LIG-MRIM at Image Photo Annotation task in ImageCLEF 2011

Rami Albatal[1], Bahjat Safadi[2], Georges Quénot[3], and Philippe Mulhem[3]

[1] LIG-UPMF
[2] LIG-UJF
[3] LIG-CNRS
385 av. de la bibliothèque, 38041 Grenoble Cedex
Rami.Albatal@imag.fr

Abstract. We describe in this paper the different approaches tested for the Photo Annotation task for CLEF 2011. We experimented state of the art techniques, by proposing late fusions of several classifiers trained on several features extracted from the images. The classifiers are SVMs and the late fusion is a simple addition of classification probabilities coming from the SVMs. The results obtained place our runs in the middle of the pack, with our best visual-based MAP at 0.337 We also integrated of Flickr human annotations, leading to a large increase of the MAP with a value of 0.377.

1 Introduction

This paper aims at describing the proposal and results of the LIG-MRIM research group at the Photo Annotation task for CLEF 2011. The proposal of the group focused mainly on applying a late fusion on multiple learners based on SVM. We also experimented some processes to reduce the feature space dimensions, and we made use of a simple integration with Flickr tags. The findings according to the official evaluations confirm that: late fusion of multiple features lead to good result, that dimension reduction on few features is an interesting direction to focus on, and that a simple integration of human assigned tags improves results.
The corpus [4] for this year is composed of a training set of 8,000 images and the test set is 10,000 images large. The image annotation is a multi-label classification process, where the 99 labels go from image elements (like *Flowers*), to feelings generated by the images (like *scary*). The images are possibly associated with EXIF data, as well as with Flickr tags provided by human. The main evaluation is MAP-based, and we focus here only on this measure to evaluate our runs.
The outline of this paper is the following. In section 2, we begin describe the visual feature extracted and their representation. Section 3 presents the processing applied on Flickr tags. In section 4, we focus on the classification applied on the extraction, as on the fusion processed between the different learners results. In section 5, we list our results, and we conclude in section 6.

2 Extraction and representation of visual features

We focus here the feature extracted, as well some PCA-based dimension reduction on some features. The features considered cover most the common feature we find the in literature.

2.1 Simple features

The features that were extracted are color-based as well as texture based. Some features are extracted globally from the whole image, and others are extracted from image regions, before being aggregated to represent one image. In the following, we give an identifier for each feature before explaining the extracted feature. Such identifiers will be reused in section 5.

Global features

- h3d_64: normalized RGB Histogram. Such color-based histogram is 64 dimensions large, using a simple 4 x 4 x 4 subsampling respectively the R, G and B colors components;
- gab_40: normalized Gabor transform [2]. For this texture-based feature, we select 8 orientations at 5 scales, leading to a 40 dimensions space for these histograms;
- hg_104: this feature results in a simple concatenation of the two representations above (h3d64 and gab40), generating a 104 dimensions space for the histograms.

Local features The local features extracted are SIFT-like. They are extracted for regions of the images, resulting from dense sampling of harris-laplace region of interest detection. Each of these features are represented as bag of visual word, similarly to [1]; the visual vocabulary is generated using a Kmeans algorithm on a sample of the features extracted from the training set;

- opp_sift_har_1000 and opp_sift_har_4000: opponent sift features with Harris-Laplace region of interest detector, generated using Koen Van de Sande's software [5]. Two representations are considered: one of 1000 and one of 4000 dimensions.
- opp_sift_dense_1000: features similar to above, except that the regions or obtained by dense sampling every 8 pixels of the images. The bag of word representation generates 1000 dimensional histograms;
- rgSift_har_4000: rgSIFT features are extracted based on regions obtained by the Harris-Laplace detector. The same tool than above is used to generate the 1000 dimensions histograms;
- rgSift_dense_4000: the rgSIFT are extracted with dense sampling selection. The size of the histograms is 4000 dimensions;

All the features described above are based on 1 nearest neighbor assignment for the generation of the bag of visual words histograms. As described in [6], softer assignments may be used. We experiment those on opponent sift features:

- opp_sift_har_unc_1000 and opp_sift_har_unc_4000: opponent sift features extracted from region generated by Harris Laplace detectors, with soft assignment. The space dimensions are respectively here 1000 and 4000;
- opp_sift_dense_unc_1000: opponent sift features extracted from dense sampling, with soft assignment. The space dimensions are 1000.

2.2 Dimension reduction on features representations

As shown by [3], some space dimension reduction do not necessarily degrades the results, and has a large advantage during the learning phase. That is why we applied PCA-based dimension reduction on some of the large spaces defined in the previous subsection. First, to modify the values in histograms bin we apply a power law normalization, similar to [3], so that the normalized value v_{norm} for each bin of the histograms is: $v_{norm} = v^{\alpha}$, with v the initial value of the bin, and α a float number depending on the collection. On the normalized histograms, we reduce the dimensions to a fixed number by using Principal Component Analysis (PCA). The resulting features are generated using the same a power law normalization with $\alpha = 0.500$ or $\alpha = 0.450$ (according to the _pw in the identifier) and PCA reduction to 400 dimensions, leading to:

- rgsift_har_4000_pw0.500p400: from rgsift_har_4000;
- rgsift_dense_4000_pw0.500p400: from rgsift_har_4000;
- opp_sift_har_1000_pw0.450_p400 and opp_sift_har_4000_pw0.450_p400: from respectively opp_sift_har_1000 and opp_sift_har_4000;
- opp_sift_dense_1000_pw0.450_p400: from opp_sift_dense_1000;
- opp_sift_har_unc_1000_pw0.450_p400: from opp_sift_har_unc_1000;
- opp_sift_dense_unc_1000_pw0.450_p400: from opp_sift_dense_unc_1000;

For the “low dimensional” features h3D64, gab40 and hg104, similar techniques lead to h3d_64_pw0.250_32, gab_40_pw0.500_20 and hg_104_pw0.375_54, when considering reducing the dimensions by a half.

3 Extraction and representation of Flickr tags

As Image annotation collection is an excerpt from Flickr, the human generated tags are available. We know that such manually input tags are not always easy to process (typos, jokes, etc.), but we propose a simple way to handle some of them. First, for each image, we split the tags into words, and we apply a Porter stemmer in a way to group similar words into classes. In a second step, if one stemmed tag equals one of the 99 stemmed labels, then the label is selected for the image. The resulting representation is a 99 dimensions binary vector, with 1 if the label describes the image and 0 otherwise.

4 Classification

4.1 Visual only

All the classification processes on the visual features use Multiple-SVM classifiers based on Radial Basis Funcion (RBF) kernels, since it was

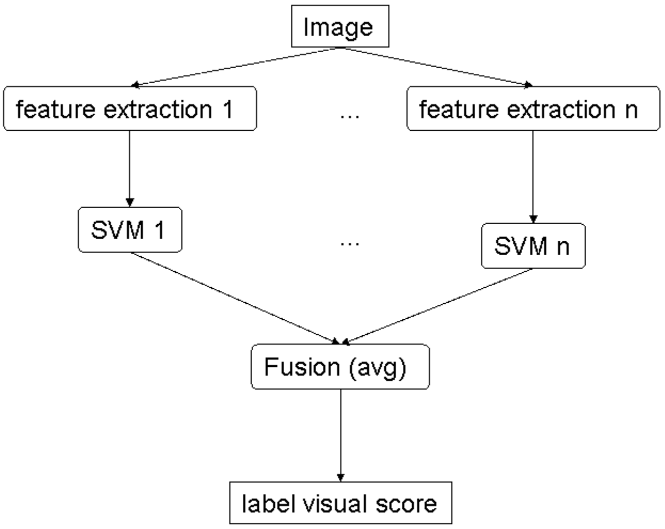

Fig. 1. Global classification process for visual features

proved to be a good solution for data imbalance problems. Such problems occur for many labels in the collection under consideration here. So, for each label, we get positive and negative samples that are used as input for the learning of the Support Vectors.
During the classification, the image representation is input to the SVM using each model, and a binary classification is processed. We assume here that each classifier outputs a probability of classification in [0,1]. The final score for each label is then the average of each individual score from each classifier of the label, as shown in figure 1.

4.2 Visual + Flickr tags

For integrating Flickr tags and visual elements, we also use a late fusion approach. In this case the visual classification result for each label is fused using a max with the label value for the image according to the Flickr tags processing described earlier. The overall process is described in figure 2.

5 Validation set Results

We present the MAP results obtained on a validation set. Our training set is composed of 2/3rd of the official training set, generated randomly with a post processing ensuring a similar distribution of the tags that on the official training set. This last point is important, especially for the

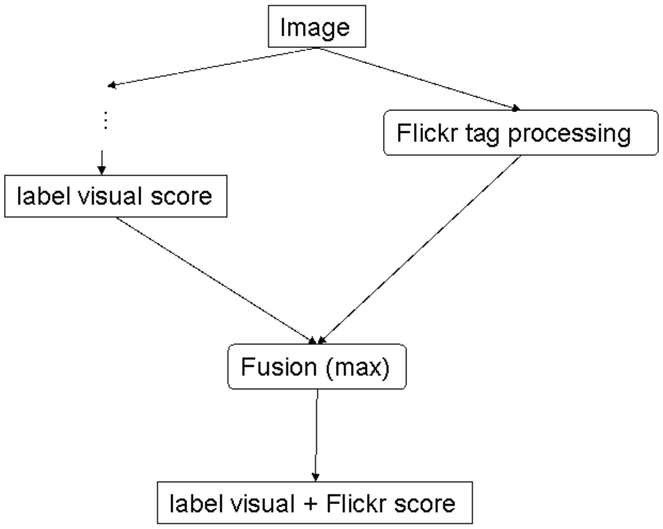

Fig. 2. Global classification process for visual + Flickr features

labels that have only few samples (like *skateboard* with only 12 positive samples). The validation set in composed of the 1/3rd remaining images of the official training set.

The table 1 presents the results obtained feature by feature for each visual feature listed in section 2.1. This table shows that all the SIFT-based features with hard assignment behave consistently, with MAP values between 0.246 and 0.258. The soft assignment opp_sift_har_unc_1000 outperforms slightly the hard assignments, but only marginally. We notice also that the hg_104 features behave surprisingly well compared to SIFT-like features.

The table 2 focuses on the results obtained when considering the dimension reduction process depicted in part 2.2. In the last column of this table, we list the percentage of increase compared to the original (i.e., not reduced) features. This table shows that the reduction of dimension proposed always outperforms the original features. This result is especially visible with the opponent sift features with strict assignment. In any cases, the dimension reduction seems effective for harris laplace features, and less for dense sampling-based features. For the "low dimansional" features, we notice also a large imrpovement with one one half reduction, leading to very good results for hg_104_pw0.375_54, which has onlyt 54 dimensions.

The last table, 3 of this section deals with the results obtained after fusing the results, according to the explanations of section 4. We chose three fusions, which correspond to the configuration of the official run submitted:

Table 1. MAP results on the validation set

Descriptor identifier	MAP
h3d_64	0.186
gab_40	0.213
hg_104	0.243
opp_sift_har_1000	0.252
opp_sift_har_4000	0.253
opp_sift_dense_1000	0.255
rgSift_har_4000	0.246
rgSift_dense_4000	0.258
opp_sift_har_unc_1000	0.262
opp_sift_dense_unc_1000	0.255

Table 2. MAP results on the validation set for reduced feature representations

Descriptor identifier	MAP (increase vs. no reduction)
opp_sift_har_1000_pw0.450_p400	0.273 (+ 8.33%))
opp_sift_har_4000_pw0.450_p400	0.282 (+ 11.46%)
opp_sift_dense_1000_pw0.450_p400	0.267 (+ 4.71%)
rgsift_har_4000_pw0.500p400	0.264 (+ 7.3%)
rgsift_dense_4000_pw0.500p400	0.270 (+ 4.7%)
opp_sift_har_unc_1000_pw0.450_p400	0.280 (+ 6.9%)
opp_sift_dense_unc_1000_pw0.450_p400	0.267 (+ 1.9%)
h3d_64_pw0.250_32	0.211 (+ 13.44%))
gab_40_pw0.500_20	0.215 (+ 0.94%))
hg_104_pw0.375_54	0.259 (+ 6.58%))

- msvm: the late fusion of all the 20 visual features considered earlier in the paper;
- msvm_tags: the late fusion of the visual scores and the Flickr tags scores;
- msvw_two_desc: the late fusion of the two best features according the table 2, but considering two different kinds of regions for the features (i.e. one Harris-Laplace based, and one dense sampling based) to ensure variability in the fused results: opp_sift_har_unc_1000pw0.450p400 and opp_sift_dense_1000pw0.450p400.

The conclusions drawn from this table is that the fusion always outperforms each f its components (such result is well known in the community). We see here that Flickr tags integration, even is the processing is quite straightforward, leads to an important increase of the results.

These three configurations are the ones used for the official submissions.

Table 3. MAP results on the validation set for three late fusions

Descriptor identifier	MAP (increase vs. best visual feature in the fusion)
msvm	0.314 (+ 11.35%)
msvm_tags	0.357 (+ 26.60%)
msvw_two_desc	0.297 (+ 6.07%)

6 Official Results

We present here the official MAP results obtained from our runs in 4. This table shows also in the last column the rank obtained in comparable lists (i.e., list of visual results for msvm and msvw_two_desc, and list of multi-modal results for msvm_tags). The results obtained place our best visual run, msvm with a MAP of 0.336, in the first tier of the list, and above the average and the median values respectively of 0.289 and 0.323. For the multimodal run, msvm_tags with a MAP of 0.378, the rank is above the middle, and also above the average and the median values of respectively 0.370 and 0.371 .

Table 4. Official MAP results for the submitted MRIM runs

Descriptor identifier	Official Id	MAP	rank (in comparable list)
msvm	1308318230664	0.336	15/46
msvm_tags	1308226825708	0.378	11/25
msvw_two_desc	1308318529187	0.324	23/46

7 Conclusion

This paper presented the worjk of the LIG-MRIM team for the Photo Annotation task for CLEF 2011. We used a large set of 20 features, with or without strict bin assignment, dimension reductions, with and without integrating Flickr tags. The results obtained place our run in the first tier for the visual runs, and in the first half for the multimedia runs. In the future, we will focus on dimension reductions to find out what reductions are useful.

Aknowledgements

This work was partly supported by: a) the Quaero Programme, funded by OSEO, French State agency for innovation and b) the Région Rhones Alpes (projet LIMA).

References

1. G. Csurka, C. R. Dance, L. Fan, J. Willamowski, and C. Bray. Visual categorization with bags of keypoints. In *In Workshop on Statistical Learning in Computer Vision, ECCV*, pages 1–22, 2004.
2. B. S. Manjunath and W. Y. Ma. Texture features for browsing and retrieval of image data. *IEEE Trans. Pattern Anal. Mach. Intell.*, 18:837–842, August 1996.
3. T. Mensink, G. Csurka, F. Perronnin, J. Sánchez, and J. Verbeek. Lear and xrces participation to visual concept detection task - imageclef 2010. In *Working Notes for the CLEF 2010 Workshop*, page 48, sep 2010.
4. S. Nowak, K. Nagel, and J. Liebetrau. The clef 2011 photo annotation and concept-based retrieval tasks. In *CLEF 2011 working notes*, 2011.
5. K. E. A. van de Sande, T. Gevers, and C. G. M. Snoek. Evaluating color descriptors for object and scene recognition. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 32(9):1582–1596, 2010.
6. J. C. van Gemert, C. J. Veenman, A. W. Smeulders, and J.-M. Geusebroek. Visual word ambiguity. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 32:1271–1283, 2010.