Using Visual Concept Features in a Multimodal Retrieval System for the Medical collection at ImageCLEF2012

A. Castellanos[1], J. Benavent[2], X. Benavent[2], A. García-Serrano[1], E. de Ves[2]

[1] Universidad Nacional de Educación a Distancia, UNED
[2] Universitat de València

xaro.benavent@uv.es,{acastellanos,agarcia}@lsi.uned.es

Abstract. The main goal of this paper is to present our experiments in the classification modality and in the ad-hoc image retrieval tasks with the Medical collection at ImageCLEF 2012 Campaign. This edition we focus on applying new strategies for both the textual and the visual subsystems included in our multimodal retrieval system. The visual subsystem has focus on extending the low-level features vector with concept features. These concept features have been calculated by means of a logistic regression model. The textual subsystem has focus on applying a query reformulation to remove general and domain stop-words, trying to produce a query with only medical-related terms. We have not obtained the results as good as obtained at the Photo annotation retrieval subtask using similar techniques. Therefore, a deep analysis for the Medical collection will be done.

Keywords: Multimedia Retrieval, Concept Features, Low-level features, Logistic regression relevance feedback.

1 Introduction

In this paper we present our experiments in ImageCLEF 2011 Campaign at Medical Image retrieval task [**¡Error! No se encuentra el origen de la referencia.**1]. In this campaign, we participate in two sub-tasks of the Medical Retrieval Tasks: Image Modality Classification and Ad-hoc Image Retrieval. The work done in this edition is building on the knowledge acquired in previous participations both at Medical Retrieval Task [6] and at Wikipedia Retrieval Task [3,10], using multimodal retrieval approaches.

Regarding the textual retrieval subsystem, we apply partially the successful technique tested last year [10] (2nd in textual category). This is based on the preprocessing of the query in order to delete common and domain stopwords (i.e generic terms not related to medical domain like image, photo and so on). Unlike the work presented in [10], in this year we have decided not to use the modality classification of the images. This is due to that the possible improvements are highly dependent of the query type and query content; as was shown in our in-depth analysis of the results of last year, presented in [7].

Concerning to the visual retrieval subsystem it uses the low-level features for image retrieval. This low-level information although gives quite enough results depending on the visual information of the query is not able to reduce the "semantic gap" in a semantic complex query. Our proposal [4] is to generate concept features extracted from the low-level features to obtain the probability of the presence of each trained category. We call this new vector, the expanded low-level concept vector that is calculated for each image of the collection and also for the example images of the query to process the retrieval task. A model for each category is trained using a logistic regression [12]. We use these regression models to extract the concept features from the low-level features and construct the expanded concept features vector for the retrieval process.

It is our first participation at the classification task with four visual runs submitted. We have adapted our regression model to act as a classifier for the classification task. A model for each of the categories have been trained and tested.

Section 2 describes the visual approach based on a regression model acting as a classifier for the modality classification subtask. Section 3 explains our multimodal retrieval system use for the ad-hoc image-based retrieval subtask. After that section 4 shows the submitted runs and the results obtained for modality classification and retrieval. Finally, in section 5 we extract conclusions and outlines possible future research lines.

2 Modality classification

We train a logistic regression model [12] for each of the 31 categories given by the 2012 medical classification subtask. Each trained model predicts the probability that a given image belongs to a certain category.

The medical classification task gives to the participants a training set, I_s, for each of the categories. Being I_s^P the training image set for each category (the relevant images), and I_s^N the set that not belong to a certain category (non relevant images). The logistic regression analysis calculates the probability for a given image to belong to a certain category. Each image of the training set, I_s is represented by a K-dimensional low-level features vector $\{x_1, .., x_i, .., x_k\}$. The relevance probability for a certain category c_i for a given image I_j will be represented as $P_{c_i}(I_j)$. A logistic regression model can estimate these probabilities. Let us consider for a binary Y, and k explanatory variables $x = (x_{1,} \dots, x_k)$, the model for π(x) = P(Y=1| X) (probability $Y = 1$) for the x values $logit\ [\pi(x)] = \alpha + \beta_1 x_1 + \cdots + \beta_k x_k$, where logit (π(x))=ln(π(x) / (1-π(x)). The model parameters are obtained by maximizing the likelihood estimator (MLE) of the parameter vector β by using an iterative method.

We have a major difficulty when having to adjust an overall regression model in which we take the whole set of variables into account because the number of selected images (the number of positive plus negative images, k) is typically smaller than the number of characteristics (k < p). In this case the adjusted regression model has as many parameters as the amount of data and many relevant variables could be not considered. In order to solve this problem our proposal is to adjust different smaller regression models: each model considers only a subset of variables consisting of semantically related characteristics of the image. Consequently each sub-model will associate a different relevance probability to a given image x and we have to combine them in order to rank the database according to the image probability or image score (Si).

The explanatory variables $x = (x_{1,} \ldots, x_k)$ to train the model are the visual low-level features based on color and texture information that are calculated by our group. We have a low-level features vector of 293 components divided by five different visual information families.

- Color information: We calculate global and local histograms of the image.
  - Global color: It is a feature vector of 30 components represents the color information of the complete image. Each of these components represents a bin on a HS (hue-saturation) histogram of size 10 x 3.
  - Local color: Local histograms have been calculated by dividing the images into four fragments of the same size. A bi-dimensional HS histogram with 12x4 bins is computed for each patch, being 48 components for each patch, and a total of 192 components.
- Texture information: Two types of texture feature are computed:
  - The granulometric distribution functions [2], using the coefficients that result in fitting the distribution function with a B-spline basis. We calculated for two different structuring elements: horizontal and vertical segment. We have 31 components for granulometric distribution with horizontal segment and 31 components for vertical segment.
  - The Spatial Size Distribution [2] using a horizontal segment as structuring element. We have a 9 components vector for the spatial size distribution.

3 Ad-hoc image-based retrieval subtask

The overall system includes three main subsystems: the TBIR (Text-Based Image Retrieval), the CBIR (Content-Based Image Retrieval), and the Fusion subsystem (see Fig. 2). Both the textual (TBIR) and the visual subsystem (CBIR) obtain a ranked list of images based on similarity scores (St and Si) for a given query. Firstly, TBIR uses the textual information from the annotations (metadata and articles) to obtain these scores (St). This textual pre-filtered list is then used by the CBIR sub-system. It extracts the visual information from the given example images of the query and generates a similarity score (Si). The fusion sub-system is in charge of merging these two lists of results, taking into account the scores and rankings, in order to obtain the final result list.

3.1.1 Text-based Index and Retrieval

This module is in charge of the textual-based indexing and retrieval, using the text associated with each image in the collection.

In order to be able to manage the textual information of the collection, a preprocessing step is carried out, before of the indexing. Later, it has been carried out the indexing of the images for their subsequent retrieval. To indexing the collection, Solr[1], a search platform from Lucene[2] project, is used. The retrieval process is done through Solr too. The result of this retrieval process is a normalized image list for each query. Below, is explained in more detail each of the different stages performed by TBIR module:

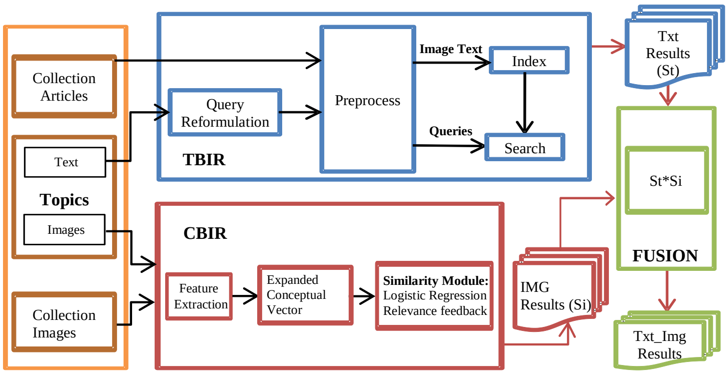

Fig. 1. - System Overview for the ad-hoc content image retrieval subtask.

- **Query Reformulation:** The original queries are reformulated in order to remove common and domain stopwords (e.g: image). No other process is done.
- **Preprocess:** Textual information (both at images description and queries description) is preprocessed : 1) special characters deletion: characters with no statistical meaning, like punctuation marks or blanks, are eliminated; 2) stopwords detection: deletion of semantic empty words in English language (e.g: the, an…), 3) stemming: reduction of word to their base form, for this purpose we use a Porter Algorithm implementation provided by Solr and, finally, 4) convert all words to lower case.
- **Indexing:** Because the collection of this edition is different from the last edition, it was necessary to index the new collection. The indexing is done automatically by Solr, using Lucene operation.
- **Searching:** The search process is also automatically done by Solr over Lucene operation. The score function used for calculating the similarity between a given query and the documents is BM25. The results are transformed to the TRECEval format, in order to merge these textual results with visual results and check the results using the UV tool [7].

[1] http://lucene.apache.org/solr/
[2] http://lucene.apache.org/java/docs/index.html

3.1.2 Content-Based Information and Visual Retrieval

The work of the **CBIR subsystem** is based on three main stages: Extraction of the low-level and the concept features of the images, and the calculation of the similarity (Si) of each of the images to the image examples given by a query.

1. **Extraction of low-level features:** The first step in the CBIR system is to extract the visual low-level and the concept features for all the images of the database as well as from the example images given in each question. The low-level features we use are calculated by our group and give color and texture information about the images. These features are the same that we have used for the modality classification task (see section 2.1 for more detailed information).
2. **Calculating the Concept features vector:** The regression models trained for each of the concepts gives for each image on the database and for the example query the probability of the presence of each concept $P_{c_i}(I_j)$. With this probability information for each concept, we extend the low-level features vector to m components, being m the number of concepts trained. Each image I_j on the database is described by the extended vector $F(I_j) = (x_1, \dots x_k, c_1, .., c_m\} \in R^{k+m}$.
3. **Similarity Module:** The similarity module instead of using the classical distance method to calculate the similarity of each of the images of the database to the example images for a given topic uses our own logistic regression relevance algorithm to get the probability of an image belonging to the query set. The sub-models regressions are set to five features inside each features family that are the number of example images given for each topic (see more details of the regression method at section 2.1.). The relevant images are the example images, and the non-relevant images are randomly taken from outside the pre-textual filtered list.

3.1.3 The fusion sub-system

The **fusion subsystem** is in charge of merging the two score result lists from the TBIR and the CBIR subsystem. In the present work we use the product fusion algorithm (Si*St). The two results lists are fused together to combine the relevance scores of both textual and visually retrieved images (St and Si). Both subsystems will have the same importance for the resulting list: the final relevance of the images will be calculated using the product.

4 Experiments and results

4.1 Modality classification experiments

In this our first participation on the medical modality classification subtask, we have only participated with visual modality runs. Our objective for this edition has been to test the behavior of our logistic regression model for the classification task, and to adjust the parameters for the regression model explained at the section 2. The parameters to be defined to model each of the categories are:

- *The automatic election for the relevant and non-relevant images for the model to train each of the categories (positive and negative images).*
  The organization gives a training set for each of the categories being these training sets the relevant images for our logistic regression model. The number of images for each category differs from 5 images at the lower range (DSEC and DSEM categories) to 49 images at the highest range (COMP, DRCT and GGEL categories). The non-relevant images are the nearest N image to the centroid images of the set of images of the other categories different to the one being trained. The number of non-relevant images will be the double of the number of relevant images.
  We present two approaches for the number of relevant images to be used: for the first approach all available images for each given category are taken as relevant images (runs 1, 3 and 4), and for the second approach we limit the number of relevant images to a MAX number of images. The MAX number chosen is 30 because is the average low-level features components for each visual information family (run 2).
- *The different subgroups to adjust smaller regression models.*
  As it has been explained above the number of positive plus negative images, k (5 + 5*2 for the minimum set of training image category is smaller than the number of characteristics p (292 low-level featured vector) (k < p). We present four different approaches to group the low-level features: a regression model for each family low-level vector (run1), a regression model for each 30 components (run2) being 30 images the number of relevant images, a regression model for the lowest number of relevant images given that for this collection is 5 images (run3), and an adaptive regression model strategy different for each category depending on the minimum number of given relevant images or to the minimum number of components for the low-level featured family vector (run4). For all runs, the different submodels are merged by the average function.

Table 1. – Detailed information and results of the submitted visual modality class. runs.

		Regression parameters						Results
				Color features		Texture features		
Run	Description	# relevant images	Global color [30 comp.]	Local color 4 patches of [48 comp.]	Granulo-metric line horizontal [31 comp.]	Granulo-metric line vertical [31 comp.]	Ssdl [9 comp.]	Correctly classi-fied (%)
RUN1	A model for each family vector.	All	[30]	4*[48]	[31]	[31]	[9]	11,9
RUN2	A model each 30 components.	30	[30]	6*[30]+[12]	[31]	[31]	[9]	13,1
RUN3	A model each 5 components.	All	6*[5]	38*[5]+[2]	5*[5]+[6]	5*[5]+[6]	[5]+[4]	13,4
RUN4	Adaptative to the minimum of the number of relevant images or components vector family.	All	[30]	4*[48]	[31]	[31]	[9]	15,7

Table 1 shows the detail information of the submitted runs and the results obtained by means of the percentage of correctly image classified. Our results for the test set at classification task are much lower than the results we get at the training set. We must study the query-by-query results to determine how to improve the performance of the regression model as a classifier. Analyzing the four different tuning parameters for the regression method (see Table 1), the one that better performs is the adaptive model to the minimum of the number of relevant images or the number of components of the vector family features.

4.2 Ad-hoc image-based retrieval experiments

Table 2 shows the submitted runs for the ad-hoc image-based medical 2012 edition. The first run is the textual baseline run that is used as the pre-filtered textual list for the following multimodal experiments (run 2 to 9). For the textual baseline, run UNED_UV_01, besides the general preprocess presented before, the text of each query is reformatted in order to remove domain stopwords (i.e. meaningless terms in the medical domain like *images*)

Original query:	*thyroid CT images*
Reformatted query:	*thyroid CT*

Table 2. – Detailed information of the Submitted runs at the retrieval task.

		TBIR	CBIR
Run	Modality	Method	Features Vector
UNED_UV_01_TXT_EN	Textual	Remove domain stopwords	
UNED_UV_02_IMG_LOW_FEATURES	Visual	-	[LF]
UNED_UV_03_TXTIMG_LOW_FEATURES	Mixed	-	[LF]
UNED_UV_04_IMG_LOW_FEAT_2VECT	Visual	-	[LF]*[CF]
UNED_UV_05_IMG_EXPAND_FEAT_1VEC	Visual	-	[LF … CF]
UNED_UV_06_TXTIMG_EXPAND_FEAT_2VECT	Mixed	-	[LF]*[CF]
UNED_UV_07_TXTIMG_EXPAND_FEAT_1VECT	Mixed	-	[LF … CF]
UNED_UV_08_IMG_CONCEPT_FEAT	Visual	-	[CF]
UNED_UV_09_TXTIMG_CONCEPT_FEAT	Mixed	-	[CF]

The multimodal experiments (runs 2 to 9) have been designed to test the behavior of the expanded concept features vector. The runs marked as visual at the modality column at Table 2 use only the visual score, Si, to re-rank the final list. Meanwhile, those marked as Mixed use both textual and visual score to re-rank the final list by the product, St*Si. The third column shows which features vector has been used by the CBIR system to obtain the visual score, Si, with the following codes meaning: [LF], uses only the low-level features vector $F'(I_i) = (x_1, \dots x_k\} \in R^k$;[CF], using only the concept/category features for visual information $F''(I_i) = (x_1, \dots x_m\} \in R^m$; [LF…CF], uses the extended concept vector $F(I_i) = (x_1, \dots x_k, c_1, \dots, c_m\} \in R^{k+m}$ as a unique vector; and finally, [LF]*[CF], uses the extended concept vector as two different vectors obtaining two probabilities, $S_x\,(I_i)$ for the low-level features vector, and $S_c\,(I_i)$ for the concept vector that are merged by the product $S(I_i\ ) = S_x\,(I_i) * S_c\,(I_i)$.

Table 3. Results for the submitted experiments at the ad-hoc image-base retrieval subtask.

Run	Modality	MAP	bpref	P@10	P@30
UNED_UV_01_TXT_EN	Textual	0.0039	0.0055	0.0091	0.0076
UNED_UV_02_IMG_LOW_FEATURES	Visual	0.0034	0.0114	0.0455	0.0273
UNED_UV_03_TXTIMG_LOW_FEATURES	Mixed	0.0015	0.0037	0.0045	0.0061
UNED_UV_04_IMG_LOW_FEAT_2VECT	Visual	0.0400	0.0104	0.0409	0.0258
UNED_UV_05_IMG_EXPAND_FEAT_1VEC	Visual	0.0036	0.0111	0.0455	0.0303
UNED_UV_06_TXTIMG_EXPAND_FEAT_2VECT	Mixed	0.0013	0.0034	0.0091	0.0045
UNED_UV_07_TXTIMG_EXPAND_FEAT_1VECT	Mixed	0.0015	0.0036	0.0045	0.0061
UNED_UV_08_IMG_CONCEPT_FEAT	Visual	0.0033	0.0104	0.0227	0.0197
UNED_UV_09_TXTIMG_CONCEPT_FEAT	Mixed	0.0021	0.0050	0.0091	0.0061

Table 3 shows the results obtained at the ad-hoc image-based retrieval subtask by means of the MAP (Mean Average Precision), bpref (binary preference) and the precisions at the first 10 and 30 image retrieved (P@10 and P@30 respectively). The textual baseline has very poor results with a MAP of 0.0039. As the multimodal approach relies on the textual pre-filtered list, the multimodal runs do not outperform the textual baseline as we have already tested in other collections [11]. This low MAP is due to a low recall value that means that an important set of the relevant images are not selected by the textual system and then are not processed by the visual system. The visual approaches that re-rank the final score list using only the visual score Si, marked as visual at table 3, get better results by means of MAP and precision @10 than runs using both textual and visual scores St*Si, runs marked as mixed at table 3. This behavior is opposite as other results in which the multimodal approaches outperform the textual baseline [11] due to the performance of the textual system.

Analyzing the multimodal experiments, we can observe that runs using the expanded conceptual vector, runs UNED_UV_04 and UNED_UV_05 obtain better results by means of MAP (0.0040 and 0.0036 respectively) than runs that only use the low-level features vector, run UNED_UV_02 (0.0034). These results confirm our idea that the expanded concept vector adds information about the categorization of the image for the retrieval process. About using a unique vector for the expanded concept features vector or two vectors, we can not extract a definitive conclusion so that the results obtains by the two runs are very close, and can be also mask by the noise introduced by the textual prefiltered approach.

5 Remarks and Future Work

The textual retrieval approach we have proposed this time, based on a query reformatted process, which focuses on the semantic of the queries by try to use only medical terms, has not obtained the expected results. We will analyze this bad performance of the textual retrieval process at the Medical collection, given that this technique was successfully tested last year at the Wikipedia collection [10] (2nd in textual category).

For the multimodal approaches presented for the ad-hoc image-based retrieval subtask, our combination of the textual pre-filtered list as input to the visual system does not outperform the textual baseline, as it has already been tested in other ImageClef collections, Wikipedia [3,10] due to the fact of the performance of the textual approaches. Focusing on the visual system, the expanded concept vector presented outperforms the use of the low-level features vector in the Medical collection as in the Flickr photo subtask [5].

The results obtained at the classification modality subtask suffered from the fact that our visual approach is a retrieval approach adapted for the classification modality task. Nevertheless, the regression model system proposed as a modality classifier will be analyzed query-by-query to improve its classification performance.

Acknowledgments. This work has been partially supported for Regional Government of Madrid under Research Network MA2VIRMR (S2009/TIC-1542), for Spanish Government by project BUSCAMEDIA (CEN-20091026) and by project MCYT TEC2009-12980.

6 References

1. Alpokocak, A., Ozturkmenoglu, O., Berber, T., Vahid, A.H., Hamed, R.G.: DEMIR at ImageCLEFmed 2011: Evaluation of fusion techniques for multimodal content-based medical image retrieval. In *CLEF 2011 Working Notes*. 2011
2. Ayala, G.; Domingo, J. Spatial Size Distributions. Applications to Shape and Texture Analysis. IEEE Transactions on Pattern Analysis and Machine Intelligence. 2001. Vol. 23, N. 12, pages 1430-1442.
3. Benavent, J. Benavent, X. de Ves, E. Granados, R. Garcia-Serrano, A.: Experimentes at ImageCLEF 2010 using CBIR and TBIR Mixing Information Approaches. In M. Braschler, D. Harman, E. Pianta, *CLEF 2010 LABs and Workshops, Notebook Papers.* Padoua, Italy. 2010.
4. Benavent, J., Benavent, X., de Ves, E. Recuperación de Información visual utilizando descriptores conceptuales. In Conference Proceedings of the Conferencia Española de Recuperación de Información, CERI 2012, Valencia, 2012.
5. J. Benavent , A. Castellanos, X. Benavent, E. De Ves, Ana García-Serrano. Visual Concept Features and Textual Expansion in a Multimodal System for concept annotation and retrieval with Flickr photos at ImageCLEF2012. *In CLEF 2012 Working Notes, 2012.*
6. Castellanos, A. Benavent, X. Benavent, J. Garcia-Serrano, A.: UNED-UV at Medical Retrieval Task of ImageCLEF 2011. In *CLEF 2011 Working Notes*. 2011.

7. Castellanos, A., Benavent, X., García-Serrano, A., Cigarrán, J.: Multimedia Retrieval in a Medical Image Collection: Results Using Modality Classes. In *Workshop of Medical Content-based Retrieval for Clinical Decision Support (MCBR-CDS 2012)*. To be published. 2012.
8. Csurka, G., Clinchant, S., Jacquet, G.: XRCE's participation at medical image modality classification and ad-hoc retrieval task of ImageCLEFmed 2011. In *CLEF 2011 Working Notes*. 2011.
9. Depeursinge, A. Müller, H.: Fusion techniques for combining textual and visual information retrieval. In: *ImageCLEF, The springer international series on information retrieval, vol. 32*. Springer, Berlin Heidelberg, pages 95–114. 2010.
10. Granados, R. Benavent, J. Benavent, X. de Ves, E. Garcia-Serrano, A.: Multimodal Information Approaches for the Wikipedia Collection at ImageCLEF 2011. In *CLEF 2011 Working Notes*. 2011.
11. Henning Müller, Alba Garcia Seco de Herrera, Jayashree Kalpathy-Cramer, Dina Demner Fushman, Sameer Antani, Ivan Eggel, Overview of the ImageCLEF 2012 medical image retrieval and classification tasks, CLEF 2012 working notes, Rome, Italy, 2012.
12. Leon T., Zuccarello P., Ayala G., de Ves E., Domingo J.: Applying logistic regression to relevance feedback in image retrieval systems, *Pattern Recognition, V40*, p.p. 2621, 2007.
13. Kalpathy-Cramer, J. Müller,H. Bedrick, S. Eggel, I. García-Seco de Herrera, A. Tsikrika, T.: Overview of the CLEF 2011 medical image classification and retrieval tasks. In *CLEF 2011 Working Notes*. 2011.
14. Torjmen, M. Pinel-Sauvagant, K. Boughanem, M.: Methods for Combining Content-Based and Textual-Based Approaches in Medical Image Retrieval. In *Evaluating Systems for Multilingual and Multimodal Information Access*. 2009.
15. Tsikrika, T. Popescu, A. Kludas, J.: Overview of the Wikipedia Image Retrieval Task at ImageCLEF 2011. In: *CLEF 2011 Working Notes*. 2011.