtopic) we first selected a set (a few hundred at most) of potentially relevant documents using pure text based search. Then we computed topic dependent mono-modal similarities $S^V_q$ and $S^T_q$ using only the preselected documents. Then we applied successfully (see (Ah-Pine et al, 2009b, 2010) and results in Section 1.5.2) our cross-similarity measures to re-rank those documents based both on visual and textual similarities of documents leading to a best combination of having both precision and diversity at top results. Indeed, diversity seeking was a key issue in the sessions 2008 and 2009, and we will see in the next section that the visual and hence cross-modal similarity had an important role from this point of view too.

## 1.5 Diversity focused Multimedia Retrieval

In the 2008 and 2009 sessions an additional sub-task to multimedia retrieval was asked to be adressed by the participants. It concerns diversity-focused multimedia retrieval and typically, the participants not only needed to provide relevant items to the topics but they also had to promote diversity so that the first retrieved items should be both relevant and thematically different from each other. Diversity-focused retrieval tasks can be encountered in different scenarios. First, we can imagine a user that has a rather general query and providing her with diverse retrieved items in the top-list is very beneficial since she can have a quick overview of the different themes related to her query. Second, we can consider a user that have a text query that is rather ambiguous and thus, she can give some information about the different sub-topics that she wants to retrieve using an image that illustrates each of them. In that case, the system should provide her with a top-list of items that retrieves items that are relevant to all the sub-topics. The first case is the kind of topics that constituted the task in the 2008 session while the second case is rather the type of scenario that was targeted in the 2009 session.

To promote diversity we basically apply a two-step approach. In the first step, we ignore the issue of diversity. In other words, we first try to find the most relevant documents using the material introduced in the previous section. Then, in a

second step, we re-rank the first relevant items by taking into account their mutual similarities in order to avoid redundancy and thus to promote diversity.

Along the two last sessions we tested different methods that we thought would be well-adapted to the kinds of topics and tasks we needed to address. We present each of these methods in subsection 1.5.1. Then, we briefly present in subsection 1.5.2 the main observations we made to evaluate them on the basis of the different runs we submitted.

### *1.5.1 Re-ranking Top-Listed Documents to Promote Diversity*

Among the four methods that we are going to introduce, the first three of them are re-ranking methods that aim at changing the order of the first items of a given top list so that they are not similar to each other according to a given similarity matrix.

The last method, is the Round Robin heuristic. It designs a simple way to combine different lists into one. This approach is used when we want to combine different methods that are assumed to provide different relevant lists to a same topic or when we want to combine different lists that are relevant to several given sub-topics of a topic.

#### 1.5.1.1 Maximum Margin Relevance

"Maximum Margin Relevance" (MMR) proposed by Carbonell and Goldstein (1998) is a re-ranking algorithm which aims at avoiding redundancy among the first elements. It has been successfully applied in different fields such as active learning in information retrieval (Shen and Zhai, 2005; Huang et al, 2008) or in document summarization (Lin et al, 2005; Boudin et al, 2008).

We suppose that we are given a relevance score vector $s_q$ (for a given query $q$) as well as a similarity matrix $S$ (for each pair of documents of the collection). The MMR framework supposes that the elements $d_i$ should be ranked according to both $s_q$ and $S$. It is a greedy algorithm: at each step (rank) $r$ we choose the element $d_i$ that maximizes the following re-ranking criterion:

$$MMR_q(d_i) = \beta(r)s_q(d_i) - (1-\beta(r))\max_{j\in P_r} S(d_i,d_j) \tag{1.16}$$

where $\beta(r)$ is a mixture parameter (between 0 and 1) depending on the rank and $P_r$ is the set of documents already selected (rank lower than $r$).

Traditionally, $\beta$ is kept constant, but we proposed a more efficient variant, where $\beta(r)$ linearly increases between $\beta(1) = \alpha$ ($< 1$) and $\beta(k)$=1 for some $k$ (typically $k$=100), before saturating at value $\beta = 1$.

Regarding the choice of $s_q$, we adopted the (best) combination of mono-media and cross-media similarity measures. For $S$, we can take any similarity matrices

(mono or cross-media) but basically we rely on the similarity matrix defined by Equation (1.14).

#### 1.5.1.2 Clustering Based Re-ranking

We assume here that we are given an ordered top-list of documents $P$ and a similarity matrix $S$ between these items (both $P$ and $S$ could be visual, textual or cross-modal based). $S$ is normalized such that for each row, the maximal element takes the value 1 and the minimal element the value 0. We apply the Relational Analysis (RA) approach for the clustering step in order to find homogeneous themes among the set of items (Marcotorchino and Michaud, 1981; Ah-Pine et al, 2008; Ah-Pine, 2009).

The clustering function that we want to optimize with respect to $X$ is the following one:

$$C(S,X) = \sum_{i,j=1}^{|P|} [S(d_i,d_j) - \underbrace{\frac{1}{|\mathbb{S}^+|} \sum_{(d_i,d_j)\in\mathbb{S}^+} S(d_i,d_j)}_{\text{constant threshold}}] X(d_i,d_j) \tag{1.17}$$

where $X(d_i,d_j) = 1$ if $d_i$ and $d_j$ are in the same cluster and $X(d_i,d_j) = 0$ otherwise and $\mathbb{S}^+$ is the set of pairs of documents which similarity measure is strictly positive: $\mathbb{S}^+ = \{(d_i,d_j) \in P \times P : S(d_i,d_j) > 0\}$.

From Equation (1.17), we can see that the more the similarity between two items exceeds the mean average of strictly positive similarities, the greater the chances for them to be in the same cluster. This clustering function is based upon the central tendency deviation principle proposed by Ah-Pine (2009). In order to find a partition represented by $X$ that maximizes the objective function we used the clustering algorithm described in (Ah-Pine et al, 2008; Ah-Pine, 2009). Notice that this approach doesn't require to fix the number of clusters. This property turns out to be an advantage for finding diverse relevant themes among the documents since we do not know the number of themes for each topic.

After the clustering step, we have to define a re-ranking strategy which takes into account the diversity provided by the clustering results. The main idea of our approach is to represent, among the first re-ranked results, elements which belong to different clusters until a stopping criterion is fulfilled. The strategy employed is described in Algorithm 1.

The stopping criterion in Algorithm 1 we used is related to a parameter denoted $nbdiv \in 1,\ldots,c$, where $c$ is the number of clusters found during the clustering process. It is the maximal number of different clusters that must be represented among the first results. Let us assume that $nbdiv = 10$. Then, this implies that the first 10 elements of the re-ranked list have to belong to 10 different clusters (assuming that $c \geq 10$). Once 10 different clusters are appended, the complementary list (from the $11^{th}$ rank to the $|P|^{th}$ rank), is constituted of the remaining multimedia documents

**Algorithm 1** *Re-ranking strategy for a (sub-)topic*

**Require:** A (sub-)topic $q$, an ordered list $P$ according to some relevance score between $q$ and $P_i; i = 1, \ldots, |P|$ and $R$ the clustering results of objects in $P$.
Let $L1$, $L2$, $L3$ and $CL$ be empty lists and $i = 2$.
Add $P_1$ as first element of the re-ranked list $L1$ and $R(P_1)$ (the cluster id of $P_1$) to the cluster list $CL$
**while** $i \leq |P|$ and Stopping criterion is not fulfilled **do**
  **if** $R(P_i) \in CL$ **then**
    Append $P_i$ to $L2$
  **else**
    Append $P_i$ to $L1$ and add $R(P_i)$ in $CL$
  **end if**
  $i = i + 1$
**end while**
Put if not empty the complementary list of objects from $P_i$ to $P_{|P|}$ in $L3$.
Extend $L1$ with $L2$ then with $L3$ and return $L1$.

sorted with respect to the original list $P$ without taking into account the cluster membership information anymore.

#### 1.5.1.3 Density Based Re-ranking

This approach consists in identifying among a top-list, peaks with respect to some estimated density functions. As a density measure *dens*, we used a simple one which is the sum of similarities (or distances) of the $k$ nearest neighbors. Thus, given an object $d_i$, we define:

$$dens(d_i) = \sum_{d_j \in kNN_i} S(d_i, d_j) \tag{1.18}$$

where $kNN_i$ is the set of the $k$ nearest neighbors of $d_i$ and $S$ is a given similarity matrix which could be the visual-based one given by Equation (1.7) or the text-based one based on Equation (1.9) or cross-media similarities as described by Equation (1.14).

Finally, we re-rank the documents according to this measure by ranking first the items that are the most "dense" and by discarding the near duplicates of these latter elements added to the list.

#### 1.5.1.4 Round Robin

This method is a simple meta-heuristic approach that consists in combining multiple ranked lists into one final list. The main idea is the following one: each ranked list takes its turn (the order of the list is chosen arbitraly) and at each turn we take the top element of the list and we append it to the final list. When a top element of a list is appended to the final list we remove it from its original list and take the next item as the new top element. The new appended documents can belong to other lists, and

if it is the case we remove it from the corresponding lists so that we avoid duplicates in the final list.

The Round Robin method can be applied in the context of different scenarios. First, in the case where we have multimedia topics that are made of several sub-topics for example, we can consider for each of the latter a list of retrieved items and thus combine them by using the Round Robin method. Second, a more general scenario is when we have different systems that give different top-lists that we would like to merge. In that case too, the Round Robin approach can be used to combine the different results.

### *1.5.2 Diversity focused retrieval at ImageCLEF Photo Retrieval Tasks*

As precised previously, the Round Robin method is a kind of meta-heuristic which aims at combining different lists. It is different from the other methods that we introduced previously. The three other approaches rely on the use of a similarity matrix and seeks to re-rank one top-list so that topically-diversed documents are rapidly proposed to the user.

Consequently, the results that are provided by the MMR, the density-based and the clustering based re-ranking methods are comparable to each other though we did not apply all of them to both sessions. On the contrary, they are not directly comparable to the Round Robin technique.

In the following we discuss the comparisons that we made during the two last sessions on the different strategies we used to promote diversity in the multimedia retrieval results.

#### 1.5.2.1 Year 2008

In the 2008 session we mainly applied the MMR and the clustering-based approaches to re-rank a relevant list in order to promote diversity. We recall in Table 1.9 some of the best runs we obtained. The baseline given by the third line is the run provided by Equation (1.15) with parameters $\alpha^T = \alpha^{VT} = 0.5$ and $\alpha^V = \alpha^{TV} = 0$. No re-ranking methods was applied to this run. However, it provides the top-list that we aim at re-ranking in order to avoid redundancy among the first elements. Accordingly, line 1 of Table 1.9 is the run that re-rank the baseline with respect to the clustering-based technique we described previously while line 2 used the MMR method. For both runs the similarity matrix which was used to measure the thematic proximity between documents was the fused cross-media simiarity given by Equation (1.14) with the same aforementioned parameters (see (Ah-Pine et al, 2008) for more details).

We can observe in Table 1.9, that any diversity-focused method fails to increase, on average, the P@20 measure. However, any method performs better than the ba-

**Table 1.9** XRCE's best runs in the 2008 session in terms of Precision at 20 (P@20) and Cluster Recall at 20 (CR@20).

| Run Description | CR@20 | P@20 |
|---|---|---|
| With clust-based re-rank. (using cross-media similarities) | 0.4111 | 0.5269 |
| With MMR re-rank. (using cross-media similarities) | 0.4015 | 0.5282 |
| Without any re-ranking method (baseline) | 0.3727 | 0.5577 |

sic run regarding the CR@20 measure. In other words, by trying to eliminate redundancy among the first retrieved objets, unfortunately, we might push relevant objects out of the 20 first re-ranked elements and on the contrary, we might put into this final top list some irrelevant objects.

In (Ah-Pine et al, 2009c), we analyzed the behavior of the MMR and the clustering-based re-ranking methods and refer the reader to this paper for more details. Here, we briefly underline the main observation that we made by looking at the assessment measures per query. The clustering-based strategy exhibits a consistent, stable behaviour, where it systematically gives slightly lower or equal P@20 performances than the basic list, while offering CR@20 performances that are superior or equal to the baseline. The MMR method does not offer such a stability in its behaviour. In fact, this method seems to take more risk in the re-ranking process with a diversity seeking goal than the former method, with a consequence of increased variance in the performances.

Therefore, despite comparable P@20 and CR@20 measures, the MMR technique and the clustering-based methods do not show the same behaviour.

#### 1.5.2.2 Year 2009

In the 2009 session we applied the density-based, the clustering-based and the Round Robin methods.

In this session many topics were constituted of several sub-topics which basically expressed different aspects of the main topic and gave the participants the definition of the different clusters to retrieve (the topic *"brussels"* for example had sub-topics *"brussels airport"*, *"police brussels"*, ‘*fc brussels"* among others). Those cases represent the topics in part 1. In this case, we treated each sub-topic as if they were independent and combined them using the Round Robin method so as to produce a single list of retrieved diverse items. The method we used to produce the top-list for each sub-topic is described in (Ah-Pine et al, 2009b). It is important to underline the fact that we first used a text-based retrieval for all sub-topics using the images' caption. In other words, the results we are going to mention used a pre-filtering step which aimed at determining a preliminary set of relevant documents from a textual standpoint. After this first pass, we then used different types of similarity in order to re-rank the documents of this preliminary set by taking into account either text (in that case there is no re-ranking) or visual or cross-media similarities. We refer the reader to (Ah-Pine et al, 2009b) for more details.

**Table 1.10** XRCE's runs in the 2009 session on topics of part 1 in terms of Precision at 10 (P@10), Cluster Recall at 10 (CR@10) and F1

| Run Description | CR@10 | P@10 | F1 |
|---|---|---|---|
| Text pre-filt. (captions as queries) | **83.9** | 78.4 | 81.0 |
| Text pre-filt. (captions as queries) + visual re-rank. | 75.2 | 60.8 | 67.2 |
| Text pre-filt. (captions as queries) + cross-modal re-rank. | 83.7 | **79.6** | **81.6** |

Since the Round Robin method is the only strategy that we used to combine different lists into one, we cannot sketch any results analysis about this technique. However, we can comment the results we obtained focusing on the media that performed well. Accordingly, in the case of topics in part 1, we found that the media that gave the best results regarding the diversity assessment measure is the one based on text only. Nevertheless, the F1 measure that combines both the precision and the diversity criteria is better for the fused cross-media similaries as given by Equation (1.14) with the parameters $\alpha^T = 5/12$, $\alpha^V = \alpha^{VT} = 1/4$ and $\alpha^{TV} = 1/12$. In brief, text-based retrieval is far the most important tool to achieve good performances in the multimedia task designed for the 2009 session. Re-ranking the documents using the visual similarities after a text-based pre-filtering doesn't increase the results. However, combining visual and textual similarities using our cross-media techniques and re-ranking the documents with respect to the fused cross-media similairy after the text-based pre-filtering, allows us to slightly improve the P@10 and F1 measures without hurting the CR@10. Those observations are numerically illustrated in Table 1.10.

If topics in part 1 were already well-detailed from a diversity viewpoint since we were provided with their different sub-topics, topics in part 2 were more challenging when seeking to promote diversity. Indeed, in that case, we were only given a text query and three different images. That type of multimedia topic is the kind of topics we had to deal with in the 2008 session. For topics in part 2, we assumed that the three image queries represented three different sub-topics though it was specified that there might be more clusters to find than those three. We computed for each of them a basic top-list (in a similar way we did for topics in part 1, see (Ah-Pine et al, 2009b) for more details) and to each of the top-list we applied a density-based or a clustering-based re-ranking technique before fusing them by means of the Round Robin method. Those types of run are denoted basic runs in (Ah-Pine et al, 2009b).

Regarding the comparison between the two re-ranking techniques on the basis of basic runs assessment measures, we observed on the one hand that density and clustering are comparable when image similarity is used, but on the other hand, with text similarity or fused cross-media similarity, clustering generally gives better results.

Another important observation that we drew out of the different experiments we made is that combining different types of basic runs by means of the Round Robin heuristic, allows us to significantly enhance the retrieved list. This is depicted in Table 1.11 where we can see that combining two different basic runs can lead to more than a 7 point increase in terms of F1 measures.

**Table 1.11** Some of XRCE's best runs in the 2009 session on topics in part 2.

| Run id | Run Description | CR@10 | P@10 | F1 |
|---|---|---|---|---|
| 1 | Text pre-filt. (captions as queries) + cross-modal re-rank | 76.8 | 72.4 | 74.6 |
| 2 | Text pre-filt. (enriched query) + Clust-based re-rank. (vis. sim.) | 65.8 | 78.0 | 71.4 |
| 3 | Text pre-filt. (enriched query) + Dens-based re-rank. (vis. sim.) | 62.6 | **83.2** | 71.4 |
| 4 | Round Robin of 1 and 2 | 82.4 | 78.8 | 80.6 |
| 5 | Round Robin of 1 and 3 | **82.5** | 81.6 | **82.0** |

## 1.6 Conclusion

As a conclusion, we would like to underline the main lessons that we learned along three participations in the Photo Retrieval tasks of ImageCLEF:

- When dealing with multimedia text-image documents, it is very beneficial to combine the text information with the visual information. A simple combination strategy such as the late fusion approach already allows one to obtain much better results than mono-media based retrieval.
- The cross-media technique that we designed as a way to combine multimedia information allowed us to perform better than the late fusion approach. The very good performances we reached in the three last sessions of ImageCLEF show the efficiency and robustness of this method. We believe that it allows one to better handle the semantic gap between different media.
- Text-based retrieval is fundamental as long as we have a good textual description of the images. It performed much better than the visual-based retrieval and for the 2009 session, we could have never been able to obtain such good results if we had not used the text as a pre-filtering before using cross-media techniques. However, visual similarities allow us to significantly gain in terms of precision and recall providing that we combine them with the text similarities in an efficient way.
- When using the fused cross-media similarities as given by (1.15), we consistently observed that one should give more weight to textual similarities than to visual similarities if the former performs much better than the latter, otherwise the equal weigthing works pretty well. Furthermore, generally the image-text cross-media similarities performs better than the text-image cross-media ones, and in most cases it is better not to consider it (null weight). We can see that the cross-media image-text is really beneficial in our strategy allowing to better bridge the gap between image and text. Finally, it is also important to mention that our cross-media similarities are dependent on a parameter $k$, the number of nearest neighbors considered by the pseudo-relevence feedback. Generally, considering a relatively low $k$ (typically $< 5$), we avoid the risk to introduce too much noise in the cross-media similarity. This is particularly true for $(s^{VT})$ while the effect of this number seems to be smoother in the case of $(s^{TV})$.
- As far as the text-based retrieval is concerned, we used standard language models to compute textual similarities. In order to overcome the issue with the sparsity

of texts data and particularly in the context of ImageCLEF collections, we tried to enrich both the queries and the documents by using external resources or co-occurrences techniques. Despite the fact that we used different approaches, we showed that enriching texts data is always beneficial.

- Regarding the diversity seeking retrieval sub-task, we applied a two-step scheme which first focus on retrieving relevant documents and then, re-rank the top-list in the goal of avoiding redundancy among the very first showed items. This approach allowed us to indeed promoting diversity without hurting too much the relevance of the re-ranked top-list. We used different approaches as re-ranking techniques and each of them gave interesting results while presenting different behaviors. The clustering-based methods showed a stable behavior enabling us to consistently improve the diversity assessment measures while decreasing a little bit the precision measures. The MMR method "takes more risk" and if on average it allows an increase of the diversity assessment measure, it also presents a less stable behavior since we observed more variablility of the evaluation measures at the level of queries. The density-based approach also provided good results and particularly when it is applied on the visual similarities.
- In the last session, we further investigated the combination of several methods which resulted from different techniques either at the level of the features we used for mono-media similarities or at the level of the similarities used to locally re-rank a top-list to favour diversity. It appeared that as simple combination methods such as the Round Robin technique generally allow one to improve both the precision and the diversity. Therefore, using different text representations or different enrichment techniques or different similarties to re-rank objects; and combining the resulted top-lists using the Round Robin method is beneficial.
- While in 2007 and 2008, the collection was of around 20K multimedia documents, in 2009, it was constituted of more than 500K items. In last year session, we thus had to deal with more scalability issues than for previous sessions. Indeed, it is not easy to compute the whole visual similarity matrix and an "online" method had to be designed. As mentioned previously, we first applied a text-based pre-filtering step. This strategy turned out to be a winning one since not only we were able to adress the scalability issue of computing visual similarities by pre-selecting a relevant set of documents given a topic, but this text-based pre-filtering was also an efficient way to obtain a very good baseline retrieval that we were able to improve further in a second step.

**Acknowledgements** This work was partially supported by the European Projects PinView FP7-216529 and Sync3 FP7-231854, and the French projects Omnia ANR-06-CIS6-01 and Fragrances ANR-08-CORD-008.

## References

Ah-Pine J (2009) Cluster analysis based on the central tendency deviation principle. In: Proceedings of ADMA

Ah-Pine J, Cifarelli C, Clinchant S, Csurka G, Renders J (2008) Xrce's participation to imageclef 2008. In: Working Notes of the 2008 CLEF Workshop, Aarhus, Denmark

Ah-Pine J, Bressan M, Clinchant S, Csurka G, Hoppenot Y, Renders J (2009a) Crossing textual and visual content in different application scenarios. Multimedia Tools Appl 42(1):31–56, DOI http://dx.doi.org/10.1007/s11042-008-0246-8

Ah-Pine J, Clinchant S, Csurka G, Liu Y (2009b) XRCE's participation to ImageCLEF 2009. In: Working Notes of the 2009 CLEF Workshop, Crete, Greece

Ah-Pine J, Csurka G, Renders JM (2009c) Evaluation of diversity-focused strategies for multimedia retrieval. In: Evaluating Systems for Multilingual and Multimodal Information Access, vol LNCS 5706

Ah-Pine J, Clinchant S, Csurka G (2010) Comparison of several combinations of multimodal and diversity seeking methods for multimedia retrieval. In: Multilingual Information Access Evaluation

Barnard K, Duygulu P, D Forsyth DB N de Freitas, Jordan M (2003) Matching words and pictures. J of Machine Learning Research 3

Blei D, Michael, Jordan MI (2003) Modeling annotated data. In: ACM SIGIR

Boudin F, El-Bèze M, Torres-Moreno J (2008) A scalable MMR approach to sentence scoring for multi-document update summarization. In: COLING

Carbonell J, Goldstein J (1998) The use of MMR, diversity-based reranking for reordering documents and producing summaries. In: SIGIR

Carbonetto P, de Freitas N, Barnard K (2004) A statistical model for general contextual object recognition. In: ECCV

Clinchant S, Renders J, Csurka G (2007) Xrce's participation to imageclef 2007. In: Working Notes of the 2007 CLEF Workshop, Budapest, Hungary

Clinchant S, Renders JM, Csurka G (2008) Trans-media pseudo-relevance feedback methods in multimedia retrieval. In: Advances in Multilingual and Multimodal Information Retrieval, vol LNCS 5152

Csurka G, Dance C, Fan L, Willamowski J, Bray C (2004) Visual categorization with bags of keypoints. In: ECCV Workshop on Statistical Learning for Computer Vision

Duygulu P, Barnard K, de Freitas J, Forsyth D (2002) Object recognition as machine translation :learning a lexicon for a fixed image vocabulary. In: ECCV

Everingham M, Sivic J, Zisserman A (2006) "hello! my name is... buffy" – automatic naming of characters in TV video. In: BMVC

Feng S, Lavrenko V, Manmatha R (2004) Multiple bernoulli relevance models for image and video annotation. In: CVPR

Huang T, Dagli C, Rajaram S, Chang E, Mandel M, Poliner G, Ellis D (2008) Active learning for interactive multimedia retrieval. In: IEEE

Iyengar G, Duygulu P, Feng S, Ircing P, Khudanpur S, Klakow D, Krause M, Manmatha R, h Nock, Petkova D, Pytlik B, Virga P (2005) Joint visual-text modeling for automatic retrieval of multimedia documents. In: Proceedings of ACM Multimedia

Jaakkola T, Haussler D (1999) Exploiting generative models in discriminative classifiers. In: Advances in Neural Information Processing Systems 11

Jeon J, Lavrenko V, Manmatha R (2003) Automatic image annotation and retrieval using cross-media relevance models. In: ACM SIGIR

Lavrenko V, Manmatha R, Jeon J (2003) A model for learning the semantics of pictures. In: NIPS

Lavrenko V, Feng S, Manmatha R (2004) Models for automatic video annotation and retrieval. In: ICASSP

Li J, Wang JZ (2003) Automatic linguistic indexing of pictures by a statistical modeling approach. PAMI 25:9

Lin Z, Chua T, Kan M, Lee W, Qiu L, Ye S (2005) NUS at DUC 2007: Using evolutionary models of text. In: Document Understanding Conference

Maillot N, Chevallet JP, Valea V, Lim JH (2006) Ipal inter-media pseudo-relevance feedback approach to imageclef 2006 photo retrieval. In: CLEF 2006 Working Notes

Marcotorchino J, Michaud P (1981) Heuristic approach of the similarity aggregation problem. Methods of operation research 43:395–404

Monay F, Gatica-Perez D (2004) Plsa-based image auto-annotation: Constraining the latent space. In: ACM MM

Mori Y, Takahashi H, Oka R (1999) Image-to-word transformation based on dividing and vector quantizing images with words. In: In MISRM'99 First International Workshop on Multimedia Intelligent Storage and Retrieval Management

Pan J, Yang H, Faloutsos C, Duygulu P (2004) Gcap: Graph-based automatic image captioning. In: CVPR Workshop on Multimedia Data and Document Engineering

Perronnin F (2010) Large-scale image retrieval with compressed fisher vectors. In: CVPR

Perronnin F, Dance C (2007) Fisher kernels on visual vocabularies for image categorization. In: CVPR

Shen X, Zhai C (2005) Active feedback in ad hoc information retrieval. In: SIGIR

Sivic JS, Zisserman A (2003) Video google: A text retrieval approach to object matching in videos. In: ICCV, vol 2

Vinokourov A, Hardoon DR, Shawe-Taylor J (2003) Learning the semantics of multimedia content with application to web image retrieval and classification. In: Fourth International Symposium on Independent Component Analysis and Blind Source Separation

Zhai C, Lafferty JD (2001) Model-based feedback in the language modeling approach to information retrieval. In: CIKM, pp 403–410