Comparison of Several Combinations of Multimodal and Diversity Seeking Methods for Multimedia Retrieval

Julien Ah-Pine, Stephane Clinchant, Gabriela Csurka

Comparison of Several Combinations of Multimodal and Diversity seeking Methods for Multimedia Retrieval

Julien Ah-Pine, Stephane Clinchant and Gabriela Csurka

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

Abstract. The aim of this paper is to analyze the technologies designed and used in the context of XRCE's participation in the Photo Retrieval Task of ImageCLEF 2009 [1]. We evaluate and compare different mono and multimedia retrieval methods and two distinct diversity-seeking strategies as well. Our analysis allows us to better understand which combinations of basic approaches are the best ones. It appears that taking advantage of the multimodal nature of the data by means of our cross-modal similarities technique and leveraging different text representations of the topics in the goal of covering distinct related subtopics, allow us to tackle the Photo Retrieval Task effectively.

1 Introduction

Given a collection of text/image objects and a set of multimedia topics and subtopics, the aim of the challenge was to produce for each topic, a ranked list of images holding both relevant and diverse objects. However, the definition of what constitutes diversity varied across topics. Basically there are two kinds of topics with respect to this aspect: part 1 and part 2. In the first part, for each subtopic of a topic, in addition to the query title, the "cluster title" field, clearly indicated what the clustering criteria and the "cluster description" field gave even more precision. However, we did not use either of them. In the second part of the challenge, only three relevant illustrative images were given with the query title of the topic, without any other indication concerning the clustering criteria. In that case, participants were encouraged to decide on how broad the results should be for each of these topics.

In what follows, we first briefly introduce the underlying technologies that we used to find relevant and diverse multimedia objects for each topic, then we perform an analysis of the results and a comparison of different strategies as well.

2 The underlying technologies

Image representation. As image representation, we use the Fisher Vector[1] proposed in [2]. This is an extension of the bag-of-visual-words representation. The main idea is to represent the visual vocabulary with a Gaussian Mixture Model (GMM) $\lambda = \{w_i, \mu_i, \Sigma_i, i = 1, ..., N\}$ where each Gaussian corresponds to a visual word and to characterize the image I with the gradient of the normalized log-likelihood according to the GMM model: $\mathbf{f}_I = F_\lambda^{-1/2} \nabla_\lambda \log P(I|\lambda)$; where F_λ is the Fisher Information matrix. The similarity between two images I_1 and I_2, is computed as follows: $2 - ||\mathbf{f}_{I_1} - \mathbf{f}_{I_2}||_1$; where the $\mathbf{f}$ are first normalized to 1.

Text representation. Two information retrieval models were considered: a standard language model and an information based model with a log-logistic distribution, after lemmatization of the texts. We refer to the working notes paper [3] for more information. Furthermore, we decided to use query expansion or enrichment, in order to find new clusters in addition to those represented by the example images for topics in part 2. The Chi-Square statistic was used to enrich the query title words with their top ten most similar terms.

Cross-media similarity. The information fusion technique used to combine textual and visual similarities can be understood as a score regularization through a two-step diffusion process, the first step being performed in one mode and the second step being performed in the other one [4, 3]. Let S_t and S_i respectively be the textual and the visual similarity matrices over the same set of multimedia objects, that are normalized to obtain a similarity value distribution between 0 and 1 for each row. The cross-media similarity matrices that combine two monomedia similarity matrices are defined as follows:

$$Sim_{img-txt} = \kappa(S_i, k_i)S_t \text{ and } Sim_{txt-img} = \kappa(S_t, k_t)S_i; \quad (1)$$

where $\kappa(S, k)$ is a 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 (see [3] for more details). Let us precise that in the more specific case of information retrieval, given a multimedia query q (q_t denoting the text part and q_i the image part of q), we similarly have the following cross-media scores: $Score_{img-txt}(q_i) = \kappa(s_i, k_i).S_t$ and $Score_{txt-img}(q_t) = \kappa(s_t, k_t).S_i$; where s_t is the similarity row vector of a given textual query q_t with a set of multimedia objects (their text part) and s_i is respectively, the similarity row vector of a given image query q_i with the same set of multimedia objects (but their image part). Finally, given a multimedia query, the final relevance score can be computed as follows:

$$Score(q) = \alpha_t s_t + \alpha_i s_i + \alpha_{it} Score_{img-txt}(q_i) + \alpha_{ti} Score_{txt-img}(q_t) \quad (2)$$

where the weight distribution was set heuristically to $\alpha_t = 5/12, \alpha_i = 1/4, \alpha_{it} = 1/4, \alpha_{ti} = 1/12$.

[1] The authors also want to thank Florent Perronin for his code allowing to compute the Fisher Vectors and to Yan Liu for his help in preprocessing the visual data.

3 Runs description and analysis

Since part 1 and part 2 constitute two different kinds of topics, we designed two slightly different approaches that we briefly describe[2] before analysing the results they can provide.

3.1 Part 1

Description. For each topic in part 1, we start by analyzing its subtopics individually as if they were independent. To this end, we first used text similarity to retrieve relevant multimedia objects. More precisely, we used the image's caption (**ICPT**) of the subtopic as a text query. This allows us to find a set of M relevant objects based on textual similarities between the text query (image's caption of the subtopic) and the captions of images in the database. Note that since the image subtopic is itself within the database, it is also retrieved in the top M list.

From these M retrieved multimedia objects we can compute three similarity matrices (textual, visual and cross-modal) as described in section 2. Based on those matrices, the top M retrieved objects can be re-ranked using either textual, visual or cross-modal similarities. Indeed, given a similarity matrix, we extract from the latter the row similarity corresponding to the subtopic and we re-rank the retrieved objects according to this similarity distribution. In that case, we can see that using only visual and cross-modal similarities actually result in a re-ranking of the top M list since the original ranking was produced by text similarities.

We thus obtain three top M lists from the three subtopics. We finally combine the latter into a single list using a Round Robin merging technique[3].

Analysis. Table 1 shows the results[4] with the three modalities: textual (**T**), visual (**I**) and cross-modal (**X**). In this part the diversity (cluster recall measure) was in principle ensured by the fact that each cluster was represented by a subtopic. Therefore we did not apply any additional diversity-seeking re-ranking strategy to the topics of part 1. As far as the notations are concerned, this is denoted by a **no** in a run's name.

From Table 1, we see that even if the cross-media run ICPT_no_X reaches the best performance, the improvement over the pure text run, ICPT_no_T is rather weak. The main reason might be that *using the subtopics' image caption and merging the resulting subtopics' top M list with a Round Robin procedure, is already an effictive strategy to address both the precision and the cluster recall.*

[2] For a more detailed description, see Algorithm 1 and Algorithm 2 in the working notes paper [3].

[3] The images of the subtopics and their exact duplicates are first removed from the different lists.

[4] The results are slightly better than the official runs. This is due to the correction of a small bug in the code.

Table 1. Results for part 1 using different modalities.

Modality	CR10	P10	F1
ICPT_no_T	**83.9**	78.4	81.0
ICPT_no_I	75.2	60.8	67.2
ICPT_no_X	83.7	**79.6**	**81.6**
(Old) ICPT_no_X	82.9	76.8	79.7

3.2 Part 2 - basic runs

Description. For topics in part 2, we assumed that the three image queries represented three different subtopics. As a consequence, we first applied the same technique as for part 1. It is worth mentioning here that, similarly to part 1, neither the cluster title nor the cluster description were exploited. Therefore, we can obtain the runs ICPT_no_T, ICPT_no_I and ICPT_no_X by considering the Round Robin fusion of the three (re-)ranked lists (using textual, visual or cross-modal similarities), in the same manner as we described previously.

Assuming that the three image subqueries were relevant, we can expect that their lists lead to high precision measures (P10). However, there might be other subtopics that are related to the topic but which are not conveyed by the image subqueries we were given[5]. To tackle such an issue, our main idea was to enrich the above obtained lists with objects that remain relevant to the general topic but which are distinct from the given image subqueries. To this end, *we proposed to use different sources of information in order to have different aspects of the topic and thus to promote diversity.* Accordingly, we used, in addition to the images captions (ICPT), two other query types: the query title (**QRW**) and the query title enriched with the most similar words (**ENT**). Again, for these extra queries, the textual information was employed as a pre-filter before using any visual information. In both cases, we first build a top list of M objects by using textual similarities[6]: ENT_no_T and QRW_no_T respectively.

Aiming to bring further diversity, *we can also use different diversity re-ranking strategies on any of the above lists* similarly to our last year's participation in the task. We mainly experimented with two different methods (see [4, 3] for more details):

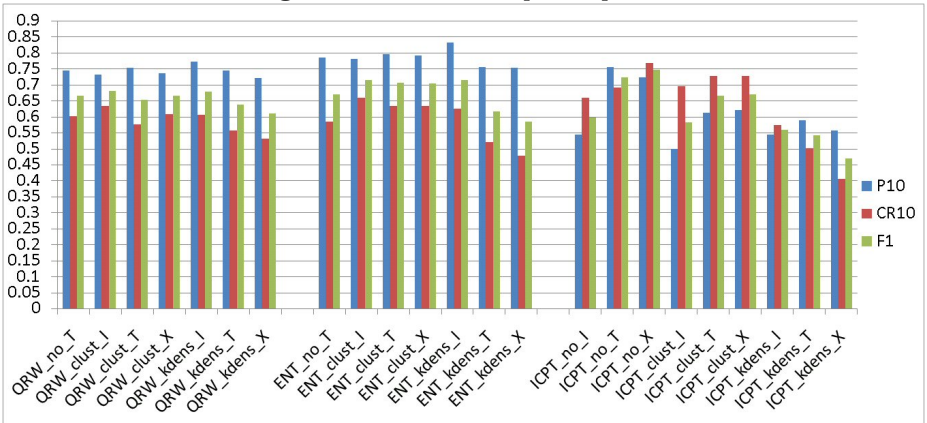
- A density based re-ranking of the top elements (**kdens**). The main idea is to select elements that have high density (similar elements) around them as representatives. We used the sum of k nearest neighbors distances as an example measure.
- A clustering based re-ranking of the top elements (**clust**). Similarly to the previous method, we search images that represent a group of images, except that we use clustering techniques.

Indeed, considering the M retrieved multimedia objects, we can first compute three similarity matrices (T, I and X) and then, re-rank the top list with either the density based method (kdens) or the clustering based method (clust)

[5] As an example, in topic 37, images about Paris Hilton, Paris-Brest and Paris-Nice clearly did not address all the aspects of the topic Paris.

[6] Note that we have a single topic list for ENT and QRW as no subtopic is assumed.

Fig. 1. Basic Runs for topics in part 2.



described above. This leads to six additional runs for each query type (see for example Figure 1).

Analysis. While these extra lists (ENT and QRW) were designed to be combined with the ICPT list in order to bring new clusters at the top of the final list, it is interesting to analyze and compare them individually in a first time. Accordingly, Figure 1 shows the performances of all the basic models we used (called basic runs). In the five first rows of Table 2, we recall some of the best basic runs we obtained. Let us analyze these results from different point of views:

- **QRW vs ENT:** Surprisingly, we can see that *query enrichment mostly benefits precision rather than cluster recall.* Query enrichment does benefit cluster recall, but to a lesser extent that we would have expected. Besides, *re-ranking techniques for query title runs (clust, kdens) do not significantly improve the F1 measure over the baseline run (QRW_no_T).* On the other hand, *a larger improvement is obtained when the same techniques are applied to ENT runs.* For example ENT_kdens_I obtains a large improvement in F1 measure, increasing both precision and cluster recall measures. Overall, we observe that enriching the query allows us to improve the performances in most cases since *ENT runs are in general better than QRW ones.*
- **ICPT:** The cross-media *ICPT_no_X is the best of our basic runs*, largely outperforming any of the other runs. Another important observation is that *no re-ranking techniques* (clust, kdens) actually *help when using ICPT.* In fact. *using the Round Robin fusion between subtopics, there may be already enough diversity represented such that the re-ranking techniques bring more noise than useful information*
- **ENT vs ICPT:** Overall, *enriched queries based runs are precision oriented.* On the contrary, *images' caption based runs are rather recall oriented.* As a result, since *those runs are complementary*, we could expect their combination to be promising.

Table 2. Best basic runs and some of their combinations for part 2 topics.

Run name	CR10	P10	F1
ENT_clust_I	65.8	78.0	71.4
ENT_clust_T	63.4	79.6	70.6
ENT_kdens_I	62.6	**83.2**	71.4
ENT_no_T	58.5	78.4	67.0
ICPT_no_X	76.8	72.4	74.6
ICPT_no_X_ENT_clust_X	**84.0**	78.0	80.9
ICPT_no_X_ENT_clust_I	82.4	78.8	80.6
ICPT_no_X_ENT_kdens_I	82.5	81.6	**82.0**
ICPT_no_X_ENT_no_T	83.1	78.8	80.9

• **Multimodal Runs:** There are evidences that taking advantage of the multimodal nature of the objects allows us to outperform monomedia based retrieval. As mentioned previously, the cross-media ICPT_no_X is a good example but there are more ones. For example, the run ENT_clust_I, first retrieves a top list of objects on the basis of textual similarities and then uses the visual similarities between those objects and a clustering technique in the goal of avoiding redundancy. We can actually observe that this run performs better than the monomedia run ENT_no_T in terms of F1. Overall, these observations demonstrate that *multimedia and cross-modal techniques are valuable and effective.*
• **kdens vs clust** Regarding the comparison between the two re-ranking techniques, we observe on the one hand that density and clustering are comparable when image similarity is used, but on the other hand, with text similarity or cross-media similarity, clustering generally gives better results. Those re-ranking techniques are really effective on the enriched queries however, as mentioned beforehand, none of these techniques helps the ICPT runs.

3.3 Part 2 - combined runs

Description. As mentioned previously, our aim when designing different runs based on different information sources, was to better promote diversity by combining several kinds of results. In that perspective, we used again the Round Robin method to combine several runs. There are many possible combinations among the basic runs that we described previously and no room for an exhausted study. Therefore we selected some of them to be analyzed and compared.

Analysis. Table 2 shows the results of our best basic runs, and some of their combinations. Since we observed that ENT basic runs are generally better than QRW ones, we thus rather consider the use of the ENT runs than the latter ones in the combinations. The best basic ICPT_no_X run reaches a 74.6% F1 value. The combination of this run with others leads to roughly 80% F1 runs. The run ICPT_no_X_ENT_no_T is particularly interesting because it is high performing and does not use any re-ranking techniques. Our *best combination is ICPT_no_X_ENT_kdens_I (F1=82%).* Therefore, we analyze these two basic runs and their combination at the topic level.

In the top chart of Figure 2, we show the F1 measures of ICPT_no_X, ENT_kdens_I and their combination. In the bottom chart we show the cluster recall at 20 (CR20) of the combination against the CR10 of the two basic runs. Our motivation is to analyze whether the retrieved subtopics in the basic runs complement each other[7]. We can make the following observations:

• Our system seems to fail on topic 43 but the relevance judgment reveals that there are only 2 relevant images for this topic and only one cluster. We actually found one relevant image so the failure is not really dramatic.

• On average, ICPT_no_X gets better results than the ENT_kdens_I, as their respective F1 scores are 74.6% and 71.4%. However, for topics where ENT_kdens_I results are better, they are much better than ICPT_no_X ones. In other words, there are a few topics where ENT_kdens_I makes a real difference over ICPT_no_X. As we already noticed above *enriching queries does not help cluster recall so often*. Thus, only queries 27, 29, 35, 37, 40, 45, 48, 49, after fusion, show some improvements in terms of cluster recall. If we look[8] at the precision of the different topics of part 2, enriched queries get often better precision than the cross-media ICPT results. We introduced query enrichment with the hope to find new clusters, but it turns out that most of the time the images' caption and the cross-media are already able to find most of the clusters[9].

• Lastly, our fusion strategy seems robust: the combination does worst than the worst basic run only for one query (query 26). Otherwise, the combination does better than the worst run. This actually confirms that those runs, ICPT_no_X and ENT_kdens_I, are complementary. Nevertheless, as far as the combination of basic runs is concerned, we could expect to obtain better results by giving more weight to the cross-media runs than other basic runs, since the former runs are better than the latter ones most of the time.

4 Conclusion

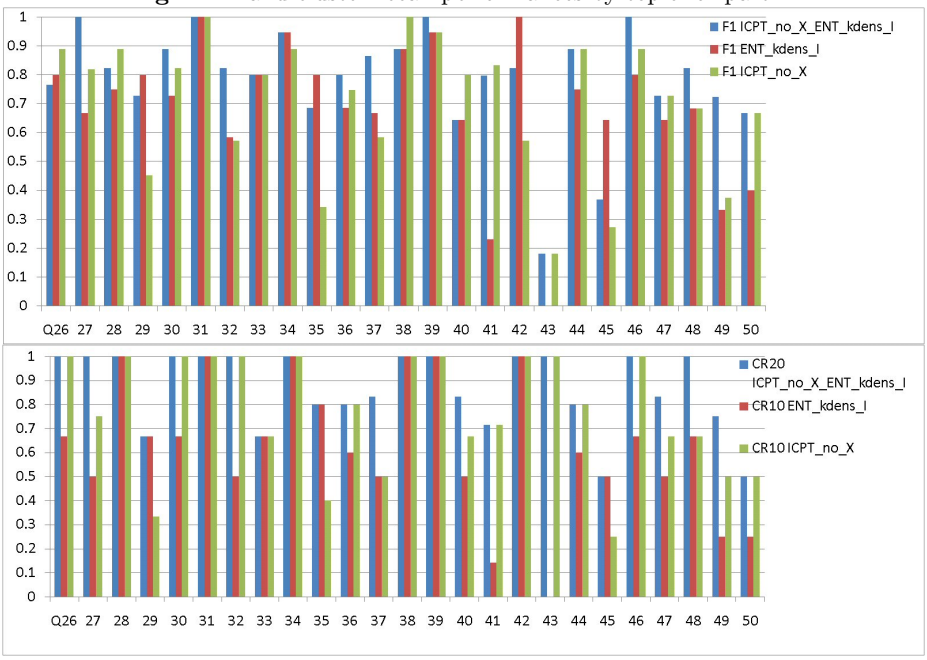
In this paper we briefly recalled the technologies we used in order to address the Photo retrieval task of ImageCLEF 2009. We then presented a detailed analysis of the different results that our methods can provide. We generated a lot of different basic runs, by varying the type of the initial query text representation and the use or not of a re-ranking technique to increase diversity. We showed that query enrichment benefits more precision than cluster recall, which was somehow unexpected. Moreover, the diversity re-ranking techniques we used, clustering and density, increase the results of enriched queries, but not the results of other text representations query title and images' caption. However, these improvements do not outperform the basic run which uses the images' caption and cross-media measures without any re-ranking method. The latter run is actually our best basic run which shows again that our cross-media technique is effective. Finally,

[7] If we only considered the CR10 of the combined list, this would take into account only the top 5 elements of each individual lists.

[8] Figure omitted due to space limitation.

[9] For 11 out of 25 topics we have CR10=1 for ICPT_no_X.

Fig. 2. F1 and cluster recall performances by topic for part 2.



textual similarities, using the images' caption particularly, allow us to obtain a first interesting baseline, that we can significantly improve in a second step, by integrating visual and cross-modal similarities. These basic runs can further be improved by combining some of them in order to increase the F1 measure. In that perspective, combining basic runs that use different information sources and media is very beneficial.

Acknowledgments This work was partially supported by the french projects Omnia ANR-06-CIS6-01 and Fragrances ANR-08-CORD-008.

References

1. Paramita, M., Sanderson, M., Clough, P.: Diversity in photo retrieval: overview of the ImageCLEFPhoto task 2009. In: CLEF working notes. (2009) http://www.imageclef.org/2009/photo.
2. Perronnin, F., Dance, C.: Fisher kernels on visual vocabularies for image categorization. In: CVPR. (2007)
3. Ah-Pine, J., Clinchant, S., Csurka, G., Liu, Y.: XRCE's participation to ImageCLEF 2009. In: Working Notes of the 2009 CLEF Workshop, Crete, Greece (September 2009)
4. Ah-Pine, J., Cifarelli, C., Clinchant, S., Csurka, G., Renders, J.: XRCE's participation to ImageCLEF 2008. In: Working Notes of the 2008 CLEF Workshop, Aarhus, Denmark (September 2008)