XRCE's Participation to ImageCLEFphoto 2007

Stephane Clinchant, Jean-Michel Renders and Gabriela Csurka
Xerox Research Centre Europe, 6 ch. de Maupertuis, 38240 Meylan, France
FirstName.LastName@xrce.xerox.com

Abstract

Our participation to ImageCLEFphoto07, for the first time, was motivated by assessing several transmedia similarity measures that we recently designed and developed. The object of investigation consists here in some "intermediate level" fusion approaches, where we use some principles coming from pseudo-relevance feedback and, more specifically, use transmedia pseudo-relevance feedback for enriching the mono-media representation of an object with features coming from the other media. One issue that arises when adopting such a strategy is to determine how to compute the mono-media similarity between an aggregate of objects coming from a first (pseudo-)feedback step and one single multimodal object. We propose two alternative ways of adressing this issue, that result in what we called the "transmedia document reranking" and "complementary feedback" methods respectively.

This year, with a "lightly" annotated corpus of images, it appears that mono-media retrieval performance is more or less equivalent for pure image and pure text content (around 20% MAP). Using our transmedia pseudofeedback-based similarity measures allowed us to dramatically increase the performance by ∼50% (relative). Trying to model the textual "relevance concept" present in the top ranked documents issued from a first (purely visual) retrieval and combining this model with the textual part of the original query turns out to be the best strategy, being slightly superior to our transmedia document reranking method. Enriching the image annotations by extra tags extracted from an external resource (namely the Flickr database) does not offer a significant advantage in the ImageCLEF07 corpus, even if we observed an improvement using other multimedia corpora and query sets. From a cross-lingual perspective, the use of domain-specific, corpus-adapted probabilistic dictionaries seems to offer better results than the use of a broader, more general standard dictionary. With respect to the monolingual baselines, multilingual runs show a slight degradation of retrieval performance (∼6 to 10% relative).

Categories and Subject Descriptors

H.3 [**Information Storage and Retrieval**]: H.3.1 Content Analysis and Indexing; H.3.3 Information Search and Retrieval; H.3.4 Systems and Software; H.3.7 Digital Libraries; H.2.3 [**Database Managment**]: Languages—*Query Languages*

General Terms

Measurement, Performance, Experimentation

Keywords

Cross-media and cross-lingual information retrieval, Trans-media relevance feedback

1 Introduction

Efficient access to multimedia information requires the ability to search and organize the information. While, the technology to search text has been available for some time - and in the form of web search engines is familiar to many people - the technology to search images and videos is much more challenging. Early systems were based mainly on visual similarity with a query image making use of lower-level features like texture, color, and shape. The only visual-based approach to retrieval has several drawbacks. It does not actually bridge the semantic gap but rather forces the user to work on low-level feature space. A gap remains between the user's conceptualization of a query and the query that is actually specified to the system.

The scientific challenge is to understand the nature of interaction between text and images: How can a text be associated with a piece of image (and reciprocally an illustrative image with text) ? How can we organize and access text and image repositories in a better way than naive late fusion techniques ? A lot of attempts have tried to find correlation between image and text features. The main difficulty is to overcome the *semantic gap* and, especially, the fact that visual and textual features are expressed at different semantic levels. As far as the "pure" visual mode is concerned, features associated to images have become more and more complex, trying to abstract their representation and to bridge the semantic gap. These trials take at least two directions. The first one is to adopt the strategy of developing more expressive visual vocabularies, potentially hierarchical, relying on latent semantic extraction techniques such as Probabilistic Semantic Analysis (PLSA) or Latent Dirichlet Allocation (LDA). The second one focus on segmentation approaches, that aim at cutting out an image into several related semantic regions.

Departing from the classical "late fusion" strategy, recent approaches have considered fusion at the feature level, estimating correspondences or joint distributions between components across the image and text modes from training data. The main idea is to enrich the images with textual data (annotations) — and vice versa – in order to facilitate their retrieval. These methods could be classify into three general families: latent variable models, graph models and cross-lingual models. Latent variable models generally extend PLSA or LDA to explain jointly image and text (e.g. [17, 1]). Graph models consider the structure of an image through a graph, e.g. with a markov network [2] or a concept graph (e.g. [18]). Finally, cross-lingual models find their inspiration in machine translation and cross lingual information retrieval (e.g. [12]).

Our work is most aligned with the third family, namely Cross-Lingual Models (CLM). The main idea of CLM applied to hybrid text-image retrieval, is to consider the visual feature space as a language constitued of blobs or patches, that we will simply call *visual words*. Unlike [5] which was inspired by machine translation, Jeon et al [12] proposed to extend cross-lingual relevance models to cross-media relevance models. This method estimates the joint probability distribution of blobs that could appear in image and words that could appear in the caption of the image assuming mutual independence between a word and the blobs given an image. These joint probabilities can be used in two ways to annotate/retrieve images. They further showed in [14] that working directly with continuous features describing the patch instead of working on discrete visual words was even better. Further extensions of this model are: an improved normalization in [13] and a Bernouilli distribution for text in Feng et al [7].

Relevance Models are a family of pseudo-feedback methods: given a query q and a database of objects o, relevance models first compute the probability $P(q|o)$ and then automatically enhance, enrich the query with textual or visual features. These models, when extended to mixed modalities, can be considered as the ancestors of methods proposed last year by [16] and [3] for instance. These last models can be called *intermedia feedback*, or *transmedia feedback* techniques. They do not rely on relevance models but act in the same spirit. For example, from a query image, a first visual similarity is computed and an initial set of (assumed) relevant objects is retrieved. As the object are multimodal, each image has also a text, and this text can feed any 'text' feedback method (others than relevance models). In other words, the modality of data is switched , from image to text or text to image, during the (pseudo) feedback process. In that sense, transmedia techniques generalize the pseudo-feedback idea present in cross-media relevance model, but are freed from the particular textual and/or visual models proposed by cross-media relevance model.

In the remaining of this report, we will first describe our mono-media similarities. The next section will explain the different cross-media similarity models we developed for ImageCLEFphoto 2007. Finally, we will describe our official runs and conclude.

2 Monomedia Similarities

2.1 Text

2.1.1 Cross-Entropy measure of similarity

Starting from a traditional bag-of-word representation of pre-processed texts (here, preprocessing includes tokenization, lemmatization, word decompounding and standard stopword removal), we adopt the language modeling approach to information retrieval as the basis of our asymmetric similarity measure:

- Let $p(w|q)$ be the multinomial language model of a text query q (obtained by maximum likelihood estimates, i.e. by simple counting and normalisation).
- Let $p(w|d)$, the multinomial language model of a document d. Documents language models are smoothed via a Jelinek-Mercer Method (other schemes are applicable, such as Dirichlet Prior or Absolute Discounting) :

$$p(w|d) = \alpha p^{MLE}(w|d) + (1-\alpha)\ p(w|Corpus) \tag{1}$$

  where $p^{MLE}(w|d)$ (resp. $p(w|Corpus)$) is simply the ratio of the number of occurrences of w in the textual object d (resp. in the global corpus) to the total document (resp. corpus) length in words.

The cross-entropy function is used as out textual similarity measure:

$$sim_{txt}(q,d) = CE(q|d) = \sum_w p(w|q)\log(p(w|d)) \tag{2}$$

This is obviously an asymmetric similarity measure. It can be trivially generalised to define the similarity between two textual objects d_1 and d_2:

$$sim_{txt}(d_1,d_2) = CE(d_1|d_2) = \sum_w p^{MLE}(w|d_1)\log(p(w|d_2)) \tag{3}$$

2.1.2 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 [8], at least for texts in English. Flickr API provide a function to get tags related to a given tag [9]. 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 the 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 rather than the queries : related tags are often at the same level or at the upper (more general) semantic level. Table 1 show some example of enrichment terms, related to the annotation corpus. We can observe the the related terms does encode a kind of semantic similarity, often towards a more abstract direction, but contains also some noise or ambiguity.

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

Table 1: 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

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. In the Language Modelling framework, the enriched terms acts as a smoothing methods: we give more weight to terms from the original text than for those added, by linear interpolation between the original document language model and the word profile derived from Flickr.

Note that this kind of semantic enrichment was done only for English documents (even if some words in other languages are automatically, and erroneously, added). As we decided to investigate the bilingual case as well (choosing German as the second language), we also built probabilitic translation matrices (ENG - GER) from standard alignment method (*Giza++*) using the small set of parallel sentences that we were allowed to exploit in the ImageCLEFPhoto 2007 corpus. It appeared that the bilingual lexicons automatically extracted from this parallel corpus provided better results than broader, but more general, standard dictionaries. It is worth to emphasize the fact that such automatically extracted lexicons have often the extra advantage to realize some semantic smoothing by side effect: related — but not strict — translations are often derived as potential candidates. Probabilistic translation dictionaries are applied on the source language models of the query to give the new target language models of the query by matrix product.

2.2 Image

The image similarity was defined from a continuous vectorial representation of the image, obtained as follows. Image patches are first extracted on regular grids at 5 different scales with a ratio of $\sqrt{2}$ between two consecutive scales. Two types of low-level features are used: grey-level SIFT-like features [15] and color features. In both cases the image patch is subdivided in $4 \times 4 = 16$ subregions. SIFT-like features are then computed as gradient orientation histograms (8 bins) collected on each subregion leading to a vector of 128 dimentions. Color features are simple means and standard deviations of the 3 RGB channels in the same subregions, which leads to a 96 dimensional feature vector. The dimensionality of both type of features are subsequently reduced down to 50 using principal component analysis (PCA).

Then, some kind of Gaussian Mixture Model (GMM) clustering [6, 20] is performed to build a visual vocabulary [21, 4] of low-level image features where each Gaussian component models a visual word (each one is characterized by $\lambda = \{w_i, \mu_i, \sigma_i,\ i = 1...N\}$ where w_i, μ_i and σ_i denote respectively the weight, mean vector and covariance matrix of Gaussian i).

Two visual vocabularies are built: one is based on texture, the other on color. Both of them have a dictionary size of 64 (meaning that we have 64 Gaussian components fore each).

Finally, we represent each image with a Fisher Kernel based normalized gradient vector as proposed in [19]. The main idea is that given a generative model (here the Gaussian Mixture

Model) with parameters λ, one can compute the gradient vector of each sample I :

$$\nabla_\lambda \log p(I|\lambda) \tag{4}$$

Intuitively, the gradient of the log-likelihood describes the direction in which parameters should be modified to best fit the data. It transforms a variable length sample I into a fixed length vector whose size is only dependent on the number of parameters in the model.

Before computing a similarity measure between images, each vector is first normalized using the Fisher Information matrix F_λ, as suggested in [11]:

$$F_\lambda = E_X \left[\nabla_\lambda \log p(X|\lambda) \nabla_\lambda \log p(X|\lambda)'\right] \ . \tag{5}$$

The normalized gradient vector, simply called Fisher vector, is thus given by:

$$\mathbf{f} = F_\lambda^{-1/2} \nabla_\lambda \log p(X|\lambda) \ . \tag{6}$$

See [19] for closed form approximations of $F_\lambda^{-1/2}$.

The Fischer vectors for color and texture respectively are then simply concatened.

To compute the similarity measure between images I and J, we simply use the the L1-norm of the difference between the Fisher vectors:

$$sim_{Img}(I, J) = norm_{max} - ||\mathbf{f}_I - \mathbf{f}_J|| = norm_{max} - \sum_i |f_I^i - f_J^i| \tag{7}$$

where f^i are the elements of the vector $\mathbf{f}$ and $norm_{max} = 2$.

3 Cross-Media Similarities

We want to define cross-media similarity measures that are more elaborated — and, hopefully, more efficient — than simple late fusion approaches. What we want to investigate here is some "intermediate level" fusion approaches, where we use some principles coming from pseudo-relevance feedback and, more specifically, use transmedia pseudo-relevance feedback for enriching the mono-media representation of an object with features coming from the other media. Our basic material is a multimedia (text + image) database $\mathcal{O}$: in other words, we assume to have at our disposal a collection of images with associated text (for each image) or, in a dual view, a collection of texts illustrated with one image (for each textual element).

3.1 Transmedia Document Reranking Approach

The main idea is the following: for a given image i, consider as new features the (textual) terms of the texts associated to the most similar images (from a purely visual viewpoint). We will denote this neighbouring set as $N_{img}(i)$. Its size is fixed a priori: this is typically the top$N$ objects returned from a retrieval system using equation 7 as ranking measure. Then we can compute a new similarity with respect to any multimodal object j of the collection $\mathcal{O}$ as the textual similarity of this new representation of the image i with the textual parts of j.

We still need to be more precise on how we compute the mono-media similarity between an aggregate of objects $N_{img}(i)$ and one single multimodal object. There are three families of approaches:

1. aggregating $N_{img}(i)$ to form a single object (typically by concatenation) and then compute standard similarity between two objects;
2. aggregating all similarity measures (assuming we can) between all possible couple of objects
3. use a method of pseudo feedback algorithm (for instance Rochhio's algorithm) to extract relevant, meaningfull features of an aggregate and finally use a mono-media similarity.

The first approach we propose belongs to the family 2 (using a simple sum or, equivalently, an arithmetic average to aggregate the individual similarity measures). The next section (Complementary Feedback) will propose an alternative approach that belongs to family 3.

More formally, if we denote by $\mathcal{T}(u)$ the text associated to multimodal object u and by $\hat{T}(i)$ the new textual representation of image i, then the new cross-media similarity measure w.r.t. the multimodal object j is:

$$sim_{ImgTxt}(i,j) = sim_{txt}(\hat{T}(i), \mathcal{T}(j)) = \sum_{d \in N_{img}(i)} sim_{txt}(\mathcal{T}(d), \mathcal{T}(j)) \tag{8}$$

where sim_{txt} is typically defined by equation 3 (e.g. the one based on Language Modelling, even if it is assymetric).

This method can be seen as a reranking method. Suppose that q is some image query; if $T(d)$ is the text of an image belonging to the initial feedback set $N_{img}(q)$, then the rank of the own neighbors of $T(d)$ in the textual sense will be increased, even if they are not so similar from a purely visual viewpoint. In particular, this allows to define a similarity between a purely image query and a simple textual object without visual counterpart. To sum up, the main idea of our method amounts to (i) perform an initial retrieval step to identify $N_{img}(q)$, (ii) to switch mode and to virtually make several queries (one for each element in $N_{img}(q)$ instead of one) and combining them afterwards. Due to renormalization effects and smoothing methods, the resulting ranking function is different from the one obtained by considering the simple concatenation of text in step (ii), since the considered models are not linear. Lastly, the values $sim_{txt}(\mathcal{T}(u), \mathcal{T}(v))$ can be pre-computed in a matrix of textual similarities between all pairs of objects in the multimedia database $\mathcal{O}$, if the corpus is of reasonable size

By duality, we can define another cross-media similarity measure: for a given text i, we consider as new features the Fisher vectors of the images associated to the most similar texts (from a purely textual viewpoint) in the multimodal database. We will denote this neighbouring set as $N_{txt}(i)$. If we denote by $\mathcal{I}(u)$ the image associated to multimodal object u and by $\hat{I}(i)$ the new visual representation of text i, then the new cross-media similarity measure is:

$$sim_{TxtImg}(i,j) = sim_{img}(\hat{I}(i), \mathcal{I}(j)) = \sum_{d \in N_{txt}(i)} sim_{img}(\mathcal{I}(d), \mathcal{I}(j)) \tag{9}$$

where sim_{img} is typically defined by equation 7

Note that we could even extend these definitions inside one mode. For instance, we have:

$$sim_{TxtTxt}(i,j) = sim_{txt}(\hat{T}(i), \mathcal{T}(j)) = \sum_{d \in N_{txt}(i)} sim_{txt}(\mathcal{T}(d), \mathcal{T}(j)) \tag{10}$$

and

$$sim_{ImgImg}(i,j) = sim_{img}(\hat{I}(i), \mathcal{I}(j)) = \sum_{d \in N_{img}(i)} sim_{img}(\mathcal{I}(d), \mathcal{I}(j)) \tag{11}$$

Once again, the process could be fast and efficient, if we can precompute the similarity matrices $sim_{img}(\mathcal{I}(u), \mathcal{I}(v))$ and/or $sim_{txt}(\mathcal{T}(u), \mathcal{T}(v))$ for all pairs (u, v) of mulmodal objects in $\mathcal{O}$.

Finally, we can combine all the similarities to define a global similarity measure between two multi-modal objects i and j: for instance, using a linear combination,

$$sim_{glob}(i,j) = \lambda_1 sim_{txt}(\mathcal{T}(i), \mathcal{T}(j)) + \lambda_2 sim_{img}(\mathcal{I}(i), \mathcal{I}(j)) + \lambda_3 sim_{ImgTxt}(i,j) + \lambda_4 sim_{TxtImg}(i,j) \tag{12}$$

3.2 Complementary Feedback

Recall that the fundamental problem in transmedia feedback is to define how we compute the mono-media similarity between an aggregate of objects $N_{img}(i)$ (or $N_{txt}(i)$) and one single multimodal object. Instead of adopting the strategy of the previous section, we would now consider the set $N_{img}(i)$ as the “relevance concept” **F** and derive its corresponding language model (LM) θ_F. Afterwards, we can use the Cross-entropy criterion between θ_F and the LM of the textual part of any object j in $\mathcal{O}$ as the new transmedia similarity. We illustrate this approach when we use $N_{img}(i)$ (using only the image part of query object i) to derive a textual LM of θ_F that can be used in conjunction with the original LM of the textual part of query i.

To this aim (build the LM of the relevance concept **F**), we use a pseudo-feedback method issued from the language modelling approach to information retrieval, namely the mixture model method from Zhai and Lafferty [23] originally designed to enrich textual queries (however, more elaborated techniques of feedback for language models can also be envisaged : e.g. [22]).
Let θ_F be a multinomial parameter, standing for the distribution of relevant terms in **F**: in other words θ_F is a probability distribution over words but peaked on relevant terms. A generative model is assumed to estimate θ_F from **F**:

$$P(\mathbf{F}|\theta) = \prod_{d \in N_{img}(i)} \prod_{w} (\lambda\theta_{F,w} + (1-\lambda)P(w|\mathcal{C}))^{c(w,d)} \tag{13}$$

where $P(w|\mathcal{C})$ is word probability built upon the corpus, λ is a fixed parameter, which can be understood as a noise parameter for the distribution of terms. $c(w,d)$ is the number of occurence of term w in document d. Finally θ_F is learned by maximum likelihood with an Expectation Maximization algorithm. Once θ_F has been estimated, a new query LM can be obtained trough interpolation:

$$\theta_{new\_query} = \alpha\theta_{old\_query} + (1-\alpha)\theta_F \tag{14}$$

where θ_{old_query} corresponds to the LM of the textual part of the query i. As mentionned, we then use the Cross-Entropy similarity measure to perform a new retrieval on the textual part of objects in $\mathcal{O}$.

Setting the value of α is done experimentally and adapted to the particular collection. The robustness of the estimation of θ_F has a significant impact on the value of α. Lastly, the value of α can be interpreted as a mixing weight between image and text.

Finally, note that we illustrated the approach using $N_{img}(i)$ to derive a textual LM of θ_F that can be used in conjunction with the original LM of the textual part of query i. But we can derive a similar scheme using $N_{txt}(i)$ to derive a new representation (actually some generalized Fisher Vectors) of the “relevance concept”, this time relying on Rocchio's method that is more adapted to continuous feature representation.

4 Description of submitted runs

For a description of the task, we refer to the overview paper [10].

Table 2 shows the name of our runs and the corresponding mean average precision measures.

Below is a detailed description of our official runs.

4.1 EN-EN-AUTO-FB-TXT_FLR

This run was a pure text run: documents were basically preprocessed and each document was enriched using Flickr database. For each term of a document, its top 20 related tags from Flickr were added to the document. Then, a unigram language model for each document is estimated, giving more weight to the original document terms. An additional step of pseudo-relevance feedback using the method explained in [23] is then performed.

4.2 AUTO-NOFB-IMG_COMBFK

This run is a pure image run: it uses Fisher Kernel metric (cf. equation 7) to define the image similarity. As a query encompasses 3 visual sub-queries, we have to combine the similarity score with respect to these 3 subqueries. To this aim, the result lists from the image sub-queries are

Table 2: Official Runs

Run Name	MAP
TextT Only	
EN-EN-AUTO-FB-TXT_FLR	0.2075
Image Only	
AUTO-NOFB-IMG_COMBFK	0.1890
Image and Text	
Transmedia Reranking	
AUTO-FB-TXTIMG_PREFFKTXT	0.2801
AUTO-FB-TXTIMG_PREFFKTXT_FLR	0.2761
EN-EN-AUTO-FB-TXTIMG_QTXT_COMBPREFFKTXT	0.3020
Complementary Feedback	
EN-EN-AUTO-FB-TXTIMG_MPRF	0.3168
DE-EN-AUTO-FB-TXTIMG_MPRF_FLR	0.2899
EN-DE-AUTO-FB-TXTIMG_MPRF	0.2776

renormalized (by substracting the mean and dividing by the standard deviation) and merged by simple sum.

4.3 AUTO-FB-TXTIMG_PREFFKTXT.off

This run uses both texts and images: it starts from query images only, to determine $N_{img}(i)$ for each query i (as in the previous run above) and then implements the method described by eq. 8. The size of the neighbouring set is 5.

4.4 AUTO-FB-TXTIMG_PREFFKTXT_FLR

It is basically the same algorithm as the preceding run, except that the textual part of the data (annotations) is enriched with Flickr tags.

4.5 EN-EN-AUTO-FB-TXTIMG_QTXT_COMBPREFFKTXT.off

This run uses the same algorithm as in *AUTO-FB-TXTIMG_PREFFKTXT* but with one more step at the end, that amounts to merge the result lists from in *AUTO-FB-TXTIMG_PREFFKTXT* and from the purely text queries (*EN-EN-AUTO-FB-TXT_FLR*), by summing the relevance scores after normalisation (by substracting the mean and dividing by the standard deviation for each list).

4.6 EN-EN-AUTO-FB-TXTIMG_MPRF.off

This run uses both texts and images: it starts from query images only, to determine the relevance set $N_{img}(i)$ for each query i (as in the run AUTO-FB-TXTIMG_PREFFKTXT.off) and then implements the method described as "the complementary (intermedia) feedback" in section 3.2. The size of the neighbouring set is 15. Refering to the notations of section 3.2, the values of λ and α are respectively 0.5 and 0.5.

4.7 DE-EN-AUTO-FB-TXTIMG_MPRF_FLR.off

This runs works with the same principle as the previous run *EN-EN-AUTO-FB-TXTIMG_MPRF.off*. The main difference is that (target) english documents have been enriched with Flickr and that the initial query — in German — was translated by multiplying its "Language Model" by the probabilistic translation matrix extracted from the (small) parallel part of the corpus. Otherwise, it uses the same parameters as previously.

4.8 EN-DE-AUTO-FB-TXTIMG_MPRF.off

This run uses the same process as in *EN-EN-AUTO-FB-TXTIMG_MPRF.off*. The difference is the starting point: english queries to search for german annotations. English queries are translated with the probabilistic translation matrix extracted from the (small) parallel part of the corpus and the translated queries follow the same process as in $EN-EN-AUTO-FB-TXTIMG\_MPRF.off$ but with different parameter : the size of the neighbouring set is 10, while the values of λ and α are respectively 0.5 and 0.7.

5 Conclusions

With a slightly annotated corpus of images, also characterised by an abstraction level in the textual description that is significantly different from the one used in the queries, it appears that mono-media retrieval performance is more or less equivalent for pure image and pure text content (around 20% MAP). Using our transmedia pseudofeedback-based similarity measures allowed us to dramatically increase the performance by ∼50% (relative). Trying to model the textual "relevance concept" present in the top ranked documents issued from a first (purely visual) retrieval and combining this with the textual part of the original query turns out to be the best strategy, being slightly superior to our transmedia document reranking method. Enriching the image annotations by extra tags extracted from the Flickr database does not offer a significant advantage in the ImageCLEF07 corpus, even if we observed an improvement using other multimedia corpora and query sets. From a cross-lingual perspective, the use of domain-specific, corpus-adapted probabilistic dictionaries seems to offer better results than the use of a broader, more general standard dictionary. With respect to the monolingual baseline, multilingual runs show a slight degradation of retrieval performance (∼6 to 10% relative).

In the future, we want to investigate more systematically and more thoroughly the ways to combine the numerous transmedia similarity measures we introduced in this report, by determining in which cases they can provide us with significant advantages with respect to more traditional "late fusion" approaches.

Aknowledgments

This work was partly funded by the French Government under the *Infomagic* project, part of the Pole CAP DIGITAL (IMVN) de Paris, Ile-de-France. The authors also want to thank Florent Perronin for his greatly appreciated help in applying some of the Generic Visual Categorizer (GVC) components in ImageCLEF07 experiments.

References

[1] D. Blei, Michael, and M. I. Jordan. Modeling annotated data. In *ACM SIGIR*, 2003.

[2] P. Carbonetto, N. de Freitas, and K. Barnard. A statistical model for general contextual object recognition. In *ECCV*, 2004.

[3] Y.-C. Chang and H.-H. Chen. Approaches of using a word-image ontology and an annotated image corpus as intermedia for cross-language image retrieval, clef 2006 working notes. In *CLEF 2006 Working Notes*, 2006.

[4] 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.

[5] P. Duygulu, K. Barnard, J. de Freitas, and D. Forsyth. Object recognition as machine translation :learning a lexicon for a fixed image vocabulary. In *ECCV*, 2002.

[6] J. Farquhar, S. Szedmak, H. Meng, and J. Shawe-Taylor. Improving "bag-of-keypoints" image categorisation. Technical report, University of Southampton, 2005.

[7] S. Feng, V. Lavrenko, and R. Manmatha. Multiple bernoulli relevance models for image and video annotation. In *CVPR*, 2004.

[8] Flickr. Flickr api. http://www.flickr.com/services/api/.

[9] Flickr. tags.getrelated. http://www.flickr.com/services/api/flickr.tags.getRelated.html.

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

[11] T. Jaakkola and D. Haussler. Exploiting generative models in discriminative classifiers. In *Advances in Neural Information Processing Systems 11*, 1999.

[12] J. Jeon, V. Lavrenko, and R. Manmatha. Automatic image annotation and retrieval using cross-media relevance models. In *ACM SIGIR*, 2003.

[13] V. Lavrenko, S. Feng, and R. Manmatha. Models for automatic video annotation and retrieval. In *ICASSP*, 2004.

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

[15] D. G. Lowe. Distinctive image features from scale-invariant keypoints. *IJCV*, 60(2), 2004.

[16] N. Maillot, J.-P. Chevallet, V. Valea, and J. H. Lim. Ipal inter-media pseudo-relevance feedback approach to imageclef 2006 photo retrieval, clef 2006 working notes. In *CLEF 2006 Working Notes*, 2006.

[17] F. Monay and D. Gatica-Perez. Plsa-based image auto-annotation: Constraining the latent space. In *ACM MM*, 2004.

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

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

[20] F. Perronnin, C. Dance, G. Csurka, and M. Bressan. Adapted vocabularies for generic visual categorization. In *ECCV*, 2006.

[21] J. S. Sivic and A. Zisserman. Video google: A text retrieval approach to object matching in videos. In *ICCV*, volume 2, 2003.

[22] T. Tao and C. Zhai. Regularized estimation of mixture models for robust pseudo-relevance feedback. In *SIGIR '06: Proceedings of the 29th annual international ACM SIGIR conference on Research and development in information retrieval*, New York, NY, USA, 2006.

[23] C. Zhai and J. D. Lafferty. Model-based feedback in the language modeling approach to information retrieval. In *CIKM*, pages 403–410, 2001.