XRCE's Participation in ImageCLEF 2009

Julien Ah-Pine[1], Stephane Clinchant[1,2], Gabriela Csurka[1], Yan Liu[1]

[1] Xerox Research Centre Europe, 6 chemin de Maupertuis 38240, Meylan France
firstname.lastname@xrce.xerox.com
[2] LIG, Univ. Grenoble I, BP 53 - 38041 Grenoble cedex 9, Grenoble France

Abstract. This paper describes XRCE's participation in Large Scale Visual Concept Detection and Annotation Task [16] and Photo Retrieval Task [17] of ImageCLEF 2009. Those tasks both use a new collection which is different and which is much larger than the ones used in past sessions. Moreover, new kinds of challenge to tackle were designed such as new categories to detect or new types of topic to deal with. Accordingly, our main motivations regarding our participation in this year's session are two fold. First, we wanted to apply ongoing work in our team in image and text processing. Second, we wanted to figure out if our cross-media approach and our diversity re-ranking techniques developped in past sessions, can perform well on a new challenging corpus. It turns out that the material that we describe in this paper both made of new and already well-established techniques allow to perform very well on those two tasks since the results we obtained with the systems we designed are top ranked.

1 Introduction

This year, we participated in two main tasks: the Large Scale Visual Concept Detection and Annotation Task (detailed in [16]) and the Photo Retrieval Task (detailed in [17]).

For the visual concept detection task, the main research questions were two-fold. First, the focus this year lied on the extension of the task concerning the amount and the diversity of concepts (53) used as types of annotation. There were also different kinds of concept such as abstract categories (landscape, family and friends, party life, ...), seasons, time of day, persons (no, single, big groups), quality (blurred, underexposed) and representation (portrait, macro image, canvas). Secondly, the aim was to assess whether a given ontology of concepts could help in classification or not.

Despite the fact that the concepts were presented in a hierarchy with a given ontology, we have not used neither the hierarchy nor the ontology in the training phase. Indeed we considered the task first as a multi-class multi-label image categorization problem and trained binary one versus all (OVA) classifiers for each concept. To train the classifiers, we used the Fisher Representation of images (as last year). However, we also used the adapted image GMMs which were combined with Fisher Representation by late fusion technique (score averaging). In the test phase and for each image, the classification scores were finally post-processed in order to ensure that the constraints of the ontology were not violated.

Similarly, the Photo Retrieval Task sets two new challenges: the scalability of the methods in order to cope with 500,000 images and the diversity of the top retrieved results. The basic methodologies used - *Language Model for text retrieval, Fisher Vector image representation, trans-media fusion based cross-media similarity and clustering based re-ranking of top results* - were ones already employed in previous years but with slight modifications and adaptations to better fit the actual task.

Furthermore, we had to design some new strategies to cope with on one hand, the large amount of data; on the other hand, the lack of information concerning the clusters in the second part of queries:

- Regarding the scalability problem, instead of pre-computing the whole cross-media similarity matrix off-line (as was done last year), we compute a cross-media similarity matrix "on-line" for each individual query (sub-)topic. This is actually done in two steps. First, we used a text-based retrieval technique. Then, we used the top list obtained from the latter in order to compute visual similarities and cross-media similarities as well. It is the computation of visual similarities that suffers the most from scalability issues. Accordingly, we abandoned the computation of similarity matrices over the whole collection of size $500,000 \times 500,000$ since that was estimated to be very costly both in speed and storage.
- Regarding the second part of queries, no cluster information was available. In that case, this is somehow similar to the approach we experimented for last year task as we did not use the cluster description. Therefore, we re-used some of our last year strategies to promote diversity among the top list namely clustering and KNN density based re-ranking. Furthermore it is worth noticing that if we use only the captions of the images queries[3], we would get results illustrating only the corresponding sub-topics. Thus, to increase our chances to find new relevant sub-topics, we combined different sources of information. Basically, we combine the query title and the captions of images. This would ensure that the retrieved elements are relevant to the *query topic* but somehow this would also allow to have more diverse elements. Furthermore, we propose to enrich the query with terms that are the most related in the corpus using some standard term similarities.

The paper is structured as follows. Section 2 described the textual retrieval models. In section 3, we present the Fisher and adapted GMM representation of images. We recall our cross-media similarity measure in section 4 and our diversity seeking methods in 5. Before concluding, a detailed description of our runs in both challenges are presented in section 6.

2 Text Retrieval

We start from a traditional bag-of-word representation of pre-processed texts: pre-processing includes tokenization, lemmatization, and standard stopword removal. Two information retrieval models were considered: a standard language models (as our previous participation) and an information based model on a log-logistic distribution. We then present a query expansion mechanism that appeared to be relevant to the Photo Retrieval Task.

2.1 Language Models

The idea of language models is to represent queries and documents by multinomial distributions [22, 20] .Those distributions are estimated by maximizing the likelihood. Then, documents distributions are smoothed by a Dirichlet Prior [22] and the Cross-Entropy[4] can be used to rank documents according to:

$$sim_{txt}(D_1, D_2) = CE(D_1|D_2) = \sum_{w} p(w|D_1)\log(p(w|D_2)) \qquad (1)$$

where D_1 and D_2 are two texts and w is a term from the bag-of-word representation. Notice that in the context of information retrieval, D_1 is the query that contains only a few keywords in comparison to a document. However, we also use language models to measure the similarity between pairs of documents in the collection so we rather introduce more general notations.

2.2 Log-Logistic Model

We used in this paper a new family of IR models similar in spirit to DFR models [4], which is based on probability distributions fitting well empirical data, and satisfying heuristic retrieval constraints [6, 7].

[3] Notice that this is systematically done through our cross-media technique.

[4] The Cross-Entropy multiplied by -1 actually.

The general idea of this family is the following one: due to different document length, discrete term frequencies are renormalized into continuous values as in [4]. Let m be the mean document length and y_D the length of document D, x_w^D is the raw number of occurrences of term w in document D. Then, the normalized frequencies are:

$$t_w^D = x_w^D \log(1 + \frac{m}{y_D})$$

For each term w, we assume that those renormalized values follow a probability distribution P. In our case, we suppose a log-logistic distribution. The log-logistic distribution is given by:

$$P_{LL}(X < t; r, \beta = 1) = \frac{t^\beta}{t^\beta + r^\beta} \tag{2}$$

and the parameter r_w is estimated with: $r_w = \frac{N_w}{N}$, where N_w is the number of document where w occurs in and N is the total number of document in the collection.

Finally, queries and documents are compared through a measure of surprise, or a mean of information of the following form:

$$sim_{txt}(D_1, D_2) = \sum_{w \in D_1 \cap D_2} -x_w^{D_1} \log(P_{LL}(X \geq t_w^{D_2}; r_w))$$

$$sim_{txt}(D_1, D_2) = \sum_{w \in D_1 \cap D_2} x_w^{D_1} \log(\frac{t_w^{D_2} + r_w}{r_w}) \tag{3}$$

2.3 Lexical Entailment / Term Similarity

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 goal 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 [13], although any other term similarity measure or lexical entailment measure could be used. Recall that for two terms u, v, one can fill in a 2 by 2 contingency table with the number of documents that contains u and v, only u, only v or neither of them.

As the number of clusters were not known before-hand for the second part of queries, we decided to use this Chi-Square measure to only enrich the text queries of the second part. 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 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 somehow also indicated in the topic images.

3 Image Representation

In this section, we first describe briefly our visual vocabulary modeled by Gaussian Mixture Model (section 3.2). Then, we present the Fisher Vector representation (section 3.2) which was used in both Visual Concept Detection Task (section 6.1) and Photo Retrieval Task (section 6.2). Finally, we present an alternative representation of images using adapted GMMs which was only used in the Visual Concept Detection Task (section 6.1).

Table 1. Query Terms and their most similar terms

1	obama 1	strike 1	euro 1
2	barack 0.98	hunger 0.04	million 0.05
3	springfield 0.16	protest 0.02	billion 0.05
4	illinois 0.16	worker 0.01	currency 0.03
5	senator 0.09	caracas 0.01	2004 0.03
6	freezing 0.08	led 0.01	coin 0.02
7	formally 0.08	venezuela 0.01	devil 0.02
8	ames 0.07	chavez 0.001	qualify 0.02
9	democrat 0.06	nationwide 0.001	qualification 0.02
10	paperwork 0.04	retaliatory 0.001	profit 0.01

3.1 The Visual Vocabulary

We model the visual vocabulary [21, 8] with a Gaussian Mixture Model (GMM) where each Gaussian corresponds to a visual word [9, 19]. Let $\lambda = \{w_i, \mu_i, \Sigma_i, i = 1, ..., N\}$ be the set of parameters of P where w_i, μ_i and Σ_i denote respectively the weight, mean vector and covariance matrix of Gaussian i and where N denotes the number of Gaussians. Let P_i be the distribution of Gaussian i so that we have

$$P(x) = \sum_{i=1}^{N} w_i P_i(x) = \sum_{i=1}^{N} w_i \mathcal{N}(\mu_i, \Sigma_i). \tag{4}$$

The visual vocabulary is supposed to describe the content of any image and, therefore, it is trained off-line on a varied set of images. Let $X = \{x_t, t = 1, ..., T\}$ be the set of training observations extracted from these images. The estimation of λ is performed by maximizing the log-likelihood function $\log P(X|\lambda)$ under an independence assumption:

$$\log P(X|\lambda) = \sum_{t=1}^{T} \log P(x_t|\lambda) \tag{5}$$

as described in [19].

3.2 Fisher Representation of Images

As image representation, we use the Fisher Vector proposed in [18]. This is an extension of the bag-of-visual-words representation. The main idea is to characterize the image I (represented by a set of samples x_t) with the gradient of the log-likelihood $\nabla_\lambda \log P(I|\lambda)$. In the case of a GMM with isotropic covariance matrix (see section 3.1 and [18]), we have the following formulas:

$$\frac{\partial \mathcal{L}(I|\lambda)}{\partial w_i} = \sum_{t=1}^{T} \left[\frac{\gamma_t(i)}{w_i} - \frac{\gamma_t(1)}{w_1} \right] \text{ for } i \geq 2 \ , \tag{6}$$

$$\frac{\partial \mathcal{L}(I|\lambda)}{\partial \mu_i^d} = \sum_{t=1}^{T} \gamma_t(i) \left[\frac{x_t^d - \mu_i^d}{(\sigma_i^d)^2} \right] \ , \tag{7}$$

$$\frac{\partial \mathcal{L}(I|\lambda)}{\partial \sigma_i^d} = \sum_{t=1}^{T} \gamma_t(i) \left[\frac{(x_t^d - \mu_i^d)^2}{(\sigma_i^d)^3} - \frac{1}{\sigma_i^d} \right] \ . \tag{8}$$

where $\gamma_t(i) = P(i|x_t, \lambda)$, the superscript d denotes the d-th dimension of a vector and σ_i^d are the diagonal elements of Σ_i. In practice, we only use the partial derivatives with respect to the means and standard deviations since adding the partial derivative with respect to the mixture weights does not improve accuracy.

The main advantage of this representation, which we will call Fisher vector, is that it transforms a variable length sample (number of local patches in the image) into a class independent fixed length representation whose size is only dependent on the number of parameters in the model (λ).

Before feeding these vectors to a classifier or computing similarities between images, each vector is first normalized using the Fisher Information matrix F_λ (see [18] for the computational details):

$$\mathbf{f}_I = F_\lambda^{-1/2} \nabla_\lambda \log P(I|\lambda) \tag{9}$$

with

$$F_\lambda = E_{XP}\left[\nabla_\lambda \log P(I|\lambda) \nabla_\lambda \log P(I|\lambda)^T\right] \ .$$

and then re-normalized to have an L1-norm equal to 1.

In the case of image retrieval, to obtain the similarity between two images I_1 and I_2, we compute the L1-norm of the two Fisher vectors:

$$sim_{img}(I_1, I_2) = norm_{max} - ||\mathbf{f}_{I_1} - \mathbf{f}_{I_2}||_1 = norm_{max} - \sum_i |f^i_{I_1} - f^i_{I_2}| \tag{10}$$

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

3.3 Images as Adapted Mixtures of Gaussian

We adapt here the method proposed in [12], where each image is represented by a GMM adapted from a common "universal" GMM (λ) using the maximum a posteriori (MAP) criterion. Let $I = \{x_t, t = 1, ..., T\}$ denote the set of "adaptation samples" extracted from one image. The goal of MAP estimation is to maximize the posterior probability $P(\lambda^a|I)$ or equivalently $\log P(I|\lambda^a) + \log P(\lambda^a)$, where λ^a is an "adaptation" of λ (see [12] for details).

The advantages of this representation are two fold. MAP provides a more accurate estimate of the GMM parameters compared to standard maximum likelihood estimation (MLE) in the challenging case where the cardinality of the vector set is small. Moreover, there is a correspondence between the Gaussians of two GMMs adapted from a common distribution and one can take advantage of this fact to compute efficiently the probabilistic similarity.

To measure the similarity between two images, we use the Kullback-Leibler Divergence (KLD) between two continuous distributions P and P':

$$KL(P||P') = \int_x P(x) \log \frac{P(x)}{P'(x)} dx \tag{11}$$

However, there is no closed-form expression for the KLD between two GMMs $P(x) = \sum_{i=1}^N \alpha_i P_i(x)$ and $P'(x) = \sum_{j=1}^N \beta_j P'_j(x)$. Therefore, we use the KLK approximation proposed by Liu et Perronnin [12]:

$$KL(P||P') \approx \sum_{i=1}^N \alpha_i \left(KL(P_i||P'_i) + \log \frac{\alpha_i}{\beta_i} \right) \ . \tag{12}$$

4 Cross-media similarity

4.1 Cross-media similarities

This section deals with the intermediate fusion we used in our experiments. It aims at combining visual and textual similarities in an efficient manner.

The method is similar to the one we introduced in the 2007's session [5] and used in last year session [3]. This method has proved to significantly leverage multimedia retrieval performances.

Moreover, cross-media similarities that we obtain using this approach will be the basis for diversity-based extensions that will be discussed in the next section. As a result, we recall the basic concepts of this approach in the following.

In general terms, we assume that S_t and S_i are two similarity matrices over the same set of multimedia objects denoted $O_i; i = 1, \ldots, K$. The former matrix is related to textual based similarities whereas the latter matrix is rather based on visual similarities. Typically, S_t is related to Eq. (1) or Eq. (3) and S_i to Eq. (10). We actually use a linear transformation of the latter matrices in order to normalize the proximity measures distribution of each row so as to obtain a similarity value distribution between 0 and 1.

Then, let $\kappa(S, k)$ be 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:

$$Sim_{img-txt} = \kappa(S_i, k_i).S_t \tag{13}$$

$$Sim_{txt-img} = \kappa(S_t, k_t).S_i \tag{14}$$

where the . symbol designates the standard matrix product.

In that context, 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 [5, 3, 2] for further details).

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_i the image part of q). In that context, as far as the notations are concerned, we rather have the following cross-media score definition:

$$Score_{img-txt} = \kappa(s_i, k_i).S_t \tag{15}$$

$$Score_{txt-img} = \kappa(s_t, k_t).S_i \tag{16}$$

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 similarly, the similarity row vector of a given image query q_i with the the same set of multimedia objects (but their image part).

4.2 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 our experiments in previous sessions of ImageCLEF Photo Retrieval Task.

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

$$Sim = \alpha_t S_t + \alpha_i S_i + \alpha_{it} Sim_{img-txt} + \alpha_{ti} Sim_{txt-img} \tag{17}$$

where $\alpha_t, \alpha_i, \alpha_{it}, \alpha_{ti}$ are four weights that sum to 1. In our experiments we used the following distribution: $\alpha_t = 5/12, \alpha_i = 1/4, \alpha_{it} = 1/4, \alpha_{ti} = 1/12$.

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

$$Score = \alpha_t s_t + \alpha_i s_i + \alpha_{it} Score_{img-txt} + \alpha_{ti} Score_{txt-img} \tag{18}$$

where the weight distribution is set in the same manner as previously.

5 Promoting diversity

One of the aims of ImageCLEFPhoto 2009 is to promote diversity in the top search results so that the first retrieved elements are not redundant. Therefore, we investigated different strategies to promote such diversity:

1. Using different sources of information for the same query: for example the caption of images, the query title and an enriched query. We can combine these different top lists using the Round Robin principle.
2. Using a density based re-ranking of the top elements.
3. Using a clustering based re-ranking of the top elements.

In this section we introduce these latter techniques.

5.1 Round Robin Approach

The main idea of Round Robin principle is that each individual (in our case the different top lists) takes its turn to make a single list. We take for each top list the elements with respect to their rank and we update a single list by taking the following element top list after top list. A same item can appear in different top lists. Thus, we first verify whether a coming element is already in the merged list or not. If not it is appended to the latter.

5.2 Density Based Re-ranking

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

$$d(O_i) = \sum_{j=1}^{k} S(O_i, O_j^i) \tag{19}$$

where $O_j^i; j = 1, \ldots, k$ are the k nearest neighbors of O_i and the proximity measure between two multimedia objects can be based on visual similarity[5] given by Eq. (10), textual similarity (between their captions[6]) based on Eq. (3) or cross-media similarities (as described in section 4.1).

Finally, we re-rank the objects according to this measure by ranking first the objects that are the most "dense" and by discarding the nearest neighbors[7] of these latter elements added to the list.

5.3 Clustering Based Re-ranking

Our clustering based re-ranking approach we tested is similar to last year's session so hear we will only briefly recall (see [3] for more details).

We assume here that we are given an ordered top list of objects Q and a similarity matrix S between these objects (both Q 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 objects [15, 14, 3, 1].

[5] In that case, we use the image part I_i of the multimedia object O_i.

[6] In that case, it is the text part D_i of the multimedia object O_i.

[7] 2 nearest neighbors.

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

$$C(S,X) = \sum_{i,i'=1}^{|Q|} \left(S(O_i,O_{i'}) - \underbrace{\frac{1}{|\mathbb{S}^+|} \sum_{(O_i,O_{i'})\in\mathbb{S}^+} S(O_i,O_{i'})}_{\text{constant threshold}} \right) X(O_i,O_{i'}) \tag{20}$$

where $X(O_i, O_{i'}) = 1$ if O_i and $O_{i'}$ are in the same cluster and $X(O_i, O_{i'}) = 0$ otherwise; and $\mathbb{S}^+$ is the set of pairs of objects which similarity measure is strictly positive: $\mathbb{S}^+ = \{(O_i, O_{i'}) \in Q \times Q : S(O_i, O_{i'}) > 0\}$.

From Eq. (20), we can see that the more the similarity between two objects 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 [1]. In order to find a partition represented by X that maximizes the objective function we used the same clustering algorithm described in [3, 1]. 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 objects.

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.

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

```
Require: A (sub-)topic $q$, an ordered list $Q$ according to some relevance score between $q$ and $Q_i; i = 1, \dots, |Q|$ and $R$ the clustering results of objects in $Q$.
  Let $L1$, $L2$, $L3$ and $CL$ be empty lists and $i = 2$.
  Add $Q_1$ as first element of the re-ranked list $L1$ and $R(Q_1)$ (the cluster id of $Q_1$) to the cluster list $CL$
  while $i \leq |Q|$ and Stopping criterion is not fulfilled do
    if $R(Q_i) \in CL$ then
      Append $Q_i$ to $L2$
    else
      Append $Q_i$ to $L1$ and add $R(Q_i)$ in $CL$
    end if
    $i = i + 1$
  end while
  Put if not empty the complementary list of objects from $Q_i$ to $Q_{|Q|}$ in $L3$.
  Extend $L1$ with $L2$ then with $L3$ and return $L1$.
```

The stopping criterion in Algorithm 1 we used is related to a parameter denoted *nbdiv* $\in 1, \dots, \kappa$[8]. 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 $\kappa \geq 10$). Once 10 different clusters are appended, the complementary list (from the 11^{th} rank to the $|Q|^{th}$ rank), is constituted of the remaining multimedia objects sorted in the original list without taking into account the cluster membership information.

6 Runs description

In this year, we have participated in two tasks of ImageCLEFPhoto 2009: Large Scale Visual Concept Detection and Annotation and Photo Retrieval Tasks. In this section, we describe in more details our runs and show the obtained results.

[8] κ is the number of clusters found during the clustering process.

6.1 Large Scale Visual Concept Detection and Annotation Task

The Visual Concept Detection and Annotation Task (described in details in [16]), had the objective to identify 53 visual concepts in users' photos organized in a hierarchy. The main goal was to indicate the presence or the absence of these concepts ensuring some given ontological constraint between categories (eg. a picture cannot be labeled with summer and spring in the same time, etc).

In spite of the fact that the concepts were presented in a hierarchy with a given ontology, we have not used neither the hierarchy nor the ontology in the training phase but considered the task as a multi-class multi-label image categorization problem and trained binary one versus all classifiers for each concept. Then for each image the classification scores were post-processed in order to ensure that the constraints of the ontology were not violated.

To do this, two types of low-level local features were extracted: SIFT-like features (referred here as texture) and local RGB statistics (referred as color). Both type of features were extracted on a multi-level image grid and the dimensionality of the feature vectors was reduced to 50 through Principal Component Analysis. We built a visual vocabulary in both feature spaces and used these "'universal" vocabularies to define higher level representations (Fisher Vectors and Adapted Mixtures of Gaussian) as described in section 3. Hence, using the training data we built four one-against-all discriminative classifiers for each concept using alternatively color and texture and Fisher Vectors and image-GMMs.

To train the classifiers, we used our own implementation of Sparse Logistic Regression (SLR) [11], (i.e. logistic regression with a Laplacian prior), L1-norm for Fisher Vectors (see section 3.2) and the KLK approximation proposed by Liu et Perronnin [12] (see section 3.3).

We transformed the classifier scores s in "probabilities" with the sigmoid mapping $(1 + exp(s))^{-1}$ and for each concept simply averaged the four scores.

Finally, in order to ensure the constraints of the ontology were not violated, we post-processed the scores as follows:

1. For concepts that were modelled as disjoint to each other, we normalized their "probabilities" to sum to 1.
2. For a parent concept, we ensured that its score is higher or equal than the maximum score of his children.
3. For a concept that implies another concept, we corrected only if their scores were conflicting to each other.

Two measures were used to determine the quality of the annotations. One for the evaluation per concept and the other one for the evaluation per photo. The first one was the Equal Error Rate (EER) and the Area under Curve (AUC). The second measure is the proposed hierarchical measure (HM) that considers the relations between concepts and the agreement of annotators on concepts. Figure 1 plots all the results submitted to the challenge (blue) and in red our results. We can see that our system was among the top performing systems using the hierarchical measure (with or without annotator agreements) and averaging the four scores are at third position (with a score of 0.789).

6.2 Photo Retrieval Task

The Photo Retrieval Task of ImageCLEF 2009 (described in details by [17]), was intended to provide a further study of the importance of diversity in image search results. This aspect was already one of the goal of last year's session. However, this years task elevates the research scope by using a new data set containing half a million images.

Participants had to produce a ranking holding both relevant and diverse objects.

The definition of what constitutes diversity varied across topics. In the first part of the challenge, the "cluster title" fields, clearly indicated what the clustering criteria were. An additional "Cluster description" tag gave even more precision however, we haven't used it. In the second part of the challenge, only three relevant example images were given with the query title, without any

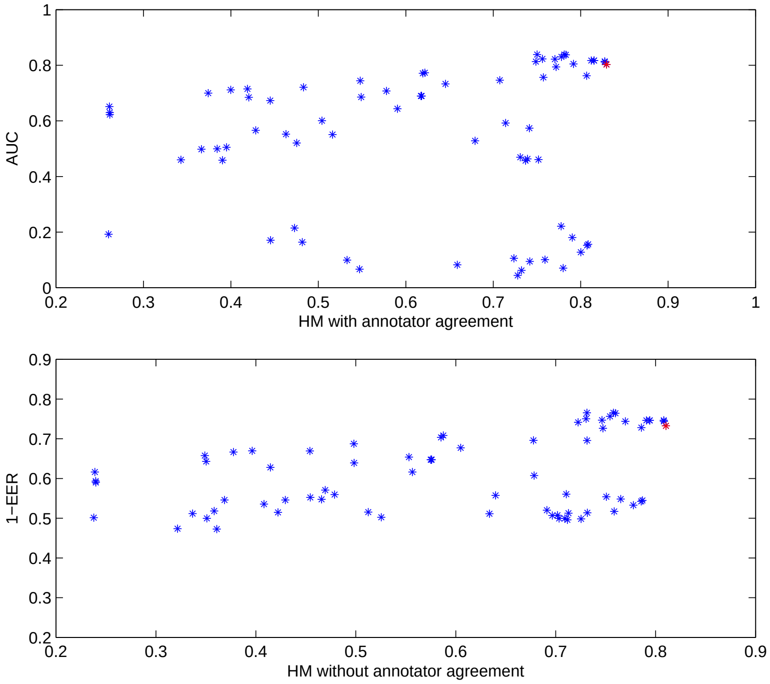

Fig. 1. ImageClef Large Scale Visual Concept Detection Task results, the red star represents our system

other indication concerning the clustering criteria. Participants were encouraged to decide on how broad the results should be for each of these topics.

Therefore, in our experiments we designed two different strategies to handle the two distinct parts. The steps of the first part are given by Algorithm 2. In the experiments, all runs used the same output for this part 1.

Algorithm 2 *Image retrieval part 1*

```
Require: query title and query images with their captions.
  for Each topic do
    for Each sub-topic $m$ do
      Rank the images according to the text similarity between the "the caption of the query image"
      (we assume that it is relevant to the query sub-topic) and the captions of all other images.
      Retain the top $M$ images and compute the $M \times M$ cross-media similarity matrix using Eq. (17).
      Re-rank the images according to their cross-media similarity with the query image (using the
      image and its caption). As the query image belongs to the top $M$ images, this is equivalent to the
      re-ranking of the row corresponding to the query image in the $M \times M$ matrix.
    end for
    Merge the $m$ ranked lists corresponding to the different sub-topics using the Round Robin technique
    (eliminating from the list the query images and the exact duplicates coming from different lists).
  end for
```

The second part was more challenging especially from the diversity viewpoint since we had no particular indication either about the number of clusters or about the clustering criteria. Therefore, we experimented different strategies for this part. We considered different sources of information (Table 2) for the query and different ranking criteria (Table 3). The steps of the second part are given by Algorithm 3.

In order to re-rank the retrieved objects, we can use either the density based measure or the clustering algorithm. For the density based, we set k=10 neighbors and the distance used was the visual similarity. However, the clustering methods used the cross-media similarity, such as defined in Eq. (17)). Nevertheless, it is worth noticing that any kinds of similarity measures (visual, textual, or cross-modal) could be used in order to assess the diversity in the aforementioned re-ranking techniques.

Table 2. Source of Information

Type	Acronym
Captions of Image Query	ICPT
Query Title	QRW
Enriched Query with most similar Terms	ENT

Table 3. Ranking Criterion

Simply Apply the Cross Media Similarity Measure
Density based Re-ranking
Clustering based Re-ranking

As it can be seen from Algorithm 3, we can generate a variety of different final lists by using different strategies. In the challenge we submitted 4 of them, summarized in Table 4.

Table 4. List of the combinations used for official runs and their F1-measure

Run Name	List Used for Round Robin			F1
XRCE1	ICPT_xsim_rank	ENT_txt_rank	QRW_txt_rank	74.4
KNND	ICPT_xsim_rank	ENT_knndens_rerank		76.2
XRCECLUST	ICPT_cluster_rerank	ENT_cluster_rerank		79.4
XRCEXKNND	ICPT_xsim_rank	QRW_knndens_rerank	ENT_knndens_rerank.	80.9

The results were ranked in terms of two measures, the precision at 10 (P10) retrieved items and the cluster recall at 10 (CR10). The two measures were combined using the F1-measure. According to this measure XRCEXKNND strategy was the best with $F1 = 0.8087$ followed by XRCECLUST and KNND.

XRCE1 got a much lower F1 score (0.7439) which shows that the re-ranking was important to enhance the retrieved list. In fact, XRCE1 is a run without any re-ranking to remove near-duplicate elements. It just used different representations of the query and applied a standard ranking algorithm. Regarding the second part of queries, our re-ranking strategies allow to leverage this baseline run with CR10 equal to 60% to runs with around 80% CR10.

We have not explored yet all the possibilities given by Algorithm 3. Can we reach the same performance with pure text runs ? Is the density measure better than the clustering methods, or are they complementary ?

In order to have final conclusions about the best strategy to be applied, we will compare the different individual and combined lists when the cluster judgements will be available.

7 Conclusion

We have participated in two tasks, Large Scale Visual Concept Detection and Annotation and Photo Retrieval. In both cases, we kept our leading position in spite of the fact that both tasks presented new challenges.

The basic methodologies used in our systems were mainly those used in previous years with slight modifiations and adaptations to better fit the actual tasks. No particular parameter tuning was necessary to achieve the good performances we obtained, on a new corpus. This shows the robustness of our methods.

However, we would like to better analyze our results and particularly for the Photo Retrieval task. In our approach, text-based retrieval is the cornerstone. To which extent, did the visual information leverage these performances ? Was the query expansion able to find new clusters when seeking to promote diversity ? and so on ...

Such investigations could be beneficial in order to better understand the properties of our systems and, in a more general perspective, to better tackle information access problems in large image/text collections.

Acknowledgments

This work was partially supported by the French National Project Omnia ANR-06-CIS6-01.

References

1. J. Ah-Pine. Cluster analysis based on the central tendency deviation principle. In *Proceedings of ADMA*, 2009.
2. J. Ah-Pine, M. Bressan, S. Clinchant, G. Csurka, Y. Hoppenot, and J.M. Renders. Crossing textual and visual content in different application scenarios. *Multimedia Tools Appl.*, 42(1):31–56, 2009.
3. J. Ah-Pine, C. Cifarelli, S. Clinchant, G. Csurka, and J.M. Renders. Xrce's participation to imageclef 2008. In *Working Notes of the 2008 CLEF Workshop*, Aarhus, Denmark, September 2008.
4. Gianni Amati and Cornelis Joost Van Rijsbergen. Probabilistic models of information retrieval based on measuring the divergence from randomness. *ACM Trans. Inf. Syst.*, 20(4):357–389, 2002.
5. S. Clinchant, J.M. Renders, and G. Csurka. Xrce's participation to imageclef 2007. In *Working Notes of the 2007 CLEF Workshop*, Budapest, Hungary, September 2007.
6. Stéphane Clinchant and Éric Gaussier. Bridging language modeling & divergence from randomness models: A log-logistic model for ir. In *ICTIR*, 2009.
7. Stéphane Clinchant and Éric Gaussier. Retrieval constraints and word frequency distributions a log-logistic model for ir. In *CIKM*, 2009.
8. G. Csurka, C. Dance, L. Fan, J. Willamowski, and C. Bray. Visual categorization with bags of keypoints. In *ECCV Workshop on Statistical Learning for Computer Vision*, 2004.
9. J. Farquhar, S. Szedmak, H. Meng, and J. Shawe-Taylor. Improving "bag-of-keypoints" image categorisation. Technical report, University of Southampton, 2005.
10. J.A. Hartigan. *Clustering Algorithms*. John Wiley and Sons, 1975.
11. B. Krishnapuram, L. Carin, M. Figueiredo, and A. Hartemink. Sparse multinomial logistic regression: Fast algorithms and generalization bounds. *PAMI*, 2005.
12. Y. Liu and F. Perronnin. Similarity measure between unordered vector sets with application to image categorization. In *CVPR*, 2008.
13. Christopher D. Manning, Hinrich Schiitze, and Lillian Lee. Review: Foundations of statistical natural language processing, christopher d. manning and hinrich schtze, 2000.
14. J.F. Marcotorchino and P. Michaud. Heuristic approach of the similarity aggregation problem. *Methods of operation research*, 43:395–404, 1981.
15. P. Michaud and J.F. Marcotorchino. Optimisation en analyse de données relationnelles. In *Data Analysis and informatics*. North Holland Amsterdam, 1980.
16. Stefanie Nowak and Peter Dunker. Overview of the clef 2009 large scale - visual concept detection and annotation task. In *CLEF working notes*, 2009. http://www.imageclef.org/2009/PhotoAnnotation.
17. M. Paramita, M. Sanderson, and P. Clough. Diversity in photo retrieval: overview of the imageclefphoto task 2009. In *CLEF working notes*, 2009. http://www.imageclef.org/2009/photo.

18. F. Perronnin and C. Dance. Fisher kernels on visual vocabularies for image categorization. In *CVPR*, 2007.
19. F. Perronnin, C. Dance, G. Csurka, and M. Bressan. Adapted vocabularies for generic visual categorization. In *ECCV*, 2006.
20. J. Ponte and W.B. Croft. A language modelling approach to information retrieval. In *Proceedings of the 21st Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 275–281. ACM, 1998.
21. J. S. Sivic and A. Zisserman. Video google: A text retrieval approach to object matching in videos. In *ICCV*, volume 2, 2003.
22. C. Zhai and J lafferty. A study of smoothing methods for language models applied to ad hoc to information retrieval. In *Proceedings of SIGIR'01*, pages 334–342. ACM, 2001.

Algorithm 3 *Image retrieval part 2*

```
Require: query title and query images with their captions
  for Each topic do
    Create three top lists using different text queries:
    if ICPT (image captions) then
      for Each query image $i = 1$ to 3 do
        Rank the images according to the text similarity between the "the caption of the query image"
        (we assume that it is relevant to the query sub-topic) and the captions of all other images (see
        section 2 for details).
        Retain the top $M$ images and compute the $M \times M$ cross-media similarity matrix using Eq. (17).
        Re-rank the images according to one of the following strategies:
        if ICPT_xsim_rank then
          Rank according to their cross-media similarity with the query image using Eq. (18).
        else {ICPT_cluster_rerank}
          Rank according to their cross-media similarity with the query image such as for
          ICPT_xsim_rank.
          Re-rank the images of the top list of ICPT_xsim_rank using the clustering based re-ranking
          method described in section 5.3 (using cross-modal similarities).
        else {ICPT_knndens_rerank}
          Rank according to their cross-media similarity with the query image such as for
          ICPT_xsim_rank.
          Re-rank the top list of ICPT_txt_rank using density based re-ranking described in section 5.2
          (using cross-modal similarities).
        end if
      end for
      Merge the 3 ranked lists corresponding to the 3 images using the Round Robin technique (eliminating
      from the list the query images and the exact duplicates coming from different lists).
    else {QRW (query title)}
      Rank the images according to one of the following strategies:
      if QRW_txt_rank then
        Rank the images according to the text similarity between the query title and the captions of all
        other images and consider the top $M$.
      else {QRW_cluster_rerank}
        Rank the images according to the text similarity between the query title and the captions of all
        other images and consider the top $M$ such as for QRW_txt_rank.
        Re-rank the top list of QRW_txt_rank using the clustering based re-ranking method (using cross-modal similarities).
      else {QRW_knndens_rerank}
        Rank the images according to the text similarity between the query title and the captions of all
        other images and consider the top $M$ such as for QRW_txt_rank.
        Re-rank the top list of QRW_txt_rank using the density based re-ranking method (using cross-modal similarities).
      end if
    else {ENT (enriched query)}
      Rank the images according to one of the following strategies:
      if ENT_txt_rank then
        Rank the images according to the text similarity between the query enriched with lexical entailment
        (as described in Eq. (3) and the captions of all other images and consider the top $M$.
      else {ENT_cluster_rerank}
        Rank the images according to the text similarity between the query enriched with lexical entailment
        and the captions of all other images and consider the top $M$ such as for ENT_txt_rank
        Re-rank the top list of ENT_txt_rank using the clustering based re-ranking method (with visual
        or textual or cross-modal similarities).
      else {ENT_knndens_rerank}
        Rank the images according to the text similarity between the query enriched with lexical entailment
        and the captions of all other images and consider the top $M$ such as for ENT_txt_rank.
        Re-rank the top list of ENT_txt_rank using the density based re-ranking method (with visual or
        textual or cross-modal similarities).
      end if
    end if
    Optionally, merge several of the above mentioned ranked lists using the Round Robin approach described in 5.1.
  end for
```