The PRA and AmILAB at ImageCLEF 2012 Photo Flickr Annotation Task

Luca Piras[1], Roberto Tronci[1,2], Gabriele Murgia[2], and Giorgio Giacinto[1]

[1] DIEE - Department of Electric and Electronic Engineering
University of Cagliari, Italy
{luca.piras,roberto.tronci,giacinto}@diee.unica.it
[2] AmILAB - Laboratorio Intelligenza d'Ambiente, Sardegna Ricerche, Italy
{roberto.tronci,gabriele.murgia}@sardegnaricerche.it

Abstract. This paper presents the first participation of the Pattern Recognition and Application Group (PRA Group), and the Ambient Intelligence Lab (AmILAB) at the ImageCLEF 2012 Photo Flickr Concept Annotation Task. In this task, the teams' goal is to detect the presence of 94 concepts in the images, and to provide a confidence score related to the confidence of the decision of each concept detector. We faced the challenge by relying on visual information only, combining different image descriptors by means of different score combination techniques. Experimental results show that just combining concept detectors not specifically designed for handling the large variety of concepts does not allow reaching satisfactory results.

Keywords: image annotation, dynamic score combination, SVM

1 Introduction

The visual concept annotation task is a multi-label classification challenge where the goal consists in the analysis of a collection of photos in order to detect the presence of one or more concepts. The number of selected concepts is equal to 94, and their semantics cover a wide range. They include categories related to persons (e.g. baby, child, teenager, adult), animals (e.g. cat, dog, horse), and sentiments (e.g. unpleasant, euphoric). In addition to the images and the associated concepts, participants are provided with textual features, and visual features. Our main objective to solve this task is to use the combination of outputs of visual concept detectors based on visual descriptors. A detailed overview of the data set, and the related task can be found in [4].

2 Visual features

For this task, a subset of the MIRFLICKR[3] collection has been used. This subset comprises 25 thousand images that have been manually annotated using a limited

[3] http://press.liacs.nl/mirflickr/

number of concepts. With respect to the previous editions of the competition, this year the annotation process has been carried out by resorting to crowdsourcing mechanisms. Several concepts have been reused of last year's task, and, for most of these concepts, the remaining photos of the MIRFLICKR-25K collection that had not yet been used in the previous task, have been annotated. In order to boost the quality of all 25,000 images, they have been reannotated for several concepts too. All the images have been naturally annotated for the new concepts. All images have been accompanied by different kind of features: textual, and visual features. Detailed information about the feature sets can be found in [4].

In our approach, we focused on visual descriptors only. The visual descriptors proposed for this task are the following: *sift*, *c-sift*, *RGB-sift*, and *Opponent-sift*. For each descriptor, the histogram of the occurrence frequencies has been extracted by using the **Color Descriptors** toolkit [3]. As expected, the K-means clustering used to produce a "bag of visual words" representation, is quite slow for the data sizes at hand, as clustering 250,000 points takes at least 12 hours per iteration. The solution usually proposed is to reduce the number of points to cluster. By default, the toolkit extracts 250,000 points regardless of the number of training images, thus reducing the number of point per image automatically as the number of images increase. It means that the toolkit extracts less than 17 points per image, thus loosing dozens of descriptors. At the same time, if up to 200 points are extracted for each of 15,000 training images, the K-means algorithm should cluster 3,000,000 points!

For this reasons we decided to divide the 15,000 training images into four groups, by retaining the same proportion of image per concept as in the whole training set. Then, for each group, we clustered around 750,000 points in order to obtain four different codebooks (one codebook for each descriptor) with 2,048 *visual words*. Each codebook has then be used to produce the Bag of Visual Words descriptors. This procedure allowed obtaining a large vocabulary of "visual words", and at the same time reduced the number of points to cluster.

3 Concept detection by dynamic combination of visual classifiers

We submitted three runs in total. All of these runs are based only on the bag of visual words descriptors illustrated in the previous section.

For all the runs, we used the Multiple Classifier System paradigm, and the Support Vector Machine has been used as the base classifier for its good performances on various image classification tasks [1]. We trained a single SVM for each global image descriptor and each visual concept. Thus, for each concept i, a set of four SVMs $\{S_i^{sift}, S_i^{rgb}, S_i^{color}, S_i^{opponent}\}$ is available.

We classified all the pattern of the test set by means of these sets of SVMs. Thus, for each test pattern x_j, we obtained as output the class decision d_{ij}^k taken by the classifier k (i.e., 1 if the pattern belongs or not to the concept i, 0 otherwise), and the distance from the decision border is transformed through a min-max normalization into a classification score s_{ij}^{k}, in the range $[0,1]$, of the test pattern with respect to the concept.

We used the following three combination rules:

- The Mean rule
- The Dynamic Score Selection by Majority Vote
- The Dynamic Score Selection by Mean rule

For the *Mean rule*, we computed the average of the classification scores obtained from the classifiers [2]:

$$s_{ij}^{mean} = \frac{1}{k} \cdot \sum_{k} s_{ij}^{k} \tag{1}$$

In the case of the other two combination rules, we used the Dynamic Score Selection (DSC) approach [5]:

$$s_{ij}^{dsc} = (1-\alpha) \cdot \min_{k}\{s_{ij}^{k}\} + \alpha \cdot \max_{k}\{s_{ij}^{k}\} \tag{2}$$

This combination rule is able to perform a dynamic combination at the score level, by allowing to dynamically chose the best scores and weights to be combined. In [5] different methods to compute dynamically the weights α are proposed. In these runs, we used one of those methods, and one that has been specifically designed for the task at hand.

The rule for computing α for the *Dynamic Score Selection by Majority Vote* is the following:

$$\alpha = \begin{cases} 1, & \text{if at least half of the } d_{ij}^{k} \text{ are equal to 1} \\ 0, & \text{otherwise} \end{cases} \tag{3}$$

The rule for computing α for the *Dynamic Score Selection by Mean Rule* is the following:

$$\alpha = \frac{1}{k} \cdot \sum_{k} s_{ij}^{k} \tag{4}$$

4 Results and Discussion

The performances (Interpolated Mean Average Precision (MiAP), Interpolated Geometric Mean Average Precision (GMiAP), and F1-measure on all concepts related to our runs are listed in Table 1, and they are compared to the performances obtained by the other team that used visual features only. Detailed information about the evaluation process can be found in [4].

A first conclusion that can be drawn from Table 1 is that using a combination of general purpose classifiers does not permit to obtain very satisfactory results, as we obtained just the tenth position out of thirteen participants.

The proposed results also show that the *Dynamic Score Selection by Majority Vote* does not work as expected, as it is outperformed by the *Dynamic Score Selection by Mean Rule* for the MiAP and GMiAP measures, and by the *Mean rule* for the F1-measure.

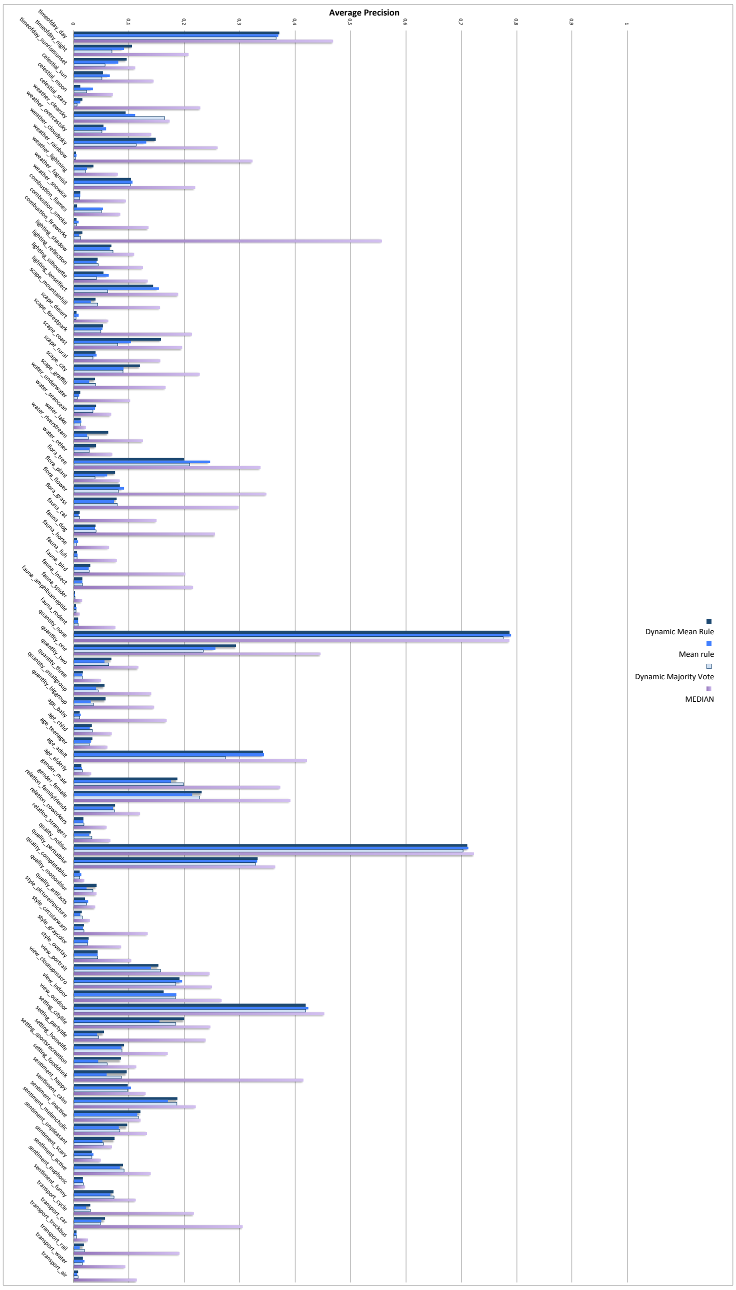

Fig. 1. Interpolated Mean Average Precision (MiAP) for each of 94 concepts

Table 1. Interpolated Mean Average Precision (MiAP), Interpolated Geometric Mean Average Precision (GMiAP), and F1-measure of teams' annotations on all concepts combined

	MiAP	GMiAP	F-ex
DBRIS	0.0976	0.0476	0.1006
Feiyan	0.0819	0.0387	0.0429
ISI	0.3243	0.259	0.5451
RTH	0.2628	0.1904	0.4838
LIRIS ECL	**0.3481**	**0.2858**	0.5437
MLKD	0.3185	0.2567	0.5534
MicroSoft ATL Cairo	0.0868	0.0414	0.1069
NPDILIP6	0.3437	0.2815	0.4199
National Institute of Informatics	0.3306	0.2694	**0.5566**
PRA (Mean rule)	0.0857	0.0417	0.3331
PRA (Dyn. Majority Vote)	0.0837	0.0403	0.3140
PRA (Dyn. Mean Rule)	0.0900	0.0437	0.2529
UAIC2012	0.2359	0.1685	0.4359
UNED	0.102	0.0512	0.1081
URJCyUNED	0.0622	0.0254	0.1984

5 Conclusions

In our participation to the ImageCLEF photo annotation task, multiple visual features has been used for representing the images. We combine the different information using the Bag-of-Words model taking care that a number of image descriptor big enough was used for each image. After the BoW extraction, we combined the four feature spaces in three different ways. The evaluation results showed that a simple combination of different feature spaces using classifiers not specifically designed for taking into account the big variety of concepts is not able to reach satisfactory results.

References

1. Cristianini, N., Shawe-Taylor, J.: An Introduction to Support Vector Machines and Other Kernel-based Learning Methods. Cambridge University Press (2000)
2. Kittler, J., Hatef, M., Duin, R.P.W., Matas, J.: On combining classifiers. IEEE Trans. Pattern Anal. Mach. Intell. 20(3), 226–239 (1998)
3. van de Sande, K.E.A., Gevers, T., Snoek, C.G.M.: Evaluating color descriptors for object and scene recognition. IEEE Trans. Pattern Anal. Mach. Intell. 32(9), 1582–1596 (2010)
4. Thomee, B., Popescu, A.: Overview of the imageclef 2012 flickr photo annotation and retrieval task. CLEF 2012 working notes, Rome, Italy (2012)
5. Tronci, R., Giacinto, G., Roli, F.: Dynamic score combination: A supervised and unsupervised score combination method. Machine Learning and Data Mining in Pattern Recognition 5632, 163–177 (2009)