UAIC participation at ImageCLEF 2012 Photo Annotation Task

Mihai Pîţu, Daniela Grijincu and Adrian Iftene
UAIC: Faculty of Computer Science, "Alexandru Ioan Cuza" University, Romania
{mihai.pitu, daniela.grijincu, adiftene}@info.uaic.ro

Abstract. This paper presents the participation of our group in the ImageCLEF 2012 Photo Annotation Task. Our approach is based on visual and textual features as we experiment with different strategies in order to extract the semantics inside an image. First, we construct a textual dictionary of tags using the most frequent words present in the user tag annotated images from the training data sets. A linear kernel is then developed based on this dictionary. To gather more information from the images we further extract local and global visual features using TopSurf and Profile Entropy Features as well as Color Moments technique. We then aggregate these features with Support Vector Machines classification algorithm and train separate SVM models for each concept. In the end, to improve our system's performance, we add a post-processing step that verifies the consistency of the predicted concepts and also applies a face detection algorithm in order to increase the recognition accuracy of the person related concepts. Our submission consists of one visual-only and four multi-modal runs. We further give a more detailed perspective of our system and discuss our results and conclusions.

Keywords: ImageCLEF, Image classification, Photo annotation, SVMs, TopSurf, Bag-of-Words model, kernel methods, PEF, Color moments

1 Introduction

ImageCLEF 2012[1] Photo Annotation Task[2] represents a competition that aims to improve the state of the art of the Computer Vision field by addressing the problem of automated image annotation [1]. The participants are asked to create systems that can automatically assign an image a subset of concepts from a list of 94 possible visual concepts.

In 2012, the organizers offered a database consisting of 15,000 training images annotated with the corresponding 94 binary labels and a set of 10,000 test images which were to be automatically annotated (see Figure 1). The images were extracted from Flickr[3] online photo sharing application and so each image had the associated EXIF data and Flickr user tags. Among the reasons that make this task of image annotation a difficult one are the diversity of the concepts simultaneously present in

[1] ImageCLEF2012: http://www.imageclef.org/2012
[2] Photo Annotation Task: http://www.imageclef.org/2012/photo
[3] Flickr: http://www.flickr.com/

an image, the subjectivity of the existing annotations in the training set (especially regarding the feelings related concepts) and the fact that the training samples are unbalanced and so there may be more examples for a concept then for another.

Figure 1: Examples of train/test images

The system we propose combines different state of the art image processing techniques (TopSurf, PEF, Color Moments) with Support Vector Machines and Kernel functions we defined in an attempt to obtain good overall performances. This was our second participation in Photo Annotation task, after our contribution from 2009 [2].

The rest of the article is organized as follows: Section 2 presents the visual and textual features we extracted to describe the images, Section 3 covers the classification and post processing modules of our system, Section 4 details our submitted runs and Section 5 outlines our conclusions.

2 Visual and Textual Features

2.1 Local Visual Features – TopSurf

TopSurf[4] [3] is a visual library that combines SURF interest points [4, 5] with visual words based on a large pre-computed codebook [6, 7] and returns the most important visual information in the image (based on assigned Tf-Idf scores[5]). SURF interest points and the associated descriptors provide (partial) invariance to affine transformations of objects in images, but the number of interest points may vary between 0 and a few thousands, depending of the size and details of a photo. Because to every SURF interest point corresponds a descriptor (a 64 dimensional array), the

[4] TopSurf: http://press.liacs.nl/researchdownloads/topsurf/
[5] Tf-Idf: http://tfidf.com/

problem of matching such descriptors arises. As matching thousands of descriptors of a given image against a large database is highly time consuming and practical infeasible, TopSurf library assigns every SURF descriptor a visual word from the pre-computed codebook and associates a limited number (the most important) of such visual words to the image. The time of the extraction process slightly increases (experiments [3] shows that for SURF interest point extraction is required on average 0.37s and 0.07s for the assignment of the visual words), but matching TopSurf descriptors improves the time complexity and quality of the overall process.

The TopSurf library assigns Tf-Idf scores [8] to every visual word in the image and returns the most important ones. In our system we use the cosine similarity to measure the distance (angle) between two given images described by their corresponding TopSurf descriptor:

$$simCos(d1, d2) = \cos(\varphi) = \frac{d1 * d2}{||d1||\,||d2||} = \frac{\sum_{i=1}^{N} d1_i\, d2_i}{\sqrt{\sum_{i=1}^{N}(d1)^2}\sqrt{\sum_{i=1}^{N}(d2)^2}}$$

The similarity score will be between 0 and 1 (because the angle of the vectors $d1$ and $d2$ is smaller than 90 degrees), with 1 for identical descriptors and 0 for absolutely different ones. The time needed to compare these descriptors is, on average, 0.2 ms (with a database of 100,000 images).

2.2 Profile Entropy Features

Profile Entropy Features (PEF) [9] is a technique of extracting global visual features which combines the texture characteristics with the shapes present in a given image by computing the simple arithmetic mean in horizontal or vertical direction.

The PEF features are computed on an image I by using the normalized RGB channels: $r = \frac{R}{l}$, $g = \frac{G}{l}$, $b = 1 - r - g$, where $l = \frac{R+G+B}{3}$. The profiles of the orthogonal projections of the pixels to the horizontal X axis is noted Π_X^{op} and to the vertical Y axis (Π_Y^{op}), where op is the projection operator (arithmetic or harmonic mean). The length of a profile is $S = C(I)$ or $S = L(I)$ (where $C(I)$ denotes I's columns and $L(I)$ denotes I's rows) and we estimate its probability distribution function ($p\,d\,f$) on $N = round(\sqrt{S})$ bins [10]. Then for each channel and operator, we compute: $\Phi_X^{op}(I) = pdf(\Pi_X^{op})$ and we set PEF components to the normalized entropy of this distribution:

$$PEF_X(I) = \frac{H(\Phi_X^{op}(I))}{\log N}$$

$$PEF_Y(I) = \frac{H(\Phi_Y^{op}(I))}{\log N}$$

$$PEF_B(I) = \frac{H(pdf(I))}{\log N}$$

The algorithm repeats for each of the 3 equal horizontal sub-images (see Figure 2) and on the whole image. The PEF descriptor is denoted by the concatenation of PEF_X, PEF_Y, PEF_B the mean and variance of the 3 channels, thus we have *4 regions* × *5 features* × *3 channels* = *60 dimensions* that describe the image *I*.

Figure 2: The 3 regions of the image

2.3 Color moments

Color moments represent a method that can be used to differentiate images based on their features of color. The main idea behind color moments is the assumption that the distribution of color can be interpreted as a probability distribution, which can be characterized by a number of moments (mean, variance, etc.). Stricker and Orengo [11] used three central moments of an image's color distribution: mean, standard deviation and skewness. The same authors showed that traditional methods like color histograms are sensitive to minor modifications in illumination or affine transformations.

A color can be abstractly represented by using color models like RGB (Red, Green, and Blue) or HSV (Hue, Saturation and Value). Thus, each of the three dimensions of the chosen color model is characterized by three moments of a color distribution, resulting in a nine dimension vector which will describe the color distribution in a given image.

$$E_i = \frac{1}{N} \sum_{j=1}^{N} p_{ij},$$

E_i is the mean or the average color value in the image, p_{ij} is the value of the j^{th} pixel in the i^{th} dimension of the color model and *N* is the number of pixels in the image.

$$\sigma_i = \sqrt{\frac{1}{N} \sum_{j=1}^{N} (p_{ij} - E_i)^2},$$

σ_i is the standard deviation (the square root of the variance) of the distribution.

$$s_i = \sqrt[3]{\frac{1}{N}\sum_{j=1}^{N}(p_{ij} - E_i)^3},$$

s_i is the skewness of the distribution which is a measure of the degree of its asymmetry.

The similarity function d_{mom} can be used to adjust the weights (w_i) of each channel, because it makes sense that, for example, the hue of a color is more important than its intensity. The function is defined as the sum of the weighted differences between the moments of the two distributions:

$$d_{mom}(I_1, I_2) = \sum_{i=1}^{3}(w_i|E_i^1 - E_i^2| + w_i|\sigma_i^1 - \sigma_i^2| + w_i|s_i^1 - s_i^2|)$$

2.4 Using Flickr user tags

In some situations, the visual information is not enough to give a semantic interpretation of an image and this is why we exploit user defined tags to improve the judgment of the whole system. The problems that arise with these approaches are the fact the number of user defined tags is relatively small (or 0), the tag can be in any language, some of them are irrelevant or they are a concatenation of words (see Figure 3). These problems make the traditional methods used in the field of natural language processing inapplicable in this situation.

Figure 3: Flickr user tags: oldbook, rarebook, latin, greek, library, bornin1550, deadlanguage, libro

The authors in [12] propose a linear SVM kernel that uses the most frequent user tags from the training set, which proved to be a good method. The idea is to construct a dictionary with user tags that appear at least *k* times (in our system we used *k = 16*) in the associated images from the training set. This process eliminates irrelevant and rare user tags and limits the dictionary to a number of *n* tags. Prior to the construction of the dictionary, we used Bing Translator[6] on every associated user tag, in order to attempt translation in English and a stemming algorithm that will reduce inflected or

[6] Bing Translator: http://www.bing.com/translator/

derived words to their root. After the dictionary is computed, an *n*-dimensional binary vector will be assigned to each image, with the i^{th} component 1 if the image is annotated with the ith user tag from the dictionary and 0, otherwise. The linear SVM kernel that classifies these vectors is:

$$K_G(t_i, t_j) = t_i{}^T t_j$$

$t_i{}^T t_j$ is the dot product between the transposed binary vector t_i and the t_j vector. The *KG* kernel counts the number of shared user tags between two associated images.

3 Classification

3.1 Classification using SVMs

Support vector machines [13, 14] proved to be one of the best classification technique used to address image classification problems as it can be very flexible and work with large amounts of data. Because this task requires multi-label classification (an image can be annotated with more than one concept), we choose to train an SVM classifier for each of the 94 concepts proposed by the ImageCLEF organizers [15] (to train a classifier for a concept *c*, we choose as positive examples the images that are annotated with the *c* concept and as negative examples the rest of the training images). Also, because of the highly unbalanced classification problem (the positive examples are usually less than the negative examples), we implemented a sampling method [16].

We propose a combined SVM kernel that makes use of all the features described above:

$$K_{combined}(x, y) = c_{ts}K_{ts}(x, y) + c_{pef}K_{pef}(x, y) + c_{ut}K_{ut}(x, y) + c_{cm}K_{cm}(x, y)$$

Where $c_{ts}, c_{pef}, c_{ut}, c_{cm} \in [0, 1]$, (such that $c_{ts} + c_{pef} + c_{ut} + c_{cm} = 1$) are weights for the following kernel functions:

- $K_{ts}(x, y) = simCos(d_{ts}(x), d_{ts}(y))$ is the cosine similarity defined in section 2.1 for the TopSurf library;
- $K_{pef}(x, y) = \exp(-\gamma||x - y||^2)$ is the RBF kernel and it is used with PEF descriptors (section 2.2);
- $K_{ut}(x, y) = \frac{t(x)^T t(y)}{n}$ is the linear kernel defined in section 2.4 normalized by the number of tags in the dictionary;
- $K_{cm}(x, y) = \exp(-\gamma\, d_{mom}(x, y))$ is the kernel that uses d_{mom} function for color moments (section 2.3) and γ is the regularization parameter.

These functions and $K_{combined}$ kernel satisfy Mercer's theorem [13] necessary to ensure SVMs convergence.

3.2 Post processing

In the post processing module of our system we ensure that the classifications made by SVMs models are correct. For example, if an image is classified with *quality_noblur* and *quality_partialblur* at the same time, we adjust the concept's probabilities so they sum up to 1. We learn about mutual exclusive concepts (CEx) from the training set. Let S_P be the set of predictions made by SVMs, with $c_1, c_2 \in CEx$ and $c_1, c_2 \in S_P$:

$$S_P = S_P \setminus \{c_1 : (c_1, c_2) \in CEx, p(c_1) < p(c_2)\}$$

We also compute the Voila – Jones face detection algorithm [17], in order to count the number of persons in a given image (the concepts regarding the number of persons in this year's competition are: *quantity_none, quantity_one, quantity_two, quantity_three, quantity_smallgroup, quantity_largegroup*) and to determine if *view_portrait* concept is present.

4 Submitted runs and results

Our system (Figure 4) has a modular and flexible structure and can easily be extended with some other feature extractors' algorithms:

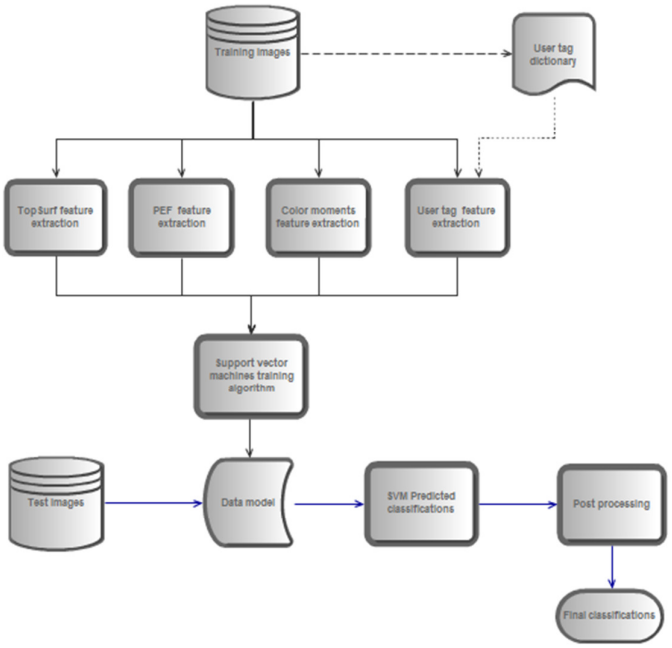

Figure 4: UAIC system

We participated at this year ImageCLEF 2012, Photo Annotation Task by submitting 5 runs with different configurations:

- *Submission1*: Visual only configuration with the following parameters: $c_{ts} = 0.6$, $c_{pef} = 0.2$, $c_{ut} = 0.0$, $c_{cm} = 0.2$ and SVM's regularization parameter: $C = 20$ with post processing step;
- *Submission2:* Multimodal run with the parameters: $c_{ts} = 0.45$, $c_{pef} = 0.1$, $c_{ut} = 0.35$, $c_{cm} = 0.1$ and SVM's regularization parameter C was chosen separately for each of the 94 classifiers, with sampling for some of the concepts;
- *Submission3:* The same configuration as for *Submission2*, with the sampling strategy applied for each of the 94 classifiers;
- *Submission4:* Multimodal run with the parameters: $c_{ts} = 0.35$, $c_{pef} = 0.25$, $c_{ut} = 0.25$, $c_{cm} = 0.15$ and SVM's regularization parameter C was chosen separately for each of the 94 classifiers, with the sampling strategy applied for each of the 94 classifiers;
- *Submission5:* Multimodal run with the parameters: $c_{ts} = 0.45$, $c_{pef} = 0.1$, $c_{ut} = 0.35$, $c_{cm} = 0.1$ with SVM's regularization parameter $C = 20$ and without the post processing step.

Table 1: Results of our submitted runs

	#Run	MiAP	GMiAP	F-ex	Features
1	1340348352281__submision1	0.2359	0.1685	0.4359	Visual
2	1340348434346__submision2	0.1863	0.1245	0.4354	Multimodal
3	1340348489605__submision3	0.1521	0.1017	0.4144	Multimodal
4	1340348583288__submision4	0.1504	0.1063	0.4206	Multimodal
5	1340348681456__submision5	0.1482	0.1000	0.4143	Multimodal

Our best run was the one with *Submission1* configuration and it was ranked 11th of a total of 18 group participants [1]. The fact that our visual-only run achieved the best of our scores shows that local invariant visual features are more appropriate for this task than other type of features. Also, we noticed that using user tags for classifying some of the concepts is, in fact, misleading. For example, for concept *weather_cloudysky* the most frequent tags were: *blue, Cannon, Nikon, clouds.*

Table 2: Results of participants in Photo Annotation task at ImageCLEF 2012

	Group name	MiAP	GMiAP	F-ex	Features
1	DBRIS	0.0925	0.0445	0.9980	Visual
2	LIRIS ECL	0.4367	0.3877	0.5766	Multimodal
3	DMS, MTA SZTAKI	0.4258	0.3676	0.5731	Multimodal
4	National Institute of Informatics	0.3265	0.2650	0.5600	Visual
5	ISI	0.4131	0.3580	0.5597	Multimodal
6	CEA LIST	0.4159	0.3615	0.5404	Multimodal
7	MLKD	0.3118	0.2516	0.5285	Multimodal
8	Multimedia Group of the Informatics and Telematics Institute Centre for Research and Technology Hellas	0.3012	0.2286	0.4950	Multimodal
9	Feiyan	0.2368	0.1825	0.4685	Textual
10	KIDS NUTN	0.1717	0.0984	0.4406	Multimodal
11	**UAIC2012**	**0.2359**	**0.1685**	**0.4359**	**Visual**
12	NPDILIP6	0.3356	0.2688	0.4228	Visual
13	IntermidiaLab	0.1521	0.0894	0.3532	Textual
14	URJCyUNED	0.0622	0.0254	0.3527	Textual
15	Pattern Recognition and Applications Group	0.0857	0.0417	0.3331	Visual
16	Microsoft Advanced Technology Labs Cairo	0.2086	0.1534	0.2635	Textual
17	BUAA AUDR	0.1307	0.0558	0.2592	Multimodal
18	UNED	0.0873	0.0441	0.1360	Visual

All participants at ImageCLEF 2012 in Photo Annotation task have submitted several runs using not only visual strategies based on features extracted from the images but as well textual ones based on user defined tags that were given alongside the images. The best results however, as it can also be observed from the table above, were achieved by the systems that managed to combine both the visual and textual features together. What our system lacked was the fact that we did not find the best balance between feature extraction algorithms (with their contribution in the learning step) and also the fact that some of them should weight more or less depending on the concept that is being learned.

5 Conclusions

In this paper we combined several different state of the art algorithms for image processing together with Support Vector Machines and kernel functions in order to approach the task of automated image annotation. As images can be annotated with more than one concept we tried to increase our system's performance by using not only local image feature descriptors (TopSurf), that for example, proved to be unpractical at detecting feelings in an image, but also try analyzing the colors (Color Moments) and the textures (Profile Entropy Features) in the image and even make use of the user defined tag semantics and face detection algorithms.

All experiments were made using the approach we presented in this paper and careful attention was given to the selection of the threshold parameters of the SVM kernel function that we used, $K_{combined}$, c_{ts}, c_{pef}, c_{ut} and c_{cm}.

As future work, we will try and set different values for these parameters taking into consideration the concept that the classifier is training for. For example, for concepts that express feelings, Color Moments technique should have the deciding weight, whereas for panoramic images a greater weight should be given to the texture descriptor (PEF).

Acknowledgement. The research presented in this paper was funded by the Sector Operational Program for Human Resources Development through the project "Development of the innovation capacity and increasing of the research impact through post-doctoral programs" POSDRU/89/1.5/S/49944.

References

1. Thomee, B., Popescu, A.: Overview of the ImageCLEF 2012 Flickr Photo Annotation and Retrieval Task. CLEF 2012 working notes, Rome, Italy. (2012)
2. Iftene, A., Vamanu, L., Croitoru, C.: UAIC at ImageCLEF 2009 Photo Annotation Task. In C. Peters et al. (Eds.): CLEF 2009, LNCS 6242, Part II (Multilingual Information Access Evaluation Vol. II Multimedia Experiments). Pp. 283-286. Springer, Heidelberg. (2010)
3. Thomee, B., Bakker, E. M., Lew, M. S.: TOP-SURF: a visual words toolkit. In Proceedings of the 18th ACM International Conference on Multimedia, pp. 1473-1476, Firenze, Italy. (2010)
4. Evans, C.: Notes on the OpenSURF Library. CSTR-09-001, University of Bristol. January 2009. (2009)
5. Bay, H., Ess, A., Tuytelaars, T., Van Gool, L.: Speeded-up robust features (SURF). Computer Vision and Image Understanding. Computer Vision and Image Understanding (CVIU), Vol. 110, No. 3, pp. 346—359. (2008)
6. Csurka, G., Dance, C., Fan, L., Willamowski, J., Bray, C.: Visual Categorization with Bags of Keypoints. In Workshop on Statistical Learning in Computer Vision, ECCV. (2004)
7. van Gemert, J., Snoek, C., Veenman, C., Smeulders, A., Geusebroek, J. M.: Comparing Compact Codebooks for Visual Categorization. Computer Vision and Image Understanding, Vol. 14, Issue 4, Pp. 450-462. (2008)
8. Salton, G., McGill, M.: Introduction to modern information retrieval. McGraw-Hill. (1983)
9. Glotin, H., Zhao, Z., Ayache, S.: Efficient image concept indexing by harmonic and arithmetic profiles entropy. In: Proceedings of 2009 IEEE International Conference on Image Processing (ICIP 2009), Cairo, Egypt, November 7-11, 14. (2009)
10. Moddemeijer, R.: On estimation of entropy and mutual information of continuous distributions. Signal Processing, vol. 16, no. 3, pp. 233–246, March 1989. (1989)
11. Stricker, M., Orengo, M.: Similarity of color images. In Storage and Retrieval for Image and Video Databases, Proc. SPIE 2420, pp. 381-392. (1995)
12. Guillaumin, M., Verbeek, J., Schmid, C.: Multimodal semi-supervised learning for image classification. IEEE Conference on Computer Vision & Pattern Recognition. pp. 902–909. Grenoble, France. (2010)
13. Cristianini, N., J. Shawe-Taylor.: An Introduction to Support Vector Machines and other kernel-based learning methods. Cambridge University Press, Cambridge, UK. (2000)
14. Tong, S., Chang, E.: Support vector machine active learning for image retrieval. In: Proc. ACM Multimedia, Ottawa, Canada. (2001)

15. Gidudu, A., Hulley, G., Marwala, T.: Image Classification Using SVMs: One-against-One Vs. One-against-All. In Proceeding of the 28th Asian Conference on Remote Sensing, Malaysia, CD-Rom. (2007)
16. Witten, I., Frank, E., Hall, M.: Data Mining – Practical Machine Learning Tools and Techniques, Third Edition. Morgan Kaufmann, 629 pages. (2011)
17. Viola, P., Jones, M.: Rapid object detection using boosted cascade of simple features. Proceedings of the Conference on Computer Vision and Pattern Recognition (CVPR), vol. 1, pp. 511–518, Kauai, Hawaii, USA. (2001)