A multimedia IR-based system for the Photo Annotation Task at ImageCLEF2013

X. Benavent[2], A. Castellanos[1], E. de Ves[2], D. Hernández-Aranda[1], R. Granados[1], A. Garcia-Serrano[1]

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

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

Abstract. The UNED-UV group at the ImageCLEF2013 Campaign have participated in the Scalable Concept Image Annotation subtask. We present a multimedia IR-based system for the annotation task. In this collection, the images do not have any textual description associated, so we have downloaded and preprocessed the web pages which contain the images. Regarding the concepts, we expanded their textual description with additional information from external resources as Wikipedia or WordNet and we generate a KLD concept model using recovered textual information. The multimedia IR-based system uses a logistic relevance algorithm to get a model for each of the concepts to be trained using visual image features. Finally, the fusion subsystem merges textual and visual scores for a certain image to belong a concept, and decides the presence of the concept in the images.

Keywords: Text-Based Information Retrieval, Content-Based Information Retrieval, Multimedia Fusion, Logistic regression relevance algorithm.

1 Introduction

The UNED-UV is a research group with researchers from two universities in Spain, the Universidad Nacional de Educación a Distancia (UNED) and the Valencia University (UV). The group is working together since ImageCLEF08 edition.

At this 2013 ImageCLEF Campaign [2], we participate at the Photo Annotation and Retrieval Task [6] in the Scalable Concept Image Annotation subtask. The motivation for this edition is focused on the development of image annotation systems that address the scalability problem in such a way that the annotation systems have to be able to adapt their behavior to take into account new concepts that can appear in the images to be annotated. There were two datasets to evaluate the systems, one containing the same concepts used for training (95 concepts) and a second one containing these 95 concepts and 21 additional ones.

As the classification-based systems, traditionally used for image annotation, are not suitable for this task, we use a multimedia IR-based annotation methodology that produces a concept model that predicts the probability that a certain concept belongs to an image. A merging algorithm fuses textual and visual probabilities, and this final score is used to decide the presence of a certain concept in an image.

Section 2 describes the system overview and the annotation methodology used for the two approaches submitted using the textual and the multimodal information. After that, section 3 shows the submitted runs and section 4 analyze the results obtained. Finally, in section 5 we extract conclusions and outlines possible future research lines.

2 System Description

The global system (shown at Fig. 1) is divided into three main subsystems: TBIR (Text-Based Image Retrieval), CBIR (Content-Based Image Retrieval) and the Merging module. The TBIR subsystem is in charge of annotating the images using only textual information selected from the web pages where they were downloaded.

As the list of concepts does not include example images to train each one of the concepts, the TBIR subsystem is in charge of generating a training set of images for each concept for the Multimedia approaches. These images are taken from the so called 3k collection images ($Devel + Test$). The CBIR subsystem generates a model for each of the concepts with the generated training set images. These concept models are used to generate the lists of relevant images for each concept. Finally, these lists are combined with the fusion subsystem following a late fusion approach based on the OWA operator [7].

2.1 Annotation using textual information

TBIR subsystem is in charge of the textual annotation of images in the collection. In this collection, the images do not have any textual description associated, so the first step is to obtain the textual information for describing them. For this task, we have downloaded and pre-processed the web pages which contain the images. Regarding the concepts, we expanded their textual description with additional information from external resources as Wikipedia or WordNet.

As an image may be annotated by several concepts, the annotation strategy is based on an information retrieval approach, which indexes the concepts and uses each image as query. The result of the retrieval process is a ranked concepts list for each image, ordered by the textual similarity or score (S_t).

The modules of the TBIR subsystem have the following functionalities:

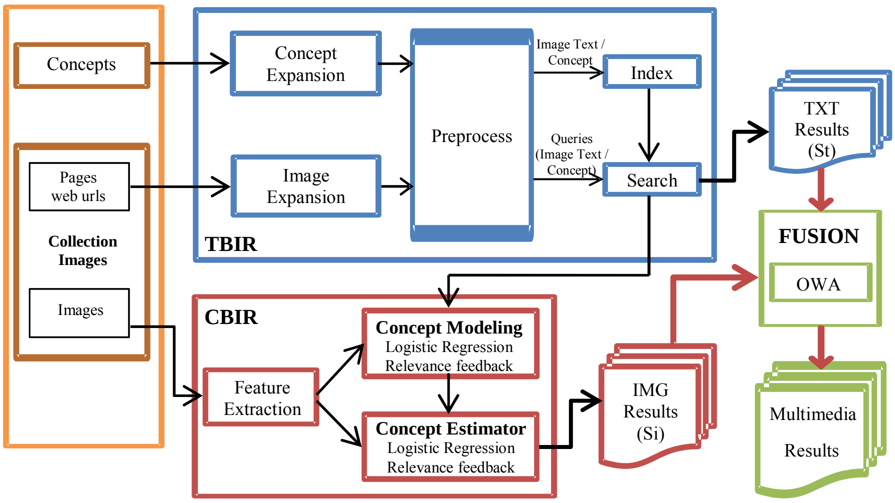

Fig. 1. System Overview

Image Expansion. This module is in charge of downloading the web pages that contain the images in the collection. Then, we extract the textual information directly related to each image taking into account the text contained in the following HTML attributes: "title" and "alt" of <img> tag; and the <a> tag if the image is within a link. An image may be contained in several web pages, so we recover the textual information of every web page. Moreover, we take into account the image name that appears in its URL.

Concept Expansion. In order to obtain additional information to describe the concepts, we use Wikipedia and WordNet as external resources, in the following way:

- Wikipedia. We extract the textual information from fields <text> and <categories> contained in the corresponding Wikipedia pages of the concepts.
- WordNet. Lexical and semantic information about the concepts is extracted: definition, synonyms, hypernyms, hyponyms and related concepts.

Additionally, we have modelled the raw text obtained from the Wikipedia concepts description to identify a list with the most representative terms (so-called the Wikipedia-KLD list). For this, we applied a divergence-based approach (Kullback Leibler Divergence or KLD [4]) to identify not only the representative terminology but also the terminology that better differentiate each concept from the rest. KLD weights each term according to their occurrence in a given content and their occurrence in the rest of the contents following the formulation in (1):

$$KLD_{pD,pC} = pD(t) \cdot \ln\left(\frac{pD(t)}{pC(t)}\right) \quad (1)$$

where $pD(t)$ is the probability of each term within a document D (frequency of t divided by the whole of terms in the document D) and $pC(t)$is the probability of the same term t within the collection C (frequency of t divided by the number of terms in the collection C).

Pre-processing. Textual information is pre-processed: 1) deletion of characters with no statistical meaning, like punctuation marks or blanks; 2) deletion of semantic empty words in English language (stopwords), 3) reduction of words to their base form by stemming, and 4) conversion of all words into lower case.

Indexing. The indexing process is carried out using Lucene. The images are indexed using only one field with the text associated to each image. The concepts are indexed using three fields, depending on the information used for expansion: Wikipedia, WordNET and Wikipedia_KLD.

Searching. This module is in charge of launching the queries against a concrete index in order to obtain the corresponding textual results (*Txt Results*). When using images as queries, a concepts list will be obtained; and when the queries are the concepts, an images list will be generated. The latter is used to fuse these textual results with the visual ones obtained from the CBIR subsystem. The applied ranking function is BM25 and its extension for structured documents BM25F [5], using the default parameters.

For the concept description several approaches were tested. Finally, we have represented (and indexed) each concept by:

- **Concept**: The name of the concept.
- **WP_Description**: Contains the raw text of the concept Wikipedia page, plus the Wikipedia categories of the concept.
- **WP_KLD_Description**: As we have modelled each concept using the raw Wikipedia text and the Wikipedia categories (WP_Desccription), the 50 most representative terms, according KLD weighting, are indexed for each concept.
- **WN_Description**: The textual information obtained for the concept at WordNet is the one included at the <definition>, <forms>, <hypernyms>,<hyponyms> and <related> WordNet components.

The different textual information describing the images is (store in the five fields):

- **Img:** The textual description initially associated to the image: *img_title, img_alt, img_link e img_name.*
- **Webpage:** Includes the general description about the webpage containing the image: *webpage_title, webpage_description y webpage_keywords.*
- **img+webpage:** The two previous fields together.
- **text**: The whole webpage text (*text* element) containing the image.
- **img+webpage+text:** The three previous fields together.

Several experiments were performed in order to compare the use of the previous alternatives for the textual description of the images when using as queries to retrieve concepts from the concepts index of the *Devel* collection. The best result was obtained using only the *img* field as a query; so this is the field to be used in the runs.

2.2 Annotation using Multimedia information

For the Multimedia approaches, the TBIR subsystem generates a training set for each concept to be annotated. These images are taken from the 3k collection images (Devel + Test). The CBIR subsystem generates a model or predictor for each of the concepts with the Logistic Regression Relevance Model algorithm [3]. Once, the concepts models are trained, these models are used to predict the probability that a given image belongs to a certain concept (Si). Both probabilities (St,Si) are combined by the Fusion subsystem that finally decides if a certain concept is present or not to be annotated.

For the Logistic Regression Relevance algorithm, each of the concept models needs two sets to be trained: a set of images that have the concept, being $I_s^P$the relevant or positive images, and a set of images that not belongs to a concept, being I_s^N, the set of non-relevant images or negative images. Each image is represented by a K-dimensional low-level features vector $\{x_1, .., x_i, .., x_k\}$. The relevance probability for a certain concept 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 $logit\,[\pi(x)] = \alpha + \beta_1 x_1 + \dots + \beta_k x_k$, 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.

The positive o relevant images ($\mathrm{I_s^P}$ set), is given by the Text-Based Information sub-system using the 3k collection images. This initial list is tailored up to the tenth top images, and the final selection is human supervised. The non-relevant images, the $\mathrm{I_s^N}$ set, is selected from the images that do not have the required concept and this list is also tailored up to the twentieth top images, being the selection also human supervised. A good selection of the images that represent a certain concept is very important to make the estimator good and robust. For this reason, we have considered important the human supervision for the training sets. Furthermore, these sets are generated only once for training, and could be used to annotate any other collection with these concepts.

The explanatory variables $\mathrm{x} = (\mathrm{x}_{1,} \dots, \mathrm{x_k})$ to train the model are the visual low-level features based on colour and texture information that are given by the organization [6]: colour histograms and GIFT shape descriptor that describes the shape in an image by calculating the Gabor transform. We have a low-level features vector of 544 components: 64 for the colour histograms and 480 for the GIFT descriptor.

2.3 Multimedia Fusion

To merge the textual and the visual information, we have followed a late fusion approach which combines the two monomodal results lists at decision level. The applied late fusion algorithm tested in previous works [1] is based on the Mathematical aggregation operator OWA [7]. The OWA transforms a finite number of inputs into a single output without associating weights to any particular input; instead, the relative magnitude of the inputs decides which weight corresponds to each input. In our application, the inputs are the textual and image scores (S_t and S_i), and this property is very interesting because we do not know, a priori, which subsystem will provide us the best information. The aggregation weights used for our experiments correspond to an $orness = 0.3$, which means that a weight of 0.3 is given to the higher probability value and a weight of 0.7 to the lower one.

Once the final fused list is obtained (containing, for each image, a ranked list of associated concepts), we have to decide how many concepts will be used to annotate each image. We consider two options: 1) select a fixed number of concepts; and 2) calculate a relevance threshold that decides whether or not an image is annotated by a concept. Both options have been evaluated with the *Devel* collection, since it has ground truth.

Relative to the first option, Fig. 2 shows the evolution of evaluation measures MAP, mFsamp (by image) and mFcnpt (by concept) depending on the number of annotated concepts (values between 1 and 20). We can see how the MAP value is higher as the number of concepts is increased, while the value of the rest of measures decreases, being the cut-point between $N = 5$ and $N = 6$. On the other hand, the mean number of concept annotation per image is 6.345 (calculated from *Devel* ground truth), so, taking into account both factors, we decide to select $N = 7$ as fixed number of concepts to annotate images.

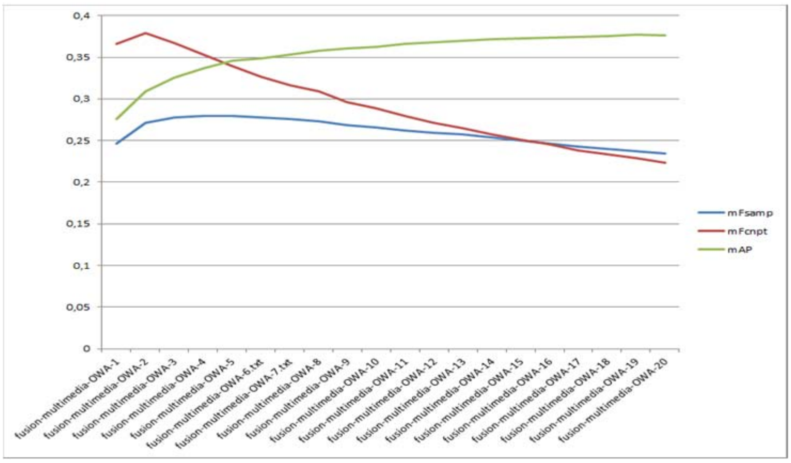

Fig. 2. Evolution of the evaluation measures depending on number of concepts

For the second option, the threshold calculation is based on the percentage of the maximum score by image. This option is more flexible, since an image is not annotated with a fixed number of concepts. Fig. 3 shows the evolution of the evaluation measures considering the percentages from 10% to 100%. We can see that the more restrictive is the threshold (low values of percentage), the more MAP value increases and the rest of measures decrease.

The cut-off point is between 70 and 80%. If the number of concepts with which an image is annotated is considered, for 70% the mean is 6.622 while for 80% is 3.356.

Therefore, we have selected the percentage of 70% as a threshold, since its mean concept number is similar to the mean calculated from *Devel* ground truth (6.345).

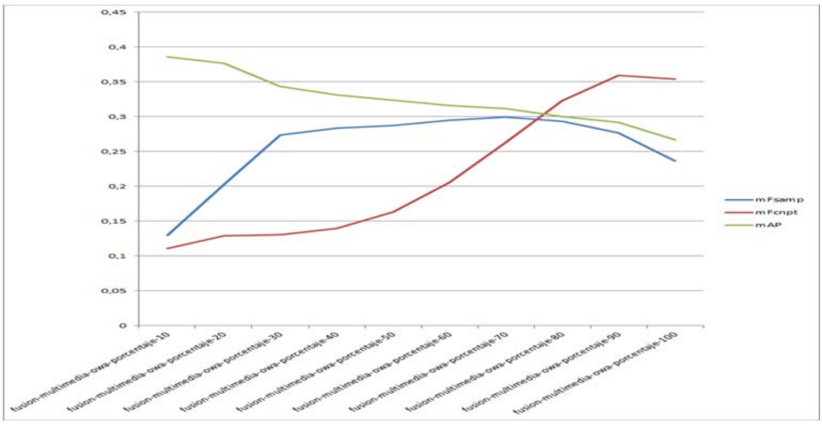

Fig. 3. Evolution of the evaluation measures depending on relevance threshold.

3 Experiments

In this section, we present the two approaches developed for the image annotation subtask: the monomodal approach (TBIR-based annotation) and the multimodal approach (TBIR and CBIR based annotation). The five runs submitted are:

- UNEDUV_1: Monomodal Approach. Query with the img field against the KLD_WP concept representation indexed.
- UNEDUV_2: Monomodal approach. Query with the img field against the KLD_WP+WN concept representation indexed.
- UNEDUV_3: Monomodal approach. Query with the img field against the WN concept representation indexed.
- UNEDUV_4: Multimodal approach. UNEDUV_2 textual run is merged with the visual run by OWA algorithm, using the seventh most representative concepts to annotate every image.
- UNEDUV_5: Multimodal approach. UNEDUV_2 textual run is merged with the visual run by OWA algorithm, using the concepts with 70% of representative concept score.

4 UNED-UV Results

In this section, we describe the results obtained with the submitted runs using the development and the test collections, measured according the mean F-measure for both the samples (MF-samples) and the concepts (MF-concepts); and the mean average precision for the samples (MAP-samples). The test set results is also evaluated with a fourth measure (MF-unseen), the mean F-measure for the concepts that are not in the development set. The values between the square brackets correspond to the 95% confidence intervals computed using Wilson's method. Table 1 shows the results for the development set, divided into textual and multimodal runs, and Table 2 shows the same results for the test set. The best obtained value for each of the measures it is also included in the table, together with the average over all the presented experiments by the rest of participants.

All our submitted runs are beyond the baseline results for both the development set and for the test set according to all measures (see the overall results at the ImageCLEF webpage [6]). Looking into the overall participant's results list, our best runs are at positions 21, 16 and 27 ordered by the MF-Samples, MF-Concepts and MAP-samples respectively for the development set, and at positions 24, 11 and 26 for the test set. It means that our best runs are at the first third top results.

Focusing on textual runs, UNEDUV_1 and 2 offer similar results according to all the measures; however, UNEDUV_3 results, which are based only in WordNET annotation, differs: values based on sample results are fewer than values of the other two approaches, and the value based on concepts (MF-concepts) improves these results (4-5 points higher). This behavior is observed in the development and in the test set.

On the other hand, the multimodal-based approach, UNEDUV_4 and UNEDUV_5 runs, improve the textual based run, the UNEDUV_2, in almost all measures, being this improvement higher in the UNEDUV_5. This means that the merging strategy of using a relevance threshold to decide if a concept should be or not annotated performs better than the one that uses a static number of concepts to be annotated. Similar behavior is observed for the two sets, development and test.

Table 1. Development Set Results

Run	Mode	MFsamples	MFconcepts	MAPsamples
UNEDUV_1	Textual	25.0 [23.8-26.4]	27.5 [24.5-32.2]	32.8 [31.4-34.4]
UNEDUV_2	Textual	24.4 [23.2-25.7]	26.1 [23.6-30.4]	32.4 [31.0-33.9]
UNEDUV_3	Textual	22.5 [21.2-23.9]	31.5 [27.9-36.5]	27.1 [25.9-28.5]
UNEDUV_4	Multimodal	**29.9** [28.7-31.3]	26.3 [23.7-30.7]	31.0 [29.8-32.3]
UNEDUV_5	Multimodal	27.6 [26.8-28.6]	**31.7** [28.3-36.5]	**35.5** [34.1-36.9]
Best	-	51.3	45.0	50.4
Average	-	27.1	24.7	34.3

An important issue to highlight for the test set results at Table 2 is the MF-unseen values. In general all our submitted approaches offer satisfactory values; especially the WordNET-based approach (UNEDUV_3) value, that is the 3rd best overall value, and the Multimedia approach (UNEDUV_5) at the 5th position. It highlights a good generalization capacity for our systems to annotate unseen concepts.

Table 2. Test set Results

Run	Mode	MF-samples	MF-concepts	MF-unseen	MAP-samples
UNEDUV_1	Textual	23.0 [22.1-23.9]	25.0 [22.8-28.7]	31.7 [25.0-44.7]	30.3 [29.3-31.4]
UNEDUV_2	Textual	22.9 [22.0-23.8]	24.0 [22.3-27.5]	30.6 [24.7-43.1]	30.6 [29.6-31.7]
UNEDUV_3	Textual	23.1 [22.0-24.2]	**31.3** [28.1-35.8]	**43.2** [33.1-55.7]	26.6 [25.6-27.7]
UNEDUV_4	Multi-modal	**30.0** [29.0-31.1]	22.8 [20.9-26.5]	24.6 [19.5-38.5]	29.8 [28.9-30.9]
UNEDUV_5	Multi-modal	24.4 [23.8-25.1]	29.2 [26.7-33.1]	35.4 [27.7-48.2]	**33.2** [32.2-34.3]
Best	-	42.6	34.1	45.3	45.6
Average	-	23.7	21.7	22.1	30.69

5 Concluding Remarks

Our best runs are at the first third top results for the different measurements. This means that the multimedia IR-based system presented has obtained quite good results regarding the current state of the art.

For the textual approaches, the WordNET-based approach for concept expansion is the one that has a better performance. As this textual baseline was not the one used for the submitted multimodal approaches, it is going to be tested in the multimodal IR-based system.

The multimedia approaches slightly outperform its textual baseline, although not in all measurements, being this behaviour needed to be further analysed. The fusion subsystem has proved that a relevance threshold to decide the annotation of a concept achieves better results than selecting a fixed number of concepts per image. It is important to highlight the good generalization capacity of our system to annotate unseen concepts.

Acknowledgments. This work has been partially supported for Regional Government of Madrid under Research Network MA2VIRMR (S2009/TIC-1542), HOLOPEDIA (TIN 2010-21128-C02) and by project MCYT TEC2009-12980.

References

1. Benavent, X., García-Serrano, A. Granados, R., Benavent, J. de Ves, E. Multimedia Information Retrieval based on Late Semantic Fusion Approaches: Experiments on a Wikipedia Image Collection. IEEE Transactions on Multimedia. DOI: 10.1109/TMM.2013.2267726.
2. Caputo, B. Muller, H. Thomee, B. Villegas, M. Paredes, R. Zellhofer, D. Goeau, H. Joly, A. Bonnet, P. Martinez Gomez, J. Garcia Varea, I. Cazorla, M. ImageCLEF 2013: the vision, the data and the open challenges. Proc. CLEF 2013, LNCS.

3. Leon T., Zuccarello P., Ayala G., de Ves E., Domingo J.: Applying logistic regression to relevance feedback in image retrieval systems, Pattern Recognition (40), pp. 2621, 2007.
4. S. Kullback R. A. Leibler. On information and sufficiency. Annals of Mathematical Statistics, 22(1). 1951.
5. Robertson, S. E. S. Walker. Some simple effective approximations to the 2-Poisson model for probabilistic weighted retrieval. In Proceedings of the SIGIR '94, W. Bruce Croft and C. J. van Rijsbergen (Eds.). Springer-Verlag. NY, USA, 232-241. 1994.
6. Villegas, M. Paredes, R. Thomee, B. Overview of the ImageCLEF 2013 Scalable Concept Image Annotation Subtask. CLEF 2013 working notes, Valencia, Spain, 2013.
7. Yager, R. On ordered weighted averaging aggregation operators in multi criteria decision making. IEEE Transactions Systems Man and Cybernetics (18), pp. 183-190. 1988.