URJC&UNED at ImageCLEF 2013 Photo Annotation Task

Jesús Sánchez-Oro[1], Soto Montalvo[1], Antonio S. Montemayor[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,juanjose.pantrigo,abraham.duarte}@urjc.es

[2] Universidad Nacional de Educación a Distancia, Madrid, Spain
{vfresno,raquel}@lsi.uned.es

Abstract. In this work we describe the URJC&UNED participation in the ImageCLEF 2013 Photo Annotation Task. We use visual information to find similar images and textual information extracted from the training set to label the test images. We propose two additional visual features apart from the provided by the organization and a method to expand the textual information available. The new visual features proposed define the images in terms of color and texture, and the textual method uses WordNet to obtain synonyms and hyperonyms of the textual information provided. The score of each concept is obtained by using a co-ocurrence matrix that matches concepts and textual information of the training images. The experimental results show that the proposal is able to obtain competitive results in all the performance measures used.

Keywords: imageclef, image classification, visual features, textual features, automatic image annotation

1 Introduction

This work proposes a new image annotation algorithm used in the Scalable Concept Image Annotation task of ImageCLEF 2013, which is one of the labs of the CLEF 2013 [4]. In this task we receive a set of training images with some visual and textual features associated to each one. The goal is the labeling of a new set of images that only contains visual features, without any textual information. We use visual information to find similar images within the training set, and then we use their associated textual information in order to annotate the test images.

Visual features provided with the images are based on the ones used in previous years, namely, GIST, Color Histograms, SIFT, C-SIFT, RGB-SIFT and OPPONENT-SIFT, all of them saved in a bag-of-words representation. The textual data of the training images contains information about the source of the images, a list of words related to the image and the words used to find the images in three different search engines. More details on the task can be found in [10]

In this work we only use the C-SIFT feature, the words related to the images and the words used to find the images in the search engines. We also propose the use of different new visual features in order to increase the performance of the annotation algorithm, as well as a procedure that extends the textual information provided with the training images.

The rest of the paper is organized as follows: Section 2 describes our proposed visual features and the algorithm to expand the textual information. The algorithm for concept annotation is described in Section 3 and in Section 4 we analyze the results obtained by our proposal. Finally, Section 5 draws the conclusions of the work.

2 Features

The algorithm proposed in this work uses only the C-SIFT visual feature provided by the organization and the list of words related to the image and words used to find it as textual features. In addition to these features we propose two more visual features and a procedure to expand the textual information provided.

2.1 Additional Visual Features

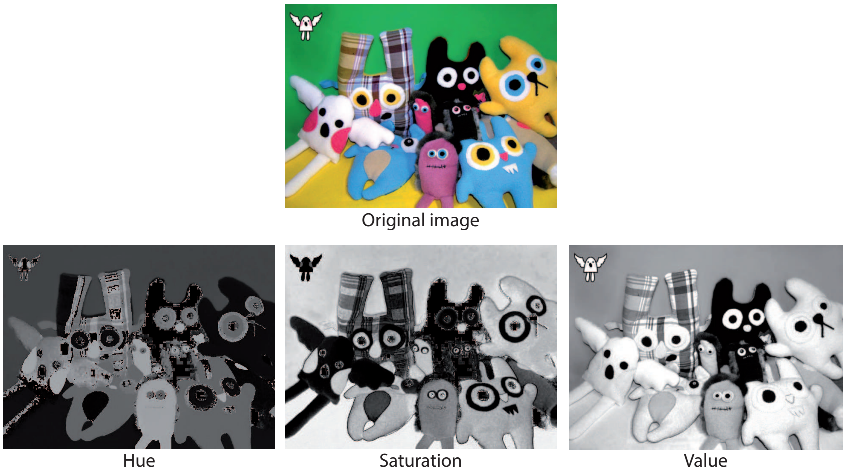
With the new visual features proposed in this work we try to increase the set of features used, that seems to be the same in the last years (SIFT with a bag-of-words representation).

The first visual feature proposed is a color histogram in the HSV space. It is common to use the color histogram of an image when we are trying to find similar images, but those histograms are usually obtained from the RGB space. HSV color space is robust against shadows in the images or changes in the lighting conditions, while in RGB color space a shadow can abruptly change the color value of an area. Specifically, we use the Hue and Saturation channels of the HSV space, and discard the Value channel, because it only stores information about the brightness, which is not useful in this task.

Figure 1 shows the result obtained extracting the channels Hue, Saturation and Value from an image of the training set. In the Hue image it is easy to see that the regions which belongs to the same or similar color are quite uniform, while the Saturation image gives information about how light or dark is the color. The Value image contains details about brightness and shadows in the image, and it is discarded in our proposal. To compare two HSV color histograms we use the Bhattacharyya distance, which is commonly used in histogram comparisons, defined as follows:

$$B(h1,h2) = \sum_{i=1}^{i=|h1|} \sqrt{\frac{h1[i] \cdot h2[i]}{\sum h1 + \sum h2}} \tag{1}$$

where $h1, h2$ are the HSV color histograms of the images that are being compared.



Original image

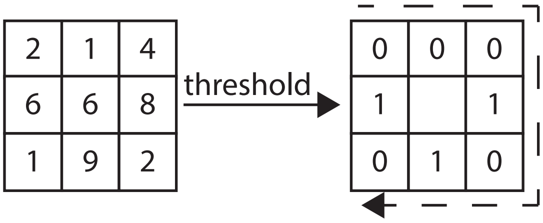
Hue

Saturation

Value

Fig. 1. Example of the extraction of Hue, Saturation and Value channels of a training image (a)

The second additional feature proposed is the Local Binary Patterns method (LBP). It was proposed in [7] and uses information about the texture of the image. The feature is based on the comparison of each pixel to its neighborhood. To analyze the local binary pattern of one pixel, we take the pixel as the center and then we threshold each pixel in the neighborhood against it, obtaining a 1 value if the intensity of the neighbor is higher or equal than the pixel in the center and a 0 value otherwise. Then we concatenate the binary values of the resulting neighbors to obtain a binary chain of 8 elements (3x3 neighborhood). That binary chain is then converted into a $[0-255]$ decimal value, which represents the local binary pattern of the center pixel.



2	1	4
6	6	8
1	9	2

0	0	0
1		1
0	1	0

Fig. 2. Example of the LBP evaluation of a pixel with respect to its neighborhood

Figure 2 shows an example of the evaluation of the LBP code of a pixel. The intensity of the center pixel is used as a threshold for the neighborhood, so values 2, 1 and 4 are converted to 0 and 8, 9 and 6 are converted to 1. The result is read clockwise starting from the upper left corner, obtaining the binary chain 00010101, which is converted in the decimal number 21. Then the intensity value of the center pixel in the new LBP image will be 21. The algorithm to obtain this new intensity value is really efficient because it is only based on threshold evaluations, and it has been successfully used in the detection and tracking of objects, biometrics, biomedical applications, video analysis, etc. [9].

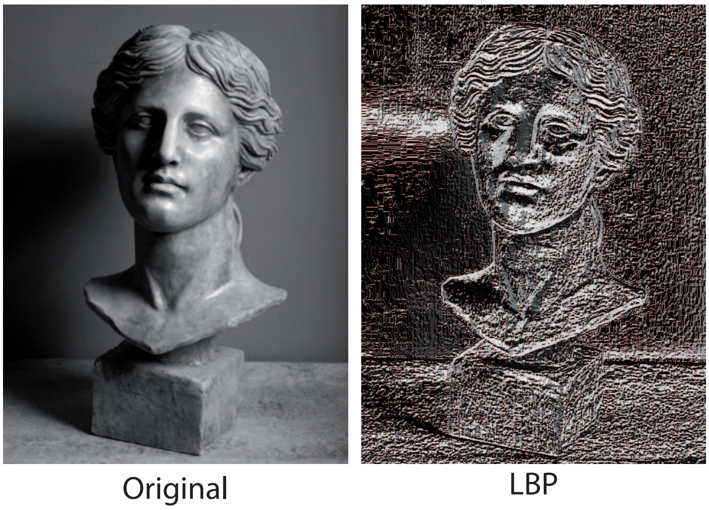


Fig. 3. Example of the LBP evaluation over a training image

Figure 3 shows the result of the LBP evaluation over one of the training images. Although the resulting image is not easy to interpret, it clearly separates the different areas of the image. For instance, it is easy to see that the texture of the face is quite different from the texture of the hair or the background. The important details of pictures are usually in the center of the image, so we remove the 15% of the image frame in order to evaluate only the center part of the image. Once the LBP image has been obtained, we calculate the histogram of resulting image, and we use it as the descriptor of the feature. Then, to compare two LBP histograms we use the Chi-Square distance, as recommended in most LBP relevant works [1, 11], defined as follows:

$$\chi^2(h1, h2) = \sum_{i=1}^{i=|h1|} \frac{(h1[i] - h2[i])^2}{h1[i] + h2[i]} \tag{2}$$

where $h1, h2$ are the LBP histograms of the images that are being compared.

2.2 Additional Textual Features

Several sets of textual features are provided, but we only use two of them. We use the text from the webpages where the images appear and the words used to find each image in the search engines. The webpages which referenced the images and the URLs of the images are not used as explicitly mentioned by the organization.

The processed text extracted from the webpages near where the images appeared are provided joined a score per word, which is derived taking into account the term frequency, the document object model attributes, and the word distance to the image. We use these scores and only select the 95% of words with the highest values. On the other hand, we select all the keywords used to find the images when querying image search engines, independently of the position given to the image in the ranking list.

We build a text representation from the textual features using the lexical database WordNet [5, 8]. WordNet is used to enrich the keywords with synonyms and hyperonyms, because we think these type of textual features could be closer to the meaning of the images than other textual features.

We represent the images by means of a co-occurrence matrix, including stop words removal and stemming for both concepts and words of selected textual features. In the matrix the columns are the concepts to be tagged, and the rows are the different words selected as textual features. Formally, the co-occurrence matrix of an image is a $N \times M$ matrix, where N corresponds to the number of unique words in the set of textual features for the image, and M corresponds to the number of concepts. A cell m_{ij} contains the number of times word w_i co-occurs with concept c_j within the set of textual features of the image. The task of building the co-occurrence matrix is quite common in corpus linguistics and provides the starting point to the algorithm to annotate the image concepts.

Usually, it can be not easy to find the suitable words to search images in a search engine. For this reason expanding the keywords with synonyms allows us to found more co-occurrences with the concepts. Different works confirm that expand with synonyms are useful for different tasks. For instance, [3] shows that indexing with Wordnet synonyms may improve retrieval results, and [2] proposes an expansion with WordNet synonyms for the task of document retrieval in Question Answering. Also, in the last edition of ImageCLEF Photo Annotation task several proposal used WordNet and synonyms to enrich the representation ([6, 12]). In the same way, expanding the keywords with hyperonyms allows us to find more concepts for the images.

3 Concept annotation

The annotation of concepts is performed in two stages. Given a test image, the first stage finds the most similar images among all the images in the training set using only visual features, while the second stage uses textual information of the training images to annotate the test image.

The first stage has been implemented using a k-nearest neighbor algorithm (KNN), with $k = 50$. We have trained three different KNN, one for each visual feature. The training has been carried out by measuring the distance from each test image (from test and devel sets) to each training image. It results in an ordered list for each test image, in which the first element is the most similar training image and the last element is the least similar training image. In order to make the labeling efficient, we saved the training for each feature in a file. With this training, to find the most similar images, we only have to read a file, instead of evaluating the distance among images for each run.

These training files contains, for each test image, the 200 most similar training images for each visual feature (i.e., C-SIFT, HSV and LBP). Then, in the first stage of the concept annotation algorithm we obtain the 50 most similar images from each feature, resulting in the union of the three features, giving the same importance to each feature. Specifically,

$$\begin{aligned} S(image_{test}) = {} & KNN(50, C-SIFT, image_{test}) \bigcup \\ & KNN(50, HSV, image_{test}) \bigcup \\ & KNN(50, LBP, image_{test}) \end{aligned} \tag{3}$$

where $s(image_{test})$ is the set of the most similar images that is going to be used in the second stage. The distance measure used depends on the visual feature used. The C-SIFT feature uses the Euclidean distance, the LBP feature uses the Chi-Square distance and, finally, the HSV feature uses the Bhattacharyya distance.

The second stage is based on the textual features of the training images. For each $image_{train} \in s(image_{test})$ we extract its co-occurrence matrix, as described in Section 2.2. Then, we sum the concept co-ocurrences (each column), constructing a vector with size equal to $n = |\text{concepts}|$ where position i contains the number of occurrences of concept i in the image textual features. Finally, we normalize this vector in the range $[0, 1]$ and use it as the output of the algorithm scores. If the score of a concept exceeds a predefined threshold, then the concept is labeled with 1, else it is labeled with 0.

We have submitted three runs, that differs in the textual features used, as described below:

- **Run 1:** Keywords and words near to the image in the website (selecting only the 95% words with the highest score) of the training images.
- **Run 2:** We add the synonyms of the keywords to the textual features of Run 1.
- **Run 3:** We add the hyperonyms of the keywords to the textual features of Run 2.

As can be seen, the textual features used grows incrementally with each run, with the aim of controlling if synonyms of hyperonyms are useful to improve the results.

4 Results

This section reports the computational experiments that we have performed to obtain the visual and textual features, train the k-nearest neighbor algorithms and finally execute each submitted run. The visual features extraction has been implemented in C++ using the Open Computer Vision 2.4.5 (OpenCV) library and the textual features, KNN training and runs has been implemented in Java 7. The experiments have been performed in an Intel Core i7 2600 CPU (3.4 GHz) and 4 GB RAM. The performance measures used to evaluate the runs are: mean F-measure for the samples (MF-samples), mean F-measure for the concepts (MF-concepts) and the mean average precision for the samples (MAP-samples).

The first experiment is oriented to evaluate the quality of the KNN training algorithm. We do not have ground truth to evaluate this experiment, so the evaluation is only qualitative. The aim of this experiment is to check whether the images obtained by the KNN training are similar to the test images or not. Figure 4 shows an example of the images extracted from each KNN training.

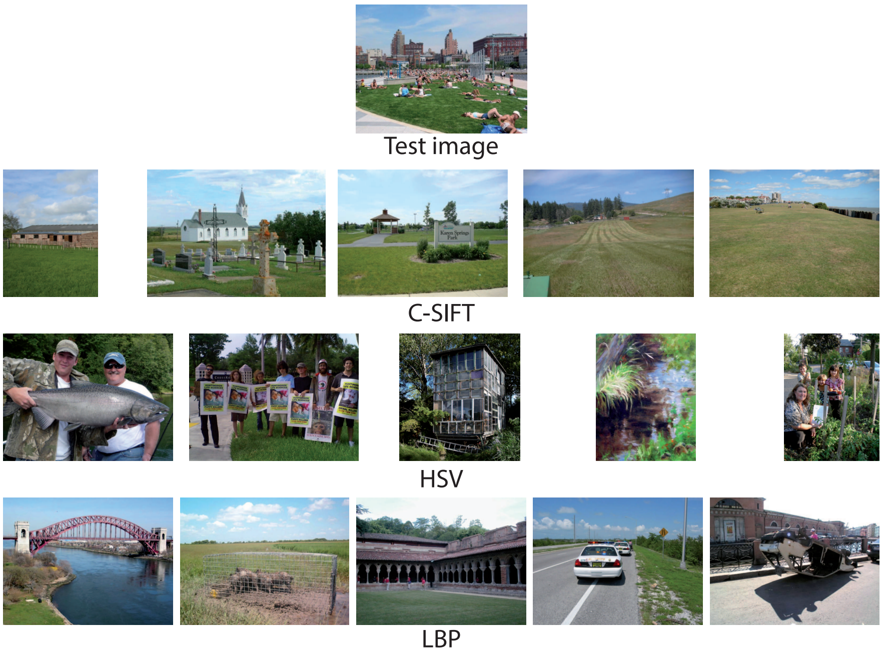


Fig. 4. Test image and the five most similar images using each visual feature (C-SIFT, LBP, HSV)

As can be seen in Figure 4, the images of the KNN training do not overlap, because the information used by each feature is quite different. For that reason we use these three visual features, each one focused in one aspect of the image: color (HSV), shape (C-SIFT) and texture (LBP).

In the second experiment we compare the three textual features proposed to evaluate which one is able to obtain a higher performance. Table 1 shows the results of the three runs using different textual features over the development set. Run 3 is the best run when comparing all the performance measures, but the difference is not big enough to discard Run 1 and Run 2.

	MF-samples	MF-samples	MAP-samples
Run 1	27.4	19.2	32.0
Run 2	27.7	19.7	32.2
Run 3	27.9	19.8	32.6

Table 1. Preliminary results of the runs over the development set

Although Run 3 is slightly better than Run 1 and Run 2 in all the performance measures used, we decided to include the three runs because there is another test set with different concepts in which Run 1 and Run 2 may be better than Run 3.

Finally, we analyze the results obtained by the three runs over the development and test set, comparing them with the other participants. Tables 2 and 3 shows the final performance of our proposal. The values between brackets indicates the ranking of the run among all the 58 participants.

	MF-samples	MF-concepts	MAP-samples
Run 1	27.4 (29)	19.2 (41)	32.0 (35)
Run 2	27.7 (27)	19.7 (40)	32.2 (34)
Run 3	27.9 (26)	19.8 (39)	32.6 (32)

Table 2. Performance measures of the runs over the development set.

	MF-samples	MF-concepts	MF-new_concepts	MAP-samples
Run 1	23.7 (29)	17.1 (41)	14.6 (45)	27.6 (36)
Run 2	23.8 (28)	17.2 (40)	14.6 (44)	27.6 (35)
Run 3	24.1 (27)	17.3 (38)	14.8 (43)	28.1 (33)

Table 3. Performance measures of the runs over the test set.

The results show that our runs are in the middle of the ranking if we analyze the MF-samples performance measure and in the third quarter if we analyze the other measures. As expected, Run 3 is our best run in all the measures, and there are not important differences between Run 1 and Run 2 in both development and test set. The union of the visual features have resulted in a good method to look for similar images. The results also show that the method to expand textual information has improved the results obtained.

5 Conclusions

In this paper we describe our participation in the ImageCLEF 2013 Photo Annotation Task. The algorithm proposed is divided in two stage. The first stage uses only visual features while the second stage take advantage of the available textual information. We propose two additional visual features apart from the C-SIFT information that are able to analyze the image focusing in different features: color, shape and texture. We also propose a method to expand the available textual information with synonyms and hyperonyms, compare that information with the concepts and give a score for each concept depending on the comparison. The results show that the best run (Run 3) takes advantage of the visual features proposed and the method to expand the textual information to improve the results of the annotation. Our submissions are in the middle of the ranking analyzing the MF-samples measure and in the third quarter analyzing the other measures. The main aim of future works is the improvement of the annotation algorithm, as well as the addition of new visual and textual features that lead us to improve our performance.

Acknowledgments

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, HiSCoP, DIETHA and SAMOA3D funded by the "Ministerio de Ciencia e Innovación" under grants TIN2010-21128-C02, TIN2008-06890-C02, TIN2012-35632 and TIN2011-28151 respectively.

References

1. Ahonen, T., Hadid, A., Pietikainen, M.: Face Description with Local Binary Patterns: Application to Face Recognition IEEE Transactions on Pattern Analysis and Machine Intelligence, 28(12):2037–2041 (2006)
2. Bernhard, D.: Query expansion based on pseudo relevance feedback from definition clusters Proceedings of the 23rd International Conference on Computational Linguistics: Posters, 54–62
3. Buscaldi D., Rosso P.: Indexing with wordnet synonyms may improve retrieval results. Proceedings of the 10th cross-language evaluation forum conference on Multilingual information access evaluation: text retrieval experiments, 128–134
4. Caputo, B., Muller, H., Thomee, B., Villegas, M., Paredes, R., Zellhofer, D., Goeau, H., Joly, A., Bonnet, P., Martínez Gómez, J., García Varea I., Cazorla, M.: ImageCLEF 2013: the vision, the data and the open challenges. Proc CLEF 2013, LNCS (2013)

5. Fellbaum C.: WordNet: An Electronic Lexical Database Cambridge, MA: MIT Press. (1998)
6. Manzato M.G.: The participation of IntermidiaLab at the ImageCLEF 2012 Photo Annotation Task. CLEF Online Working Notes/Labs/Workshop (2012)
7. Maturana, D., Mery, D., Soto, A.: Face Recognition with Local Binary Patterns, Spatial Pyramid Histograms and Naive Bayes Nearest Neighbor Classification Proceedings of the 2009 International Conference of the Chilean Computer Science Society, 125–132 (2009)
8. Miller G.A.: WordNet: A Lexical Database for English Communications of the ACM, 38(11):39–41 (1995)
9. Pietikinen, M., Hadid, A., Zhao, G., Ahonen, T.: Computer Vision Using Local Binary Patterns Springer (2011)
10. Villegas, M., Paredes, R., Thomee, B.: Overview of the ImageCLEF 2013 Scalable Concept Image Annotation Subtask. CLEF 2013 Working Notes, Valencia, Spain (2013)
11. Zhang G., Xiangsheng H., Li S., Wang Y., Wu X.: Boosting Local Binary Pattern (LBP)-Based Face Recognition Advances in Biometric Person Authentication LNCS 3338:179–186 (2005)
12. Znaidia, A., Shabou, A., Popescu, A., Borgne, H.L.: CEA LIST's Participation to the Concept Annotation Task of ImageCLEF 2012 CLEF Online Working Notes/Labs/Workshop (2012)