XRCE's Participation in Wikipedia Retrieval, Medical Image Modality Classification and Ad-hoc Retrieval Tasks of ImageCLEF 2010

Stephane Clinchant, Gabriela Csurka, Julien Ah-Pine, Guillaume Jacquet, Florent Perronnin, Jorge Sánchez, Keyvan Minoukadeh

XRCE's Participation in Wikipedia Retrieval, Medical Image Modality Classification and Ad-hoc Retrieval Tasks of ImageCLEF 2010

Stéphane Clinchant[1,2], Gabriela Csurka[1], Julien Ah-Pine[1], Guillaume Jacquet[1], Florent Perronnin[1], Jorge Sánchez[1] and Keyvan Minoukadeh[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 year, XRCE participated in three main tasks of ImageCLEF 2010. The Visual Concept Detection and Annotation Task is presented in a separate paper. In this working note, we rather focus on our participation in the Wikipedia Retrieval Task and in two sub-tasks of the Medical Retrieval Task (Image Modality Classification and Ad-hoc Image Retrieval). We investigated mono-modal (textual and visual) and multi-modal retrieval and classification systems. For representing text we used either standard language model or a power law (log-logistic or smoothed power law) distribution-based information retrieval model. For representing images, we used Fisher Vectors improved by power and L2 normalizations and a spatial pyramid representation. With theses representations and simple linear classifiers we achieved excellent image modality classification both using mono-modal and combined textual and visual information. Concerning the retrieval performances, text based runs performed very well, but visual-only retrieval performances were in general poor showing that even state-of-the art image representations are insufficient to address these tasks accurately. However, we have shown that despite poor visual retrieval results, multi-modal runs that combine both visual and textual retrieval scores, can outperform mono-modal systems as long as the information fusion is done appropriately. As a conclusion we can say that our participation in these tasks was successful, as the proposed systems obtained leading positions both in retrieval and modality classification and for each type of run: text, image or mixed.

Keywords

Cross-modal Information Retrieval, Image Modality Classification, Medical Image Retrieval, Wikipedia Retrieval, Fisher Vector, Lexical Entailment, Query Expansion

1 Introduction

This year, XRCE participated in three main tasks of ImageCLEF 2010. The Visual Concept Detection and Annotation Task is presented in a separate paper. In this working note, we rather focus on our participation in the Wikipedia Retrieval Task [19] and in two sub-tasks of the Medical Retrieval Task (Image Modality Classification and Ad-hoc Image Retrieval) [15]. In participating in these tasks, our first motivation was to investigate how well do our textual and visual retrieval and classification systems perform on Wikipedia and Medical Image Collections. Our second aim was to experiment with different information fusion strategies to cope with the multi-modal (textual and visual) aspect of the tasks.

Concerning the text-based retrieval, two information retrieval models were considered: a standard language model (similar to the techniques used in our past participation in other tasks of

ImageCLEF [14]) and an information based model employing power law (log-logistic and smoothed power law) distribution proposed in [3]. These methods have shown high performances especially when combined with pseudo-relevance feedback or query expansion mechanisms were also applied.

In the case of the Medical Retrieval Task, our aim was to combine Lexical Textual Entailment inspired by Statistical Translation Models with query expansion mechanisms using external resources (in our case Wikipedia pages).

For representing images, we used the Improved Fisher Vectors (IFV) [18, 17]. Fisher Vectors have been successfully used in our previous ImageCLEF participations [14], and for this challenge we used an improved version with power and L2 normalizations and a spatial pyramid representation as suggested in [17]. These IFVs were used as the image representation in all our tasks. They showed excellent performances in the Image Modality Classification Task both using them alone or combined with text representation. Concerning retrieval results, even with the latter recent advances in visual similarity [18], using only images was not sufficient to address the Wikipedia and Medical Image Retrieval tasks. Indeed all visual-only runs performed poorly.

As the visual retrieval is an important element for the cross-modal similarities technique we used with success in the Photo Retrieval Tasks of past ImageCLEF sessions [14], we did not experiment with these techniques for this year's challenges. We decided to use other score aggregation methods where we could easily handle the asymmetric roles of the two modalities during the fusion process. Hence we designed and compared several fusion strategies. We have shown that despite poor visual retrieval results, our proposed aggregation techniques were able to outperform both the image-only and text-only retrieval systems.

The rest of the paper is organized as follows. In Section 2 we describe our text retrieval and query expansion models. In Section 3 we briefly describe the image representation with the improved Fisher Vectors. In section 4 we present and compare different runs we submitted and we conclude in Section 5.

2 Text Retrieval

We start from a traditional bag-of-words representation of pre-processed texts where pre-processing includes tokenization, lemmatization, and standard stopword removal. However, in some cases lemmatization might lead to a loss of information. Therefore before building the bag-of-words representation we concatenated a lemmatized version of the document with the original document.

Two information retrieval models were considered: a standard language model (similar to the techniques used in our past participation in other tasks of ImageCLEF [14]) and an information-based model employing a log-logistic distribution proposed in [3]. We also present a query expansion mechanism that appeared to be relevant to the Medical Retrieval Task. Finally, we briefly describe some details specific to the individual tasks.

2.1 Information Retrieval Techniques

To take into account the fact that one is comparing documents of different lengths, most IR models do not rely directly on the raw number of occurrences of words in documents, but rather on normalized versions of it. Language models for example use the relative frequency of words in the document and the collection[3]:

$$P(w|d) = \lambda \frac{x_w^d}{l_d} + (1-\lambda)\frac{\sum_d x_w^d)}{\sum_d l_d} \quad (1)$$

where x_w^d is the number of occurrences of word w in document d, l_d is the length of d in tokens after lemmatization and C is the corpus. Then we can define the similarity between the query $q = (q_1, ..., q_l)$ and a document d using the cross-entropy function:

$$CE(q|d) = \sum_{q_i} P(q_i|q) \log(\sum_w P(q_i|w)P(w|d)) \quad (2)$$

[3] Here we use Jelinek-Mercer interpolation but we can also use *e.g.* Dirichlet smoothing instead.

Other classical term normalization schemes include the well known Okapi normalization, as well as pivoted length normalization [20]. More recently, the concept of the amount of information brought by a term in a document has been considered in several IR models, inspired by the following observations made by Harter in [11]: the more a word deviates in a document from its average behavior given the collection, the more likely it is "significant" for this particular document. This can be easily captured in terms of information: if a word behaves in the document as expected in the collection, then it has a high probability p of occurrence in the document, according to the collection distribution, and the information it brings to the document, $-\log(p)$, is small. On the contrary, if it has a low probability of occurrence in the document, according to the collection distribution, then the amount of information it conveys is greater. This is the underlying idea of the information models proposed by Clinchant and Gaussier [3]: the log-logistic and smoothed power-law models.

These models are specified in three steps: the Divergence from Randomness (DRF) normalization of terms frequencies, the choice of a probability distribution to model these frequencies in the corpus and the mean information as the Relevance Score Vector (RSV). In the case of the log-logistic model, we have (for further details see [3]):

- DFR Normalization with parameter c: $t_w^d = x_w^d \log(1 + c\frac{avg_l}{l_d})$
- $Tf_w \sim \text{LogLogistic}(\lambda_w = \frac{N_w}{N})$
- Ranking Model:

$$RSV(q,d) = \sum_{w \in q \cap d} x_w^q \left[-\log P(Tf_w > t_w^d) \right] \tag{3}$$

where x_w^d and x_w^q are the numbers of occurrences of word w in document d and query q, N and N_w are the numbers of documents in the corpus and the number of those containing w, avg_l and l_d are the average document length and the length of the document d.

In the case of the smoothed power-law model we have the same steps but the Relevance Score Vector in the Ranking Model is replaced by (see details in [3]):

$$RSV(q,d) = \sum_{w \in q \cap d} -x_w^q \log\left(\frac{\lambda_w^{\frac{t_w^d}{t_w^d+1}} - \lambda_w}{1 - \lambda_w}\right) \tag{4}$$

2.2 Lexical Entailment - Statistical Translation Model

Berger and Lafferty [2] addressed the problem of information retrieval as a statistical translation problem with the well-known noisy channel model. This model can be viewed as a probabilistic version of the generalized vector space model. The analogy with the noisy channel is the following one: To generate a query word, a word is first generated from a document and this word then gets "corrupted" into a query word. The key mechanism of this model is the probability $P(v|u)$ that term u is "translated" by term v. These probabilities enable us to address a vocabulary mismatch, and some kinds of semantic enrichments. The problem now lies in the estimation of such probability models.

We refer here to a previous work [4] on lexical entailment models to estimate the probabilities. Lexical Entailment (LE) [4, 10, 7] models the probability that one term entails another. It can be understood as a probabilistic term similarity or as a unigram language model associated to a word (rather than to a document or a query). Let u be a term in the corpus, then lexical entailment models compute a probability distribution over terms v of the corpus $P(v|u)$. These probabilities can be used in information retrieval models to enrich queries and/or documents and to give a similar effect to use of a semantic thesaurus. However, lexical entailment is purely automatic, as statistical relationships are only extracted from the considered corpus. In practice, a sparse representation of $P(v|u)$ is adopted, where we restrict v to be one of the N_{max} terms that are the closest to u using an Information Gain metric[4].

[4] The Information Gain, aka Generalised (or average) Mutual Information [6], is used for selecting features in text categorisation [21, 9] or detecting collocations [8].

To summarize our approach, documents containing images were extracted from the corpus and we computed a lexical entailment measure on this collection with the model GI-M3 as in [4] or using Chi-square statistics. Finally, we expanded the query terms with words that were the most similar to the latter in terms of the chosen lexical entailment measure.

2.3 ImageCLEF Task-specific Text Processing

In this section we give some further details on the text retrieval processes that were designed or applied to a given task.

Wikipedia The Wikipedia Corpus consisted of images with their captions extracted from Wikipedia in different languages namely French, English and/or German. In addition, the participants were provided with the original Wikipedia pages in one or several languages in wikitext format. Similarly the textual part of the query was multilingual.

For each image, in addition to the given metadata (image's captions), we extracted from the Wikipedia pages the paragraph of the Wikipedia page in which the image was mentioned. Thus, two indexes were built for the collection: one for the captions (metadata) of the images and one for the paragraphs. The combination of the retrieval scores based on those two types of text, was done by a simple late fusion (mean average) after having normalized the scores between 0 and 1.

To cope with the multilingual nature of the wikipedia collection, we adopted an early fusion approach that we previously experimented within CLEF'08 Adhoc [5]. This early fusion of text amounts to merging the different languages of a document into a single document. Different language representations of queries are also merged in a similar fashion. Multilingual queries are then matched with multilingual documents as if there was no difference between languages.

Medical Task In the Medical Retrieval Task, although participants were provided with articles containing the image, we used only the image captions to build the term frequencies and the distributions of these frequencies (see 2.1). Nevertheless, the text from the articles containing the image was used as a corpus during the topic enrichment step with the model described in 2.2.

Furthermore we also experimented with query expansion mechanisms using external data. Without such techniques, we would have vocabulary mismatches between queries and documents. For example, for the topic 16 'dermatofibroma' we found no document containing "dermatofibroma" as a word. The Lexical Entailment methods using the articles enabled us to address some of the issues, but the coverage of the retrieval system can be improved by using an external knowledge base. Ideally, the use of a thesaurus such as Mesh and the resources provides by UMLS would have been preferable. However, with the lacks of experience (this was our first participation in the Medical Task) and time for extracting useful information from these resources, we used our usual tools to improve the coverage for a given query with information extracted from Wikipedia pages.

To do so we proceeded as follows. For each query, a set of related pages in Wikipedia was found in order to cover all query terms:

```
topic-1  | thoracic_aortic_dissection
topic-2  | Acute_Myeloid_Leukemia
topic-4  | congestive_heart_failure
topic-5  | brachial plexus nerve_block
```

We filtered redundant pages in order to have a unique coverage of query terms and to disambiguate some terms by taking the sense which was the most similar to the medical collection. Concerning the coverage, if the query contained a multiword expression related to a Wikipedia page, we used this page and not those related to the words within this multiword expression. For example, in topic 1, we can find the page "thoracic_aortic_dissection" in Wikipedia, then we use it and not the page "dissection". The disambiguation was done by computing a text similarity with the medical collection and by keeping the sense which had the higher score.

Then, for each of theses pages, we extracted only the hyperlinked text embedded in the body content. For each query, all those new terms were merged and we kept only the 20 terms that were the most frequent in the collection. This can be viewed as a query expansion mechanism provided by Wikipedia pages. Although, this expansion was expected to be noisy, it had the potential to improve recall. Finally, original textual runs were combined with these expanded query runs in order to maintain a high precision (see details in 4.3).

3 Image Representation

As for the image representation, we used an improved version [18, 17] of the Fisher Vector [16]. The Fisher Vector can be understood as an extension of the bag-of-visual-words (BOV) representation. Instead of characterizing an image with the number of occurrences of each visual word, it characterizes the image with the gradient vector derived from a generative probabilistic model. The gradient of the log-likelihood describes the contribution of the parameters to the generation process.

Assuming that the local descriptors $I = \{x_t, x_t \in \mathbb{R}^D, t = 1 \dots T\}$ of an image I are generated independently by Gaussian mixture model (GMM) $u_\lambda(x) = \sum_{i=1}^{M} w_i \mathcal{N}(x|mu_i, \Sigma_i)$, I can be described by the following gradient vector (see also [12, 16]):

$$G_\lambda^I = \frac{1}{T} \sum_{t=1}^{T} \nabla_\lambda \log u_\lambda(x_t) \tag{5}$$

where $\lambda = \{w_i, \mu_i, \Sigma_i, i = 1 \dots M\}$ are the parameters of the GMM. A natural kernel on these gradients is the Fisher Kernel [12]:

$$K(I, J) = {G_\lambda^I}' F_\lambda^{-1} G_\lambda^J, \qquad F_\lambda = E_{x \sim u_\lambda} \left[\nabla_\lambda \log u_\lambda(x) \nabla_\lambda \log u_\lambda(x)' \right]. \tag{6}$$

where F_λ is the Fisher information matrix. As it is symmetric and positive definite, F_λ^{-1} has a Cholesky decomposition $F_\lambda^{-1} = L'_\lambda L_\lambda$ and $K(I, J)$ can be rewritten as a dot-product between normalized vectors $\mathcal{G}_\lambda$ with: $\mathcal{G}_\lambda^I = L_\lambda G_\lambda^I$. We will refer to $\mathcal{G}_\lambda^I$ as the *Fisher Vector* (FV) of the image I.

In the case of diagonal covariance matrices Σ_i (we denote by σ_i^2 the corresponding variance vectors), closed form formulas can be derived for $\mathcal{G}_{w_i^d}^I$, $\mathcal{G}_{\mu_i^d}^I$, $\mathcal{G}_{\sigma_i^d}^I$, for $i = 1 \dots M$, $d = 1 \dots D$ (see details in [17]). As we do not consider $\mathcal{G}_{w_i^d}^I$ (the derivatives according to the weights), $\mathcal{G}_\lambda^I$ is the concatenation of the derivatives $\mathcal{G}_{\mu_i^d}^I$ and $\mathcal{G}_{\sigma_i^d}^I$ and is therefore $N = 2MD$-dimensional.

While this representation was successfully used with the L1 normalization and L1 norm based similarity between the Fisher Vectors in our previous ImageCLEF participation [14], this time we used an improved version of it with Power and L2 normalization and simple dot product similarity between them as suggested in [17]. In order to obtain these Improved Fisher Vectors we first use a Power Normalization with $\alpha = 0.5$:

$$PN(\mathcal{G}_{\lambda_n}^I) = \text{sign}(\mathcal{G}_{\lambda_n}^I) |\mathcal{G}_{\lambda_n}^I|^\alpha, \ n = 1 \dots N. \tag{7}$$

The aim of this normalization is to make the distribution of features in a given dimension n less peaky around zero. Then, these vectors are further L2 normalized to discard image-independent (*i.e.* background) information (see further details for both normalizations in [17]).

Another improvement made to our Fisher Vector based image representation was the use of the spatial pyramid representation by Lazebnik *et al.* to take into account the rough geometry of a scene [13]. The main idea is to repeatedly subdivide the image and represent each layout as a concatenation of the representations (in our case Fisher Vectors) of individual sub-images. As we used three spatial layouts (1×1, 2×2, and 1×3), we obtained 3 image representations of respectively N, $4N$ and $3N$ dimensions.

As low level features we used our usual (see for example [14]) SIFT-like Orientation Histograms (ORH) and local color statistics (COL), *i.e.* local color means and standard deviations in the R,G and B channels, both extracted on regular multi-scale grids and reduced to 50 or 64 dimensional with Principal Component Analysis (PCA). With the three different spatial layouts mentioned above each image was finally represented by 6 different high level feature vectors (referred to as IFV for Improved Fisher Vector).

4 Runs Description

4.1 Wikipedia Retrieval

The Wikipedia Retrieval task consists of multilingual and multimedia retrieval. The collection contains images with their captions extracted from Wikipedia in different languages namely French, English and German. In addition, participants were provided with the original Wikipedia pages in wikitext format. The task consisted in retrieving as many relevant images as possible from the aforementioned collection, given a textual query translated in the three different languages and one or several query images.

We submitted different types of runs: mono-media runs and multimedia (mixed) runs with different fusion approaches.

Concerning the pure visual retrieval, we used 6 Improved Fisher Vector (IFV), corresponding to the two different low level features (ORH and COL) with the 3 spatial-layout (1x1, 2x2, 1x3) as described in section 3. The 6 IFVs were used independently to rank the Wikipedia images using the dot product as similarity measure and the 6 scores were weighted giving higher weights (0.7) to the IFV based on orientation histograms compared to the IFV based on color (0.3) before averaging them. The pure visual run results are mentioned in line 1 of Table 1. The performances are very poor and this shows that even recent advances in visual similarities [17] are not sufficient to address the Wikipedia Retrieval task efficiently.

Pure text based retrieval performs much better. Details on the definition of the methods and the underlying retrieval models used are mentioned in section 2. Accordingly, from lines 2 to 4 of Table 1, we report 3 pure text runs which are based on recent models presented in [3]. Line 2 is based on a smoothed power-law retrieval model while the latter two ones rely on a log-logistic probability distribution. It appeared that the smoothed power-law model performed better than the log-logistic ones, achieving 23.61% of Mean Average Precision (MAP). Nevertheless, in our combined modality runs, we used the former run when combining with visual retrieval. All text runs use both captions (metadata) and paragraphs as explained in section 2. However, the difference between the text runs LGD_ALL_METANOPRF_PARAGPRF20 (line 3) and LGD_ALL_META_PARAG (line 4) is that the former uses on the paragraphs, an additional pseudo-relevance feedback mechanism (PRF), introduced in [3]. This strategy appeared to be beneficial since the results obtained for the former run are better.

Runs reported from lines 5 to 16 are mixed, *i.e.* they make use of both visual and textual similarities. We grouped the runs according to the text run used in the combination. Consequently, from lines 5 to 10, we report the results of mixed runs that used the text run from line 3 (T2=LGD_ALL_METANOPRF_PARAGPRF20) whereas from line 11 to line 16 we report results of the mixed runs that used the text run from line 4 (T3=LGD_ALL_META_PARAG). The runs with a $\star$ symbol (lines 5 and 12) in Table 1 were not submitted but we added them in order to have a better comparison between the results obtained using both aforementioned text runs.

As explained earlier, T2 gave a better result than T3. Moreover, whatever the combination technique we used, we always observe that the run using the PRF on the paragraphs dominates the other one. Therefore, we will only comment on the results of mixed runs reported from lines 5 to 10 in Table 1. Similar conclusions can be deduced for the runs from line 11 to line 16.

From our past experiences with ImageCLEF [14], we observed that textual similarity played a core role, but that it could be complemented with visual similarity if combined in an appropriate way. In other words, when combining image and text similarities in the context of multimedia

retrieval, we paid attention to the fact that these two media should not be given, in general, symmetric roles during the fusion process.

Consequently, our first fusion strategy was to start with using the text similarities as a pre-filtering stage before applying the visual similarities. In practice, we first selected the top 2000 images according to the textual similarities. Then, we re-ranked this pre-filtered top-list using the visual similarities only. This pre-filtering step based on textual similarities allowed us to significantly increase the pure visual run from 5.53% (line 1) to 18.05% of MAP (line 5). In particular, it gives better results than both image and text runs in terms of Precision at 20 (P@20). Still, this run entitled SEQ_RERANKING-T2 (line 5) does not perform as well as pure textual similarities T2 (line 3) in terms of MAP.

Then, going beyond this simple re-ranking strategy, we investigated the combination of the re-ranked run SEQ_RERANKING-T2, with the pure textual run T2. We basically normalized the two runs so that the scores were between 0 and 1 and we applied different aggregation functions. The different results are shown from line 6 to line 10.

Our first approach consisted in applying a simple weighted mean average operator. In that case, the submitted runs were given by the following aggregated score:

$$w^R s_{TRV}(I) + w^T s_{TR}(I) \tag{8}$$

where I is an image of the collection, s_{TRV} is the normalized score distribution based on the visual similarities but after having selected the top 2000 images according to the textual similarities, s_{TR} is the normalized score distribution based on the textual similarities and w^R and w^T are their respective weights in the late fusion method.

Lines 6, 8 and 10 with runs' titles beginning with FUSION are the results obtained with the following parameters: line 6 with $w^R = 0.5$ and $w^t = 0.5$; line 8 with $w^R = 0.7$ and $w^t = 0.3$; and line 10 with $w^R = 0.3$ and $w^t = 0.7$.

Our second fusion technique uses an aggregation operator that was introduced in [1]. This approach attempts to reinforce the aggregated score of an image assuming that there is a strong relationship between the two scores. In that case, the submitted run reported in line 7 of Table 1 (called CONFAGmin-T2) was given by the following aggregated score:

$$s_{TRV}(I) + s_{TR}(I) + \min(s_{TRV}(I), s_{TR}(I)) \tag{9}$$

Our last combination approach follows the idea that image similarities are less reliable then text similarities and when combining both of them one should adapt the combination weights accordingly. In that perspective, we propose to set the combination weights as functions of the visual similarity values. We particularly tested linear functions with $w^R(s_{TRV}(I)) = \alpha s_{TRV}(I)$ and $w^T(s_{TRv}(I)) = 1 - \alpha s_{TRV}(I)$. This leads to the following aggregation method:

$$\alpha(s_{TRV}(I))^2 + (1 - \alpha s_{TRV}(I)) s_{TR}(I) \tag{10}$$

Line 9 corresponds to the results obtained using the aforementioned aggregation function with $\alpha = 0.7$ (CONFα-T2).

From Table 1, we can make the following observations:

- Image similarities are not sufficient to address the task accurately and text similarities perform much better.
- Using as a preliminary stage the text similarities in order to select a first relevant list of images and re-ranking the latter list on the basis of image similarities only allows us to boost the precision (P@20) but not the MAP. The important conclusion to make here and towards multimedia retrieval, is that visual similarities might be relevant but they need to be filtered. In that case, a text-based pre-filtering step is very important in order to select relevant visual similarities by removing images that are not relevant "semantically" despite of being similar to the image query "visually".

Table 1. Wikipedia retrieval: overview of the performances of our different runs. For reference we have included the best performing run from the other participants (line 17).

	RUN	Modality	MAP	P@20
1	XRCE_AUTO_IMG	Visual	0.0553	0.2686
2	ALL_DFRPMETAPRF5_DFRPARAGPRF20	Textual	0.2361	0.4393
3	LGD_ALL_METANOPRF_PARAGPRF20	Textual	0.2045	0.4200
4	LGD_ALL_META_PARAG	Textual	0.1903	0.4000
5	⋆ SEQ_RERANKING-LGD_ALL_METANOPRF_PARAGPRF20	Mixed	0.1805	0.4493
6	**FUSION_TEXTLGD_ALL_METANOPRF_PARAGPRF20**	**Mixed**	**0.2765**	**0.5193**
7	CONFAGmin-TEXTLGD_ALL_METANOPRF_PARAGPRF20	Mixed	0.2681	0.5257
8	FUSION_TEXTLGD_ALL_METANOPRF_PARAGPRF20-Rer7_Text3	Mixed	0.2627	0.5407
9	CONF0.7-TEXTLGD_ALL_METANOPRF_PARAGPRF20	Mixed	0.2532	0.4986
10	FUSION_TEXTLGD_ALL_METANOPRF_PARAGPRF20-Rer3_Text7	Mixed	0.2493	0.4743
11	SEQ_RERANKING-LGD_ALL_META_PARAG	Mixed	0.1747	0.4471
12	⋆ FUSION_TEXTLGD_ALL_META_PARAG	Mixed	0.2660	0.5193
13	CONFAGmin-TEXTLGD_ALL_META_PARAG	Mixed	0.2575	0.5164
14	FUSION_TEXTLGD_ALL_META_PARAG-Rer7_Text3	Mixed	0.2527	0.5336
15	CONF0.7-TEXTLGD_ALL_META_PARAG	Mixed	0.2424	0.4907
16	FUSION_TEXTLGD_ALL_META_PARAG-Rer3_Text7	Mixed	0.2415	0.4664
17	best non-XRCE run	Textual	0.2251	0.3871

– Combining textual similarities with text-based pre-filtered visual similarities, dramatically outperforms mono-media runs. Any of the aggregation functions used in fusing the two aforementioned runs allows us to increase both the MAP and the P@20 of the best mono-media runs. It appeared that the best aggregation strategy was the simplest one: the mean average. In that case, it improved the text run performance from MAP=20.45% to 27.65% which corresponds to more than a 35% increase.

4.2 Medical Image Modality Classification Task

Imaging modality is an important aspect of the image for medical retrieval. Therefore, within the Medical Retrieval Task in 2010 [15], a first sub-task was image modality classification. Participants were provided a training set of 2000 images that have been classified into one out of 8 modalities (CT, MR, XR etc). These images also contained captions that participants could use in addition to the visual information both at training and test time. The measure used for this sub-task was the classification accuracy.

In our experiments we investigated mono-modal and mixed modality based classification. Concerning the pure visual-based classifiers, we trained 2 linear classifiers per modality (using the one-versus-all scheme) corresponding to the two different low level features (ORH and COL). Linear SVM classifiers with hinge loss using the primal formulation were trained with the Stochastic Gradient Descent (SGD) algorithm[5] using the Improved Fisher Vectors (IFV) as described in section 3 and a single spatial layout (1x1). The SVM scores were combined by weighted averaging color and SIFT features for color images. We only used the ORH based IFV for gray-scale images. The weights for each modality were tuned by maximizing classification accuracy using a 5-fold cross validation scheme on the training set.

Concerning our text based modality classification, we used two different representations and hence two classifiers. The first one was based on a pattern matching (PM) technique where we searched for the modalities in the image captions. The second one was based on a binarized bag-of-words representation.

The pattern matching was based on the information describing the modalities in the modalityClassificationREADME.txt file. Indeed, this file provides for each modality a description that

[5] An implementation is available on Léon Bottou's web-page: `http://leon.bottou.org/projects/sgd`.

contains a list of expressions mainly related to image sub-modalities grouped in that category (i.e. the PX modality description "PX: optical imaging including photographs, micrographs..." contains expressions "PX","optical imaging", "photograph", "micrograph"...). Therefore, for each image, we detected (matched) these expressions in the corresponding captions. However, within the corpus, many documents contain several images with the same caption while the latter do not necessarily have the same modalities. Therefore, we reduced[6] the image caption for a targeted image name, to:

- sentences related to the image reference[7],
- sentences containing multiple image references including the targeted one,
- and the introductory sentences (all sentences before the first occurrence of an image reference).

Consequently, the outputs of the pattern matching (PM) decision function can be a single modality, a set of potential modalities or an empty set.

The binarized bag-of-words representation consisted in a vector indicating for each word whether it appears in this document or not (in our case image caption). Note that for this representation we did not apply the aforementioned caption reduction but used them as they were provided. The feature vectors of the training set made of 2000 items with their modality labels, were then used to train a classifier per modality (one-versus-all scheme). The aim of this representation was to go beyond pattern matching and learn other words related to different modalities. To train the linear classifiers we used the logistic regression classifier (LRC) from the liblinear toolkit[8] with Laplace Prior (L1 norm).

To combine the outputs of the two classifiers, we proceeded as follows:

- If we have no output from PM, the modality is given by:

$$c^* = \arg\max_{c \in C} s_T^c(I) \tag{11}$$

  where $s_T^c(I)$ is the output of the classifier LRC trained for the modality c.
- If PM outputs one or more modalities and c^* is amongst them, the modality is again c^*.
- Otherwise the modality amongst the PM outputs that has the highest score $s^c(I)$ was selected (C in equation 11 is reduced to the PM outputs).

In the case of mixed modalities, we first combined the LRC scores with the SGD visual scores (both normalized to get values between 0 and 1) before combining with the PM output as described above ($s_T^c(I)$ is replaced by the combined scores in equation 11).

The results of the different classifiers are shown in Table 2. As these results show, the pure text modality classifier slightly outperforms the visual only information based classifier and both of them were outperformed by the combination of visual and textual modalities.

Table 2. Modality Classification task: overview of the performances of our different runs. For reference we have included the best performing run from the other participants (fourth line).

RUN	Modality	ACC
XRCE_MODCLS_COMB	Mixed (textual + visual)	0.94
XRCE_TXT	Textual	0.90
XRCE_VIS	Visual	0.87
best non-XRCE run	Mixed	0.93

[6] When it was possible to detect automatically.
[7] For example references might be *(a)* or *(a-c)*
[8] `http://www.csie.ntu.edu.tw/~cjlin/liblinear/`

4.3 Medical Ad-hoc Image Retrieval Task

For this task [15] the participants were given a set of 16 textual queries with 2-3 sample images for each query. The queries were classified into textual, mixed and semantic queries, based on the methods that are expected to yield the best results. The results of the modality classification can be used to filter the search in this sub-task. However, instead of filtering we preferred to aggregate the classification scores with the retrieval scores as explained below.

In our experiments we wanted to investigate both mono-modal and mixed-modality retrieval.

Visual Only Concerning the pure visual retrieval, in contrast to our Wikipedia runs, we used only the 3 Improved Fisher Vector (IFV), corresponding to the orientation histograms (ORH) with the 3 spatial-layout (1x1, 2x2, 1x3). The 3 IFVs were used independently to rank the Wikipedia images using the dot product as similarity measure and the 3 scores were simply summed[9] and then normalized to be between 0 and 1. We denote this score by $s_{VR}(I)$ (visual retrieval score). This score was then further combined with the visual only modality score as follows:

- First the query images were classified by the SGD classifiers using $c_k^*(t) = \arg\max_{c \in C} s_{SGD}^c(I_k^t)$, where $c_k^*(t)$ is the modality assigned to the image I_k^t of the topic t.
- If for all images I_k^t in a topic (if multiple image query) the same modality c^* was assigned, the scores $s_{SGD}^{c^*}(I)$ were added to $s_{VR}(I)$ to rank the images of the dataset. The aim was to boost the scores of the images corresponding to that modality.
- In contrast, if we had difficulty to retain a given modality (as e.g. for the last topics) we preferred to rank the images based only on the retrieval scores $s_{VR}(I)$.

The performance of this run shown in line 1 of Table 3 is very poor, leading to the conclusion that visual only information is insufficient to address the task.

Text Only Concerning the textual retrieval, we submitted two types of text-only retrieval. For the first one, we simply ranked the text corresponding to some cross-entropy-based text similarity or Relevance Score Vector between the query text and the image caption. For the second, we aggregated these scores with text-only classification scores.

First we describe a few textual retrieval runs (see section 2 for technical details). Several such textual runs can be created depending on:

- Different information retrieval models: Language Model or log-logistic model (the latter was used if LOGIT or LGD appears in the run names, the former otherwise).
- Different smoothing parameters: we mostly used the classical Jelinek-Mercer interpolation except for the run DIR_TXT (line 2) where the Dirichlet smoothing was used.
- Different query and document enrichment models: statistical translation model (referred to as AX) and/or[10] using the Wikipedia Corpus (referred to as WIKI).
- Different statistics: using Chi2 with the log-logistic model is referred as CHI2_LOGIT and using Chi2 statistics instead of GI-M3 in the statistical translation model is referred as CHI2AX.

While only a single pure text retrieval run was submitted (line 2), nevertheless, we describe them here since they were used (referred to as textual retrieval score with $s_{TR}(I)$) in the combination with the modality classification outputs and/or with the visual retrieval scores (see below). To aggregate the textual retrieval scores $s_{TR}(I)$ with the text based modality scores, we first extracted the modality c from the query text by pattern matching (PM) and selected the corresponding $s_{LRC}^c(I)$ score[11] (see section 4.2). If no modality is found, only the $s_{TR}(I)$ is used to rank. Two such

[9] Note that, due to the linearity property of the dot product, this is equivalent to the dot product of the concatenation of all three IFV vectors.

[10] Note that WIKI_AX means that we combined the lexical entailment based scores with the scores obtained with Wikipedia based query expansion to obtain $s_{TR}(I)$ for further fusion.

[11] This is the output of the LRC image modality classifier trained on the binarized bag-of-word feature vectors for the modality c.

runs were submitted (lines 3 and 4 in Table 3). While not directly comparable, we can nevertheless see that these runs led to better performances than DIR_TXT suggesting that combining text retrieval with modality classification helps. Furthermore, WIKI_AX_MOD_late led to a much better MAP than CHI2AX_MOD_late, showing the benefit of using external data (Wikipedia) for query expansion.

Mixing Textual and Visual Information Finally, we further used both visual and textual information. We also experimented with further combining these scores with modality classification scores as follows:

$$w^V s_{VR}(I) + w^M s^c_{LRC}(I) + w^T(t) s_{TR}(I). \quad (12)$$

Here $s_{VR}(I)$ and $s_{TR}(I)$ are the normalized visual and textual retrieval scores (see above) and $s^c_{LRC}(I)$ is the output of the image modality classifier for the modality c using only textual information (LRC scores). The modality c is extracted from the query text of topic t by pattern matching (PM) and $s^c_{LRC}(I)$ is set to zero if no modality was found.

We made the weights of the textual retrieval score $w^T(t)$ dependent on the type of topic t. Hence, we used 0.5 for topics with query type $Visual$, 1 for type $Mixed$ and 1.5 for type $Semantic$, in order to increase or decrease the importance of the textual score. The weights w^V and w^M took values 1 or 0. Hence setting w^V to 0 lead to pure textual based retrieval runs discussed above (such as WIKI_AX_MOD_late and CHI2AX_MOD_late depending on the $s_{TR}(I)$ used). If we further set w^M to 0, we get runs without using the modality (e.g. DIR_TXT). If we set $w^V = 1$ and $w^M = 0$ we get the classical late fusion (e.g. AX_LGD_IMG_late) and setting $w^V = 1$ and $w^M = 1$ we get a late fusion based on textual and visual retrieval scores combined with modality classifier scores (CHI2_LOGIT_IMG_MOD_late, WIKI_LGD_IMG_MOD_late).

Finally the runs AX_rerank and AX_rerank_comb consist in ranking images based on $s_{TR}(I)$ and then re-ranking the top N=1000 relevant images using respectively $s_{VR}(I)$ and $s_{VR}(I) + w^T(t) s_{TR}(I)$.

The MAP, bPref and P10 results of the runs described above are shown in Table 3. These results show that the poor performance of image retrieval for this task, decreases the performances when the visual modality is not appropriately combined with the other modalities (as in AX_rerank or WIKI_AX_IMG_MOD_late). We can also note that the run AX_rerank_comb obtained the best performance on (MAP and bPref) while the run WIKI_AX_MOD_late has the best performing P10 among all runs.

However, it is difficult to make conclusions about these runs that performed better than others from this table. Indeed, we need to evaluate different text runs with and without combining them with modality classifier and/or visual retrieval scores. Our intention is to do such analysis as soon as the image relevance scores are made available by the organizers.

5 Conclusion

This year we have participated with success in two new main tasks, namely the Wikipedia Retrieval Task and two sub-tasks of the Medical Retrieval Task (Image Modality Classification and Ad-hoc Retrieval). In all cases, we obtained leading positions both in retrieval and modality classification, and for each type of run: text-only, visual-only and mixed. We achieved excellent text based retrieval results and despite the fact that pure visual based retrieval led to poor results, when we appropriately combined them with our text ranking we were able to outperform the latter showing that multi-modal based systems can be better than mono-modal ones.

Acknowledgments

This work was partially supported by the European Project PinView FP7/2007-2013 and the French National Project Fragrances ANR-08-CORD-008. We would like also to acknowledge Craig Saunders for his proofreading and useful comments on the paper.

Table 3. Overview of the performances of our different runs. For reference we have included the best performing run from the other participants (line 11).

	RUN	Modality	MAP	bPref	P@10
1	IMG_max	Visual	0.0029	0.0069	0.0063
2	DIR_TXT	Textual	0.2722	0.2837	0.4
3	CHI2AX_MOD_late	Textual	0.2925	3027	0.4125
4	WIKI_AX_MOD_late	Textual	0.338	0.3828	**0.5062**
5	CHI2_LOGIT_IMG_MOD_late	Mixed	0.3167	0.361	0.3812
6	AX_rerank_comb	Mixed	**0.3572**	**0.3841**	0.4375
7	AX_rerank	Mixed	0.0732	0.1025	0.1063
8	AX_LGD_IMG_late	Mixed	0.3119	0.3201	0.4375
9	WIKI_LGD_IMG_MOD_late	Mixed	0.2343	0.2463	0.3937
10	WIKI_AX_IMG_MOD_late	Mixed	0.2818	0.3279	0.3875
11	best non-XRCE run	Textual	0.3235	0.3109	0.4687

References

1. Julien Ah-Pine. Data fusion in information retrieval using consensus aggregation operators. In *Web Intelligence*, pages 662–668, 2008.
2. Adam Berger and John Lafferty. Information retrieval as statistical translation. In *In Proceedings of the 1999 ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 222–229, 1999.
3. Stéphane Clinchant and Eric Gaussier. Information-based models for ad hoc ir. In *SIGIR '10: Proceeding of the 33rd international ACM SIGIR conference on Research and development in information retrieval*, pages 234–241, New York, NY, USA, 2010. ACM.
4. Stéphane Clinchant, Cyril Goutte, and Éric Gaussier. Lexical entailment for information retrieval. In *Advances in Information Retrieval, 28th European Conference on IR Research, ECIR 2006, London, UK, April 10-12*, pages 217–228, 2006.
5. Stéphane Clinchant and Jean-Michel Renders. Multi-language models and meta-dictionary adaptation for accessing multilingual digital libraries. In *Evaluating Systems for Multilingual and Multimodal Information Access, 9th Workshop of the Cross-Language Evaluation Forum, CLEF 2008, Aarhus, Denmark, September 17-19, 2008, Revised Selected Papers*, pages 83–88, 2008.
6. Bernard Colin. Information et analyse des données. *Pub. Inst. Stat. Univ. Paris*, XXXVII(3–4):43–60, 1993.
7. Ido Dagan, Oren Glickman, and Bernardo Magnini. The PASCAL recognising textual entailment challenge. In *PASCAL Challenges Workshop for Recognizing Textual Entailment*, 2005.
8. Ted Dunning. Accurate methods for the statistics of surprise and coincidence. *Computational Linguistics*, 19(1):61–74, 1993.
9. George Forman. An extensive empirical study of feature selection metrics for text classification. *Journal of Machine Learning Research*, 3:1289–1305, 2003.
10. Oren Glickman, Ido Dagan, and Moshe Koppel. A probabilistic classification approach for lexical textual entailment. In *Proceedings of the Twentieth National Conference on Artificial Intelligence (AAAI-05)*, 2005.
11. S. P. Harter. A probabilistic approach to automatic keyword indexing. *Journal of the American Society for Information Science*, 26, 1975.
12. T. Jaakkola and D. Haussler. Exploiting generative models in discriminative classifiers. In *Advances in Neural Information Processing Systems 11*, 1999.
13. S. Lazebnik, C. Schmid, and J. Ponce. Beyond bags of features: Spatial pyramid matching for recognizing natural scene categories. In *CVPR*, 2006.
14. H. Müller, P. Clough, Th. Deselaers, and B. Caputo, editors. *Leveraging image, text and corss-media similarities for diversity-focused multimedia retrieval*, volume 32 of *The Information Retrieval Series*. Springer, 2010.
15. Henning Müller, Jayashree Kalpathy-Cramer, Ivan Eggel, Steven Bedrick, Charles E. Kahn Jr., and William Hersh. Overview of the clef 2010 medical image retrieval track. In *Working Notes of CLEF 2010, Padova, Italy*, 2010.

16. F. Perronnin and C. Dance. Fisher kernels on visual vocabularies for image categorization. In *CVPR*, 2007.
17. F. Perronnin, J. Sánchez, and Y. Liu. Large-scale image categorization with explicit data embedding. In *CVPR*, 2010.
18. Florent Perronnin, Yan Liu, Jorge Sánchez, and Hervé Poirier. Large-scale image retrieval with compressed fisher vectors. In *CVPR*, 2010.
19. Adrian Popescu, Theodora Tsikrika, and Jana Kludas. Overview of the wikipedia retrieval task at imageclef 2010. In *Working Notes of CLEF 2010, Padova, Italy*, 2010.
20. Amit Singhal, Chris Buckley, and Mandar Mitra. Pivoted document length normalization. In *SIGIR '96: Proceedings of the 19th annual international ACM SIGIR conference on Research and development in information retrieval*, pages 21–29, New York, NY, USA, 1996. ACM.
21. Yiming Yang and Jan O. Pedersen. A comparative study on feature selection in text categorization. In *Proceedings of ICML-97, 14th International Conference on Machine Learning*, pages 412–420, 1997.