DUTh at the ImageCLEF 2016 Image Annotation Task: Content Selection

Georgios Barlas, Maria Ntonti, Avi Arampatzis

Department of Electrical and Computer Engineering
Democritus University of Thrace, Xanthi 67 100, Greece
{gbarlas,mntonti,avi}@ee.duth.gr

Abstract. This report describes our experiments in the Content Selection subtask of the Image Annotation task of ImageClef 2016[7, 13]. Our approach is based on the fact that the human visual system concentrates mostly on local features [12]. In this respect, we trained an SVM classifier with descriptors that are based on the local features of the image, such as edges and corners. For the experimentation process we used the set of 500 images provided for the task, divided into training and test set. This set was particularly created for this year's new subtask, Content Selection, although the concepts are the same as last year. Through experimentation we determine which descriptors give the best results for the given task. To conduct the main experiment the SVM classifier is trained with the aforementioned set of 500 images using a subset of the top-performing features. Consecutively, the SVM processes the new set of 450 images and selects the boxes that best describe them conceptually.

1 Introduction

Content Selection is the intermediate step between identifying objects of an image and generating a natural language caption. The objective of this task is to identify which bounded objects of the image are important so as to be included in the annotator's caption. In addition, since each object is labeled, the result would be a conceptually accurate description of the image. As requested by the ImageCLEF competition we developed a system that receives as input labeled objects of an image and identifies the objects that are referred to in the corresponding image caption. Since participants of the content selection subtask have concentrated so far on using only text features or bounding box information, this paper may provide a novel contribution in exploring features based on visual keypoints only. We are not aware of another work with keypoints being used in our ways (e.g. ratio of keypoints in bounding boxes to keypoints in image).

Our approach relies on finding suitable descriptors, from the given data, in order to train an SVM classifier. In this work we followed an image-only approach without processing the textual information provided as groundtruth. Inspired by the fact that the human visual system concentrates mostly on local features, we incorporated several such descriptors using well known algorithms of image processing. Thereby we created a set of 17 descriptors and grouped

them into several subsets referring to similar features. Then, after experimenting with descriptors standalone and as groups, we arrived to a set of 9 most-useful descriptors.

Complementary to testing for different feature subsets, experiments were also conducted using various SVM kernel functions. Linear, polynomial and Gaussian radial basis function kernels were tested. Polynomial and Gaussian kernels performed similarly, but the latter one was chosen as it produced slightly better results.

The rest of this report is organized as follows. In the next section we describe the dataset provided as well as the methodology we followed to tackle the problem. Specifically, we analyze the descriptors that were used during experimentation and describe the SVM classifier. Section 3 describes the evaluation methodology along with our the experimental results. Conclusion and directions for further research are summarized in Section 4.

2 Data and Methods

2.1 Dataset

The dataset provided by the ImageCLEF competition consists of 500 images from various concepts, accompanied with a set of bounding boxes for each image and a label for each box. Furthermore, a set of textual descriptions is given for each image as ground-truth with a minimum of 5 and a mean of 9.5 sentences per image [7]. This set was initially split into a training and a validation set and was used for experimentation. After experimentally concluding to the best configuration the whole set was used to train the SVM classifier. A second set of 450 images was later released by ImageCLEF which was used as the test set.

2.2 Feature Extraction

Initially, 17 descriptors were created. After experimentation we concluded to a subset of 9 that was found to have the maximum contribution. Correlation information between the descriptors and the purpose of each one was used to cluster them into categories. In the rest of this section, we will elaborate on each descriptor category.

Position Desctiptors For each bounding box, two points are given that define its position in the image. The coordinates of these points are used individually as four values $x_{min}, x_{max}, y_{min}, y_{max}$, divided by the corresponding dimension of the image, as follows:

$$d_1 = \frac{x_{min}}{w}$$
$$d_2 = \frac{x_{max}}{w}$$
$$d_3 = \frac{y_{min}}{h}$$
$$d_4 = \frac{y_{max}}{h}$$

where w, h denote the width and the height of the image, respectively. Additionally, two more features are formed that correspond to the relative position of the center of the bounding box, abscissa and ordinate respectively:

$$d_5 = \frac{d_1 + d_2}{2}$$
$$d_6 = \frac{d_3 + d_4}{2}$$

The purpose of those descriptors is to investigate the correlation between the position of the bounding box and the importance of it [6].

Size Descriptors The descriptors of this group aim to calculate the portion of the image that is occupied by the bounding box, separately for the two dimensions and combined as well.

$$d_7 = \frac{w_b}{w}$$
$$d_8 = \frac{h_b}{h}$$
$$d_9 = d_7 d_8$$

where w_b, h_b denote the width and height of the bounding box, respectively.

Descriptors based on Local Features For the calculation of the descriptors of this group we used several well-established algorithms for local feature detection (i.e. the Canny edge detector [2], Harris and Stephens [8], BRISK [9], SURF [1] and FAST [11]). These algorithms imitate the way that human processes visual information. The calculated features are based on the hypothesis that an elevated number of the key-points or key-regions detected will be located in conceptually important boxes. Towards this direction we propose the calculation of the percentage of the image local features detected in each box.

$$d_{10} = \frac{Canny_{box}}{Canny_{image}}$$
$$d_{11} = \frac{Harris_{box}}{Harris_{image}}$$
$$d_{12} = \frac{BRISK_{box}}{BRISK_{image}}$$
$$d_{13} = \frac{SURF_{box}}{SURF_{image}}$$
$$d_{14} = \frac{FAST_{box}}{FAST_{image}}$$

Entropy Descriptors Image entropy is a quantity that is used to describe the amount of information that is contained in an image. In this regard, the motivation behind this feature group is to quantify the amount of information held by the content of a bounding box, proportionally to that of the image. As a first step we produced three different versions of the image. The first two correspond to the edge maps generated by the Canny edge detector [2] and the Structured Forests edge detection method (ESF) [5], respectively. The last one corresponds to the image reduced to gray-level pixel values. Consecutively, we calculate the entropy contained by a bounding box in all three images divided by the total entropy of the image.

$$d_{15} = \frac{E(Canny_{box})}{E(Canny_{image})}$$
$$d_{16} = \frac{E(ESF_{box})}{E_(ESF_{image})}$$
$$d_{17} = \frac{E(Grayscale_{box})}{E(Grayscale_{image})}$$

where $E(x)$ denotes the entropy of an image x, defined as

$$E(x) = -\sum_{i=1}^{N} h(i) \log_2 h(i)$$

where $h(i)$ is the count of pixels assigned to the ith bin of the image histogram and N the total number of bins.

2.3 Support Vector Machine (SVM) Classification

For the purpose of this task, we trained a binary SVM in order to classify each bounding box as 'important' or 'not important'. SVM tackles the problem of non-liner classification by determining a hyperplane that separates two classes in a space of higher dimensionality than the feature space using a kernel function [3,

4]. We used MATLAB's default function to train and use the SVM. For improved performance, we experimented with different kernel functions, such as the Linear, the Polynomial and the Gaussian radial basis kernel functions, concluding to the Gaussian kernel as the top performing.

During the experiments, the training dataset was divided into two subsets, the training and the validation subset, respectively. We experimented with training subsets of 100 and 250 images randomly selected. We investigated the performance of different sets of descriptors, concluding to a set of 9. Table 1 presents the results of the experiments with different configurations.

In the cases where the SVM classified a small number of the boxes as important, we used the SURF descriptor as criterion for the selection. Specifically, if the result contained less than two boxes, then the $\frac{n}{2} + 3$ boxes with the biggest d_{13} value were added to result set, where n is the number of boxes in the image. In case the image contained less than $\frac{n}{2} + 3$ boxes, then all of them were selected.

The number $\frac{n}{2} + 3$ was selected after experimentation. We initially started with number $\frac{n}{2}$ and then concluded to $\frac{n}{2} + 3$.

3 Experimental Evaluation

3.1 Evaluation Measures

According to the instructions, Subtask 3 is evaluated using the content selection metric, which is the F_1 score averaged across all test images. Each F_1 score is computed from the precision and recall metrics averaged over all gold standard descriptions for the image.

The precision P^{I_i} for test image I_i is computed as:

$$P^{I_i} = \frac{1}{M} \sum_{m=1}^{M} \frac{|G_m^{I_i} \cap S^{I_i}|}{|S^{I_i}|} \tag{1}$$

The recall $R^(I_i)$ for test image I_i is computed as:

$$R^{I_i} = \frac{1}{M} \sum_{m=1}^{M} \frac{|G_m^{I_i} \cup S^{I_i}|}{|G_m^{I_i}|} \tag{2}$$

where

$I = \{I_1, I_2, \ldots, I_N\}$ the set of test images
$G^{I_i} = \{G_1^{I_i}, G_2^{I_i}, \ldots, G_M^{I_i}\}$ the set of gold standard descriptions
S^{I_i} the resulting set of unique bounding box instances
M the number of gold standard descriptions for image I_i.

The content selection score F^{I_i} for image I_i, is computed as:

$$F^{I_i} = 2\frac{P^{I_i} R^{I_i}}{P^{I_i} + R^{I_i}} \tag{3}$$

The final P, R and F scores are computed as the mean P, R and F scores across all test images.

3.2 Experimental Results

For the experiments we used the 500 images from imageCLEF dataset. Firstly, the images are split randomly in two sets, the training set and the test set. As training set, 100 or 250 images are used, 20% of 50% of the set respectively. As expected, the bigger training dataset gave better results. for this reason, it was decided to use all the 500 provided images dataset as training set at the submission run. Experiments took place with various combination of descriptors. As shown in Table 2, descriptors are managed as groups. For decision criterion in the cases that the SVM classified a small number of boxes as important, we experimented with d_9 and d_{13} descriptors. The SURF descriptor (d_{13}) produced better results, as it is a well-established and robust algorithm, in contrary to descriptor d_9 which is more abstract. Table 2 shows the setup of each experiment. For example, for experiment 1 descriptors $d_{1...4}$, d_{11}, d_{13} and d_{17} were used. SVMs kernel is radial basis function, 100 images of 500 were used as training set and as criterion d_9 were used. Before the experiments each group of descriptors were tested separately, so the behavior of each was known. That means that there was not need to include or exclumide all descriptors of the group in each experiment but the best representatives of the group. Furthermore, the correlation matrix was taken into account, that is why for example d_9 is not included in our experiments, as shown in Table 2. Finally, as seen from Table 1, the F-measure was not the only one taken under consideration for our final choice but also the balance of precision and recall.

Our approach using Gaussian kernel SVM classifier and 9 descriptors as experiment 14 achieves an overall F-measure of $54.59\% \pm 15.33$, the best result of the two participating groups to the subtask.

Table 1. Experimental results for various combinations of descriptor

Experiment	F-score	Precession	Recall
1	0.5885 ± 0.2300	0.5972 ± 0.2451	0.6758 ± 0.2963
2	0.5873 ± 0.2297	0.5946 ± 0.2442	0.6783 ± 0.2946
3	0.5882 ± 0.2312	0.5991 ± 0.2465	0.6726 ± 0.2982
4	0.5888 ± 0.2294	0.5963 ± 0.2446	0.6761 ± 0.2978
5	0.5980 ± 0.2224	0.5994 ± 0.2390	0.6892 ± 0.2955
6	0.5976 ± 0.2235	0.5982 ± 0.2392	0.6893 ± 0.2962
7	0.5891 ± 0.2296	0.5969 ± 0.2446	0.6758 ± 0.2986
8	0.5889 ± 0.2301	0.5985 ± 0.2449	0.6746 ± 0.2973
9	0.5585 ± 0.2108	0.4939 ± 0.2081	0.7461 ± 0.2994
10	0.6380 ± 0.1899	0.5683 ± 0.2210	0.8203 ± 0.2253
11	0.5314 ± 0.1602	0.3865 ± 0.1620	0.9671 ± 0.0761
12	0.5709 ± 0.2175	0.5298 ± 0.2257	0.7235 ± 0.2850
13	0.5894 ± 0.2283	0.5950 ± 0.2430	0.6817 ± 0.2930
14	0.5888 ± 0.2322	0.6018 ± 0.2475	0.6663 ± 0.3042
15	0.5893 ± 0.2303	0.6001 ± 0.2457	0.6717 ± 0.3000
16	0.5893 ± 0.2312	0.6014 ± 0.2462	0.6688 ± 0.3030
17	0.5888 ± 0.2300	0.5973 ± 0.2450	0.6757 ± 0.2975

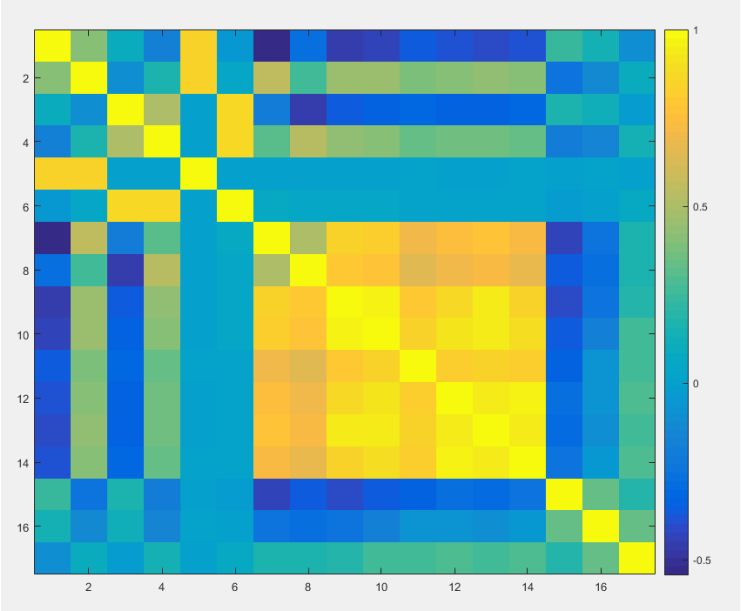

Fig. 1. Correlation calculation between every descriptor investigated

4 Conclusion

This report describes the methodology, the experimentation, and the results acquired concerning DUTh's participation to the Subtask 3 of the Image Annotation task of ImageCLEF 2016. Our novelty is that we have tackled the task using only visual keypoints/features, in contrast to other participants so far using only text features or bounding box information. We investigated the performance of seventeen image descriptors combined with three SVM configurations corresponding to different SVM kernels. Experimental evaluation highlighted the significance of a feature subset in determining the conceptually important boxes. These features mostly relate to the edges and corners detected by well-established algorithms. Our approach using Gaussian kernel SVM classifier and nine descriptors achieves an overall F-measure of $54.59\% \pm 15.33$, the best result of the two participating groups to the subtask.

Taking a step further, we believe that the proposed methodology can be improved towards two directions. The first one concerns the improvement of the feature extraction methods. Motivated by the high performance of local feature detection demonstrated by this project, we would like to additionally incorporate the local feature descriptors that correspond to them. The statistical analysis and the comparison of these descriptors may provide useful information concerning the importance of each key-point. Room for improvement also exists in the exploitation of textual analysis of the proposed annotation terms. Textual features that are based on term and document frequencies can provide useful

Table 2. Experimental results for various combinations of descriptor

#	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17
d_1	•		•		•		•							•		•	•
d_2	•		•		•		•							•		•	•
d_3	•		•		•		•							•		•	•
d_4	•		•		•		•							•		•	•
d_5							•	•								•	•
d_6							•	•								•	•
d_7	•	•	•	•	•	•	•	•	•					•	•	•	•
d_8	•	•	•	•	•	•	•	•		•				•	•	•	•
d_9																•	•
d_{10}																•	•
d_{11}	•	•	•	•	•	•	•	•				•		•	•	•	•
d_{12}																•	•
d_{13}	•	•	•	•	•	•	•	•					•	•	•	•	•
d_{14}																	
d_{15}																•	•
d_{16}								•								•	•
d_{17}	•	•	•	•	•	•	•	•			•			•	•	•	•
Kernel	rbf	rbf	poly	poly	rbf	rbf	rbf	rbf	rbf	rbf	rbf	rbf	rbf	rbf	rbf	rbf	rbf
Training set	100	100	100	100	250	250	250	250	100	100	100	100	100	100	100	100	100
Criterion	d_9	d_9	d_9	d_9	d_9	d_9	d_9	d_9	d_9	d_9	d_9	d_9	d_9	d_{13}	d_{13}	d_9	d_{13}

insight in order to determine the importance of every box label. Furthermore, textual features concerning term significance may be extracted exploiting word ontologies or semantic networks, such as WordNet [10].

References

1. Bay, H., Ess, A., Tuytelaars, T., Van Gool, L.: Speeded-up robust features (surf). Computer vision and image understanding 110(3), 346–359 (2008)
2. Canny, J.: A computational approach to edge detection. Pattern Analysis and Machine Intelligence, IEEE Transactions on (6), 679–698 (1986)
3. Cortes, C., Vapnik, V.: Support-vector networks. Machine learning 20(3), 273–297 (1995)
4. Cristianini, N., Shawe-Taylor, J.: An introduction to support vector machines and other kernel-based learning methods. Cambridge university press (2000)
5. Dollár, P., Zitnick, C.L.: Structured forests for fast edge detection. In: ICCV. International Conference on Computer Vision (December 2013), `http://research.microsoft.com/apps/pubs/default.aspx?id=202540`
6. Friedman, J., Hastie, T., Tibshirani, R.: The elements of statistical learning, vol. 1. Springer series in statistics Springer, Berlin (2001)
7. Gilbert, A., Piras, L., Wang, J., Yan, F., Ramisa, A., Dellandrea, E., Gaizauskas, R., Villegas, M., Mikolajczyk, K.: Overview of the ImageCLEF 2016 Scalable Concept Image Annotation Task. In: CLEF2016 Working Notes. CEUR Workshop Proceedings, CEUR-WS.org, Évora, Portugal (September 5-8 2016)
8. Harris, C., Stephens, M.: A combined corner and edge detector. In: Alvey vision conference. vol. 15, p. 50. Citeseer (1988)
9. Leutenegger, S., Chli, M., Siegwart, R.Y.: Brisk: Binary robust invariant scalable keypoints. In: Computer Vision (ICCV), 2011 IEEE International Conference on. pp. 2548–2555. IEEE (2011)
10. Miller, G.A.: Wordnet: a lexical database for english. Communications of the ACM 38(11), 39–41 (1995)

11. Rosten, E., Porter, R., Drummond, T.: Faster and better: A machine learning approach to corner detection. Pattern Analysis and Machine Intelligence, IEEE Transactions on 32(1), 105–119 (2010)
12. Shapley, R., Tolhurst, D.: Edge detectors in human vision. The Journal of physiology 229(1), 165 (1973)
13. Villegas, M., Müller, H., García Seco de Herrera, A., Schaer, R., Bromuri, S., Gilbert, A., Piras, L., Wang, J., Yan, F., Ramisa, A., Dellandrea, E., Gaizauskas, R., Puigcerver, K.M.J., Toselli, A.H., Snchez, J.A., Vidal, E.: General Overview of ImageCLEF at the CLEF 2016 Labs. Lecture Notes in Computer Science, Springer International Publishing (2016)