URJCyUNED at ImageCLEF 2012 Photo Annotation task*

Jesús Sánchez-Oro[1], Soto Montalvo[1], Antonio S. Montemayor[1], Raúl Cabido[1], Juan J. Pantrigo[1], Abraham Duarte[1], Víctor Fresno[2], and Raquel Martínez[2]

[1] Universidad Rey Juan Carlos, Móstoles, Spain
{jesus.sanchezoro,soto.montalvo,antonio.sanz}@urjc.es
{raul.cabido,juanjose.pantrigo,abraham.duarte}@urjc.es
[2] Universidad Nacional de Educación a Distancia, Madrid, Spain
{vfresno,raquel}@lsi.uned.es

Abstract. This paper describes the URJCyUNED participation in the ImageCLEF 2012 Photo Annotation task. The proposed approach uses both visual image features and textual associated image information. The visual features are extracted after preprocessing the images, and the textual information are the provided Flickr user tags. The visual features describe the images in terms of color and interesting points, and the textual features make use of the semantic distance between the user tags and the concepts to annotate by using WordNet. The annotations are predicted by SVM classifiers, in some cases trained separately for each concept. The experimental results show that the best of our submissions is obtained by using only textual features.

Keywords: Image classification, Visual features, Textual features, Semantic similarity, Bag of Words

1 Introduction

In this paper we describe our submission to the ImageCLEF 2012 Photo Annotation task. The main aim of this task is the analysis of a set of images in order to detect one or more visual concepts. Those concepts can be used for automatically annotate the images. This year a total of 94 concepts are available, categorized as *natural elements, environment, people, image elements*, and *human elements*, each one subdivided in different sub-categories.

The goal of this task is to annotate the images with concepts detected by using both visual and textual features extracted from the images. Participants are given 15000 training images and 10000 test images. The problem can be solved using three different methods: using visual information only, using textual

* This work has been part-funded by the Education Council of the Regional Government of Madrid, MA2VICMR (S-2009/TIC-1542), and the research projects Holopedia and SAMOA3D, funded by the Ministerio de Ciencia e Innovación under grants TIN2010-21128-C02 and TIN 2011-28151 respectively.

information only, and a hybrid approach that involves the fusion of visual and textual information.

This year a set of visual and textual features are provided by the ImageCLEF organization. On the one hand, the visual features include SIFT, C-SIFT, RGB-SIFT, OPPONENT-SIFT, SURF, TOP-SURF and GIST. On the other hand the textual features contain Flickr user tags, EXIF metadata and user information and Creative Commons license information. More details of the task and the features provided can be found in [11].

However this work only uses the Flickr user tags as textual information and the visual features are extracted by our own methods. Specifically, we extract visual features and textual features to create a global descriptor for the images. Visual features are mostly based on the color and interesting points of the images, while textual features use a similarity measure to compare the Flickr user tags with the concepts and their synonyms in the WordNet lexical database.

Analyzing last year works, most used visual features involved Bag Of Words as well as SIFT and color features [1, 12, 10]. The textual features were mostly based on similarity metrics between concepts and Flickr user tags and tags enrichment [7, 9, 14]. In our approach a preprocessing of the image is added to the visual feature extraction. This preprocessing is based on the change of the resolution and the removal of a percentage of the external part of the image.

The rest of the paper is organized as follows. Section 2 describes the visual and textual features proposed in this work. Section 3 presents the results of the submissions. Finally, Section 4 draws the conclusions of the work.

2 Features

Our approach uses visual image features and Flickr user tags as textual associated information. We did not use other textual features as EXIF metadata or user and license information. Sections 2.1 and 2.2 present, respectively, the visual and textual features considered.

2.1 Visual Features

The visual features proposed in this paper describe the images in terms of color and interesting points. The descriptor is created by joining all the features into one feature vector that is later used as input to the classifier.

Color quantization The color histogram for an image is a representation of the color distribution of the image. In this work we use the histogram of the RGB color space, in which each component is represented by three values, namely R (red), G (green) and B (blue). The RGB color histogram is three-dimensional, and each dimension is divided in N bins, being N the only parameter of the histogram. Relying in a comparison of several values for N, we have selected a value of 32. This means that each dimension of the histogram is divided into 32 discrete intervals. Once the histogram has been generated, the feature extracted

consists of the two most repeated bins in the image, setting up a vector of six components (R1,G1,B1,R2,G2,B2). We select the most repeated colors under the assumption that those are the most relevant colors for the images. Figure 1 shows an example of the quantization of an image of the training set into 32 bins.

Fig. 1. Color quantization of an image in 32 bins

Horizontal and vertical edges The edges of an image define the shape of the figures depicted in it. High frequency signals carry the most part of the information, so it seems reasonable to have a measure of edges. For that reason, edges have been commonly used in image classification in different ways. This work proposes the use of edges defining an image as the percentage of pixels belonging to horizontal and vertical edges respectively. The edges are extracted using the very common Canny edge detector [2]. With this information, the feature used is a two component vector that contains vertical and horizontal edge pixels percentages. More on edge analysis will be included with the use of Histogram of Oriented Gradients for a number of objects.

Grey color percentage Some of the concepts proposed this year are usually suggested by photographs that are partial or totally black and white pictures (melancholic, gray color, scary). Consequently, we propose an additional feature based on the percentage of gray pixels in the image. The selected pixels are not only those which are purely gray (i.e., an RGB code where R=G=B), but also those which can be considered gray within a threshold.

Face detection An adult human brain contains highly specialized neurons for the recognition of human faces. Moreover, some of the concepts contain some words that are related to people (family-friends, quantity, coworkers, ...). It would be interesting to have a face detector which can difference between pictures with and without faces, and, therefore, with and without people. The detector uses the Viola-Jones method [13] and later improved by Lienhart and Maydt [6] to store Haar features obtained from an image, using the integral image.

The method uses AdaBoost to combine several weak classifiers, resulting in a strong classifier. The output of the algorithm is the location of each face and its bounding box, but this work uses only the number of faces in the image. Figure 2 shows the face detection over a training image.

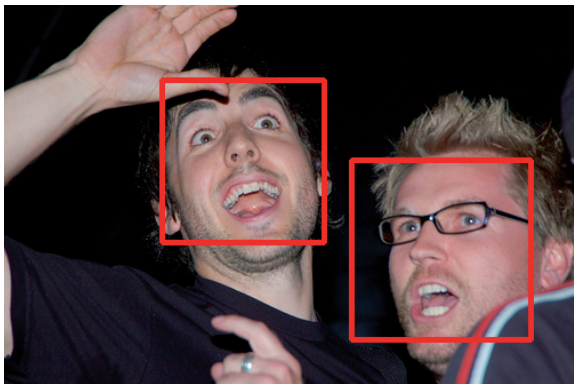

Fig. 2. Example of a face detection in a training image

Bag of Words Analyzing the previous ImageCLEF results [1] we can see that Bag Of Words (BOW) is one of the most extended features. The BOW method relies under the assumption that the spatial relationships of the key points in the image lacks of importance. Those key points can be obtained by using some well-known descriptors like SURF or SIFT, for instance. The features are vectors of real numbers, and the construction of a dictionary from all the features obtained from the training set directly can be unaffordable, both in time and in memory. The solution is based on the use of a limited number of feature vectors which represent the feature space well for constructing a dictionary, which is usually carried out with k-means clustering. Once the dictionary has been constructed, the new images can be described by extracting their features and matching them with the features in the dictionary which are closest [3].

The k-means algorithm used in this work is k-means++, which uses a heuristic for choosing good initial cluster centers, instead of the random centers chosen by the standard k-means. The encode of the images can be divided in three steps: feature detection, feature extraction and descriptor matching. The feature detection step identifies the keypoints, which are later extracted in a preset format in the feature extraction stage. Finally, the descriptor matching step matches the features extracted to features in the dictionary to construct the BOW representation of the image. We use the Good Features To Track (GFTT) detector implemented in OpenCV as feature extractor. It uses the Shi-Tomasi corner detector to detect keypoints in the images. Figure 3 shows an example of the detection of keypoints using GFTT. The feature descriptor uses SURF as descriptor

format and the descriptor matcher is FLANN based. FLANN (Fast Library for Approximate Nearest Neighbors) is a library that contains a collection of algorithms optimized for fast neighbor search. The keypoints are clustered into 50 groups, resulting in a vector of 50 elements per image.

Fig. 3. Interesting points detected using Good Features To Track

Histogram of Oriented Gradients (HoG) The Histogram of Oriented Gradients (HoG) relies on the idea that local object appearance and shape can often be characterized rather well by the distribution of local intensity gradients or edge directions. It is implemented by dividing the image into small cells. A local 1-D histogram of gradient directions or edge orientations is accumulated for each cell. The representation of the image is form by combining the histogram entries. The representation is normalized for invariance to illumination and shadowing by accumulating a local histogram of energy over larger spatial regions, called blocks [4].

The HoG templates available covers 12 of the 94 concepts of the task. The feature obtained with this method is a boolean value that indicates whether the concept appears in the image or not. Figure 4 shows the template for a bicycle and an example of the detection in one of the training images.

Resolution We propose the extraction of the previous features for different image resolutions. Reducing the spatial resolution of an image helps to reduce some unimportant details as well as noise. On the one hand, the histogram, the edges, the gray color, and the face detection work best at half resolution. On the other hand, BOW and HoG are better at the original resolution of the image. We take a trade off after experimentation.

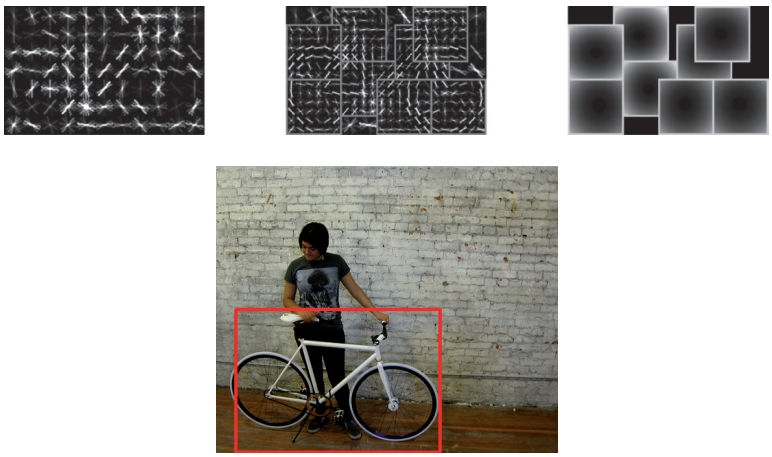

Fig. 4. HoG bicycle template and detection

Frame elimination The main elements of a composition are usually located in the center of it. For that reason it is interesting to delete the external part of the image as a way of deleting secondary elements that can interfere in the detection of the main elements. Neither BOW nor HoG improve their results with this technique, but the other features have better results by deleting the 15% of the image frame (both vertical and horizontal). Figure 5 shows an example of the frame elimination in a training image. It is easy to see that the elimination of the frame allows the feature extractor to focus on the interesting part of the image instead of looking for keypoints in non-interesting areas of the image.

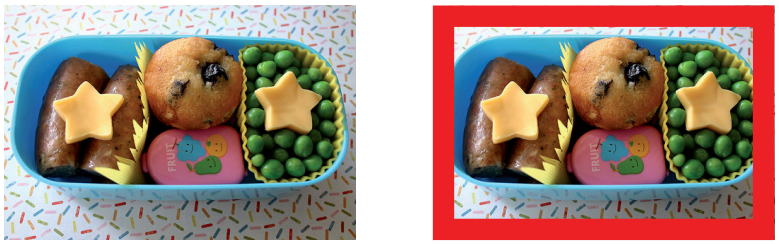

Fig. 5. Original training image (a) and frame elimination (b)

2.2 Textual Features

Our approach consists of building a text representation from Flickr user tags and using the lexical database WordNet [8]. We think these text features would allow

capturing part of the meaning of the images, so that they could provide valuable information for the concept annotation process that is not easy to extract from the image features.

WordNet is used to enrich the concepts to be tagged with synonyms. After this enrichment procedure the number of total concepts is 310. This way each image is represented by means of a vector of those 310 components that represents the semantic distance between the image tags and the concepts. Then, the vector components are the minimal semantic distances between the word tags of each image and the concept or synonym of the vector representation.

Since the Flickr user tags are written in different languages, it is necessary translate them into the same language. As WordNet is in English, this is the selected axis language. For each image and each word tag we identify whether the word is written in English or not; and if it is not, it is translated into English by means of bilingual dictionaries. The word tags that do not have translation are discarded. After the elimination of the stop-words, we calculate the semantic distance matrix between the final word tags of each image and the 310 concepts using WordNet and the Leacock-Chodorow semantic measure [5]. The Leacock-Chodorow measure is based on path length, where the similarity between two concepts c_1 and c_2 is as follows:

$$sim_{LCH}(c_1, c_2) = -\log \frac{length(c_1, c_2)}{2 \times \max\limits_{c \in WordNet} depth(c)} \quad (1)$$

That is, the number of nodes along the shortest path between them, divided by two times the maximum depth of the hierarchy (from the lowest node to the top in the taxonomy in which c_1 and c_2 occur).

Several works in the ImageCLEF 2011 also took into account text features and semantic distances [14, 7], using different semantic measures and different semantic resources.

3 Experimental Evaluation

The experimental computation has been carried out in an Intel Core i7-2600 3.40 GHz with 3 GB RAM. We have submitted a total of four runs: one visual, one textual and two mixed runs. Figure 6 shows the steps followed in the classification.

3.1 Submitted runs

– **Visual run:** Independent SVM classifiers are used for each concept. This give us the opportunity of using only the best combination of features for each concept, instead of using all the features for all concepts, which could eventually end in worse results. The features are divided in three main groups. The first group contains the color features (histogram and gray color percentage) and the face detector, the second one refers to the edges percentage and the

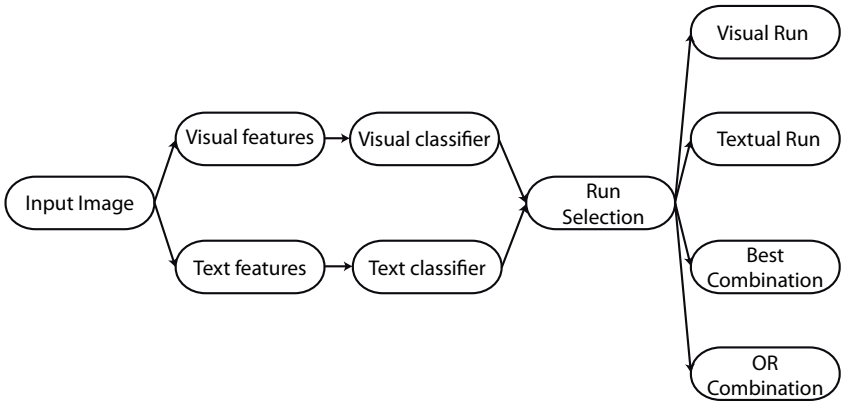

Fig. 6. General scheme of the image concept annotation process

last one contains the bag of words representation. In a preliminary experiment, we tested all combinations of these groups for each concept, storing the best combination for each one.

- **Textual run:** In textual run we also use a different SVM classifier for each concept. Specifically, each classifier uses as input vector the semantic similarity values including only the concept that is being evaluated and its synomyms (described in 2.2), but rejecting the other concepts and their synonyms.
- **"Best" combination:** This run uses the results obtained in the visual and textual runs. In a previous experiment we have identified in which concepts visual features are better than textual features, so this run selects the best option for each concept, according to the results of the previous experiments.
- **"Or" combination:** This run uses the results obtained in the visual and textual runs. Concretely, it marks a concept as relevant for an image if at least one of the classifiers (visual or textual) has marked it as relevant.

3.2 Results

The evaluation of the submission is based on three measures: Mean interpolated Average Precision (MiAP), Geometric Mean interpolated Average Precision (GMiAP) and F1 measure (F-ex). The MiAP value is the average of the average interpolated precisions over all concepts, and the GMiAP gives more importance to the most difficult concepts. Both are measures defined in Information Retrieval context. The F-ex is a metric that uses the binary score to determine how well the annotations are, and it is mostly used in automatic classification problems, where the confidence scores are not commonly used to rank predictions. It is computed by determining the number of true positives, false positives, true negatives and false negatives in terms of detected concepts.

Our submissions were obtained by using a binary classifier, and the confidence score for all evaluation are equal to 1. As MiAP and GMiAP are very dependent

on the confidence scores, our MiAP and GMiAP results are not representative for the quality of the results (0.0622 and 0.0254, respectively, in all submissions). In previous experiments with the training data, and using the MAP of the previous year, we were able to obtain a maximum MAP of 0.20, so the final results have been unexpected. The results of the present competition could be affected due to the change on the evaluation method. However, the main problem has been the confidence score, that, in our case, has been always 1, because of the classifier used.

Table 1 shows the best run of each group ordered by the F-ex measure as well as our 4 different runs. It can be seen that our textual submission is in position 13 out of 18 groups. Moreover, the rest of our submissions would be placed in position 16. Taking into account all the runs submitted by the participants our submissions would be placed in 49, 59, 60 and 64 position out of the 79 runs (and taking into account only the F-ex measure, the only representative for our evaluation).

The textual run is the best run submitted, and the visual run is the worst one. Our own ranking has been also a surprising result for us, as in the previous experimentation the visual runs had always better scores than the textual ones. Our low results on the visual runs are probably due to the structure of the image descriptor. In spite of the features extracted are very representative, the bad use of them in the classification has led us to a lower F-ex value than the one expected.

Group	**MiAP**	**GMiAP**	**F-ex**	**Features**
LIRIS ECL	0.4367	0.3877	0.5766	Multimodal
DMS. MTA SZTAKI	0.4258	0.3676	0.5731	Multimodal
National Institute of Informatics	0.3265	0.2650	0.5600	Visual
ISI	0.4131	0.3580	0.5597	Multimodal
MLKD	0.3185	0.2567	0.5534	Visual
CEA LIST	0.4159	0.3615	0.5404	Multimodal
CERTH-ITI	0.3012	0.2286	0.4950	Multimodal
Feiyan	0.2368	0.1825	0.4685	Textual
KIDS NUTN	0.1717	0.0984	0.4406	Multimodal
UAIC2012	0.2359	0.1685	0.4359	Visual
NPDILIP6	0.3356	0.2688	0.4228	Visual
IntermidiaLab	0.1521	0.0894	0.3532	Textual
URJCyUNED	**0.0622**	**0.0254**	**0.3527**	Textual
Pattern Recognition and Applications Group	0.0857	0.0417	0.3331	Visual
Microsoft Advanced Technology Labs Cairo	0.2086	0.1534	0.2635	Textual
BUAA AUDR	0.1307	0.0558	0.2592	Multimodal
URJCyUNED	**0.0622**	**0.0254**	**0.2306**	Multimodal
URJCyUNED	**0.0622**	**0.0254**	**0.2299**	Multimodal
URJCyUNED	**0.0622**	**0.0254**	**0.1984**	Visual
UNED	0.0873	0.0441	0.1360	Visual
DBRIS	0.0972	0.0470	0.1070	Visual

Table 1. Results of the best run of each group ordered by the F-ex measure

Finally, Table 2 shows the average F-ex of all runs of each group ordered in descending order. We can see that our submissions (URJCyUNED) are in position 15 out of 18 groups, with a F-ex average of 0.2529.

Group	**F-ex**
DMS. MTA SZTAKI	0.5648
National Institute of Informatics	0.5575
ISI	0.5553
LIRIS ECL	0.5414
CEA LIST	0.5131
MLKD	0.5014
UAIC2012	0.4241
CERTH-ITI	0.4173
NPDILIP6	0.4056
Feiyan	0.3662
KIDS NUTN	0.3640
IntermidiaLab	0.3461
Pattern Recognition and Applications Group	0.3000
DBRIS	0.2756
URJCyUNED	**0.2529**
BUAA AUDR	0.1656
Microsoft Advanced Technology Labs Cairo	0.1279
UNED	0.1076

Table 2. Results of the average of all runs of each group ordered by the F-ex measure

Overall results for all de groups participants and the experimental setup can be found in [11].

4 Conclusions

In this paper we describe our first participation in the ImageCLEF 2012 Photo Annotation task. We have used multiple visual features for representing the images, and also textual information, expecting that this information can be used to improve the performance of visual features. Some of the visual features have been defined taking into account the categories of concepts to extract relevant characteristics for the classification. On the other hand, we have used the Flickr user tags to measure the semantic distance between them and the concepts and their synonyms extracted from WordNet. Linear SVM classifiers have been used for the image classification in all submissions.

The evaluation results showed that the best of our submissions is obtained by using textual features only, with a F-ex of 0.35, followed by the multimodal run in which we choose the feature (visual or textual) that have been experimentally better in each concept, obtaining a F-ex of 0.231. Analyzing those results it is easy to see that the combination of visual and textual features has made the

evaluation worse. This is probably caused by the methods of combination that we have chosen. Specifically, we have given the same importance to both features, textual and visual. That kind of combination makes the evaluation almost an average of both methods (textual and visual), which have lead us to a worse F-ex value.

As we can see in the results, visual features classification has obtained lower F-ex values than textual. This result is due to the image descriptor and the classifier used, as the visual features have been experimentally tested. Our main problem seems to be that we have probably chosen a bad descriptor for each visual feature extracted, which has ended in a bad classification result according to our expectation and preliminary experimentation.

Due to the selected classifier, the output for the confidence score is always 1, which means that the classifier is always sure of the presence of the concept in the image. That result has strongly penalized us in terms of MiAP and GMiAP. But the results obtained in the F-ex metric demonstrate that our proposal is competitive and not as bad as it seems to be according to the MiAP and GMiAP metrics. The main aim of future works is the choice of a better and non-binary classifier, as well as the improvement of the visual and textual features representation.

References

1. Binder, A., Samek, W., Kawanabe, M.: The joint submission of the TU Berlin and Fraunhofer FIRST (TUBFI) to the ImageCLEF2011 Photo Annotation Task. CLEF (Notebook Papers/Labs/Workshop). (2011).
2. Canny, J.F.: A computational approach to edge detection. IEEE Trans. Pattern Analysis and Machine Intelligence. (1986) 679-698
3. Csurka, G., Dance, C.R., Fan, L., Willamowski, J. and Bray, C. Visual categorization with bags of keypoints. Workshop on Statistical Learning in Computer Vision, ECCV (2004), 1-22
4. Dalal, N.; Triggs, B.: Histograms of oriented gradients for human detection. Computer Vision and Pattern Recognition (2005). Vol: 886-893
5. Leacock, C. and Chodorow, M.: Combining Local Context and WordNet Similarity for Word Sense Identification. MIT Press, Cambridge, MA. (1998) chapter 11:265-283
6. Lienhart, R. and Maydt, J.: An Extended Set of Haar-like Features for Rapid Object Detection. IEEE ICIP (2002), (1):900-903
7. Liu, N., Zhang, Y., Dellandréa, E., Bres, S. and Chen, L.: LIRIS-Imagine at ImageCLEF 2011 Photo Annotation task. CLEF (Notebook Papers/Labs/Workshop). (2011).
8. Miller, G. A.: WordNet: a lexical database for English. Communications of the ACM. (1995) (38):39–41
9. Nagel, K., Nowak, S., K̀uhhirt, U. and Wolter K.: The Fraunhofer IDMT at ImageCLEF 2011 Photo Annotation Task. CLEF (Notebook Papers/Labs/Workshop). (2011).
10. Su, Y. and Jurie, F. Semantic Contexts and Fisher Vectors for the ImageCLEF 2011 Photo Annotation Task. CLEF (Notebook Papers/Labs/Workshop). (2011).

11. Thomee, B. and Popescu, A. Overview of the ImageCLEF 2012 Flickr Photo Annotation and Retrieval Task. CLEF 2012 working notes. (2012).
12. van de Sande, K.E.A and Snoek, C.G.M.: The University of Amsterdam's Concept Detection System at ImageCLEF 2011. CLEF (Notebook Papers/Labs/Workshop). (2011).
13. Viola, P. and Jones, M.J.: Rapid Object Detection using a Boosted Cascade of Simple Features. IEEE CVPR (2001)
14. Znaidia, A., Le Borgne, H. and Popescu, A.: CEA LIST's Participation to Visual Concept Detection Task of ImageCLEF 2011. CLEF (Notebook Papers/Labs/Workshop). (2011).