TOWARDS FULLY UN-SUPERVISED METHODS FOR GENERATING OBJECT DETECTION CLASSIFIERS USING SOCIAL DATA

Spiros Nikolopoulos, Elisavet Chatzilari, Eirini Giannakidou and Ioannis Kompatsiaris

Informatics and Telematics Institute, ITI - CERTH, 1st Km Thermi-Panorama Road, GR-57001, Greece
email: {nikolopo@iti.gr, echatzil@auth.gr, igiannak@iti.gr, ikom@iti.gr}

ABSTRACT

In this work a framework for constructing object detection classifiers using weakly annotated social data is proposed. Social information is combined with computer vision techniques to automatically obtain a set of images annotated at region-detail. All assumptions made to automate the proposed framework are driven by the reasonable expectation that due to the collaborative aspect of social data, linguistic descriptions and visual representations will start to converge on common concepts, as the scale of the analyzed dataset increases. Comparison tests performed against manually trained object detectors showed that comparable performance can be achieved.

1. INTRODUCTION

Among the various approaches that have been proposed to facilitate multimedia consumption, the ones that try to exploit the local image characteristics have attracted considerable attention. Advances in this direction have led to the employment of machine learning techniques where classifiers are trained to recognize an object using indicative instances of its visual representation. Although these techniques are known to perform well, their major drawback is that they require a large amount of region-detail image annotations, which is very expensive to obtain. On the other hand, in the context of Web 2.0, collaborative systems like Flickr, accommodate databases that are being populated with hundreds of user tagged images. This work attempts to reduce the effort required for generating region-detail image annotations by mining the necessary information from a large social corpus. Specifically, region-detail image annotations are derived from weakly annotated data for training robust object detectors. The contribution of our work concentrates on combining methods that exploit social information, with computer vision techniques.

The idea of learning object categories from weakly annotated images has attracted considerable interest in recent literature. In [1], a statistical model that integrates semantic information provided by text and images is used for organizing image collections. [2] presents an approach that learns object categories by utilizing the raw output of image search engines, while [3] performs top-down image segmentation with weak supervision. In [4] the use of Markov Filed Aspect models for performing region classification is investigate. [5] introduces a model for object recognition as machine translation, while [6] performs automatic linguistic indexing of pictures. Although these works are trying to tackle the object detection problem using weak annotations, few are the attempts that exploit the social aspect of user contributed data.

2. FRAMEWORK DESCRIPTION

A weakly annotated image is an image I_q associated with a set of tags than even if they have been contributed to describe an object of the image, there is no spatial information locating this object into the image. The goal of our framework is to identify this type of relations for the purpose of training a classifier detecting an object (from here on called target concept).

The framework's architecture, depicted in Fig.1, is actually a pipeline where the output of one module constitutes the input of the next. Specifically, the framework receives as input a set of weakly annotated images and performs the following operations: a) social and semantic-based clustering of image tags for acquiring groups of images with an increased level of semantic coherence, b) segmenting all images in a group for identifying regions that are likely to represent meaningful objects, c) extracting the visual features of these regions d) performing feature-based clustering of the identified regions, e) labeling of each cluster based on the tags associated with the original images and f) utilizing the visual features extracted from the regions belonging in a specific cluster to train an object recognition model.

Although, there are some open issues in the aforementioned approach such as a) how to select among the groups of images generated by the tag-based image clustering, the one depicting the target concept, b) how to decide the number of clusters for the feature-based region clustering or c) how to pick the cluster containing the regions depicting the target concept, our principal assumption is that since the scale of the analyzed dataset can grow arbitrary big, both linguistic descriptions and visual representations are expected to con-

verge towards on common concepts. Based on this assumption and in order to fully automate the aforementioned process, we adopt the following solutions.

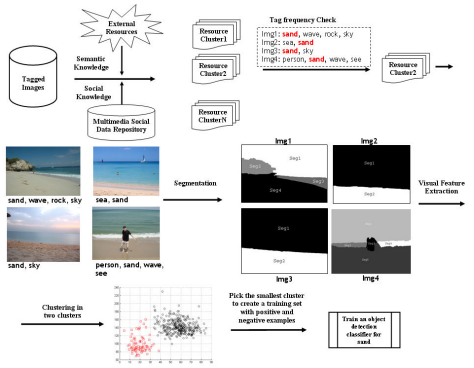


Fig. 1. Visualization of the proposed framework

Mining on Multimedia Social Data Sources: For acquiring the necessary groups of images we adopted the SEMSOC approach [7]. In this work an unsupervised model for efficient and scalable mining of multimedia social-related data is presented. The reason for adopting this approach is to overcome the limitations that characterize collaborative tagging systems such as tag spamming, tag ambiguity, tag synonymy and granularity variation. SEMSOC jointly considers social and semantic features to cluster multimedia sources and obtain meaningful groups, each corresponding to a particular topic. Let I denote the entire set of images acquired from the social tagging environment. The outcome of SEMSOC is a set of image groups $I^{L_i} \subset I, \quad i = 1, \ldots, m$, where L_i is an indicator of the group topic and m the number of groups. L_i is chosen so as the most frequent tag of all images in I^{L_i} to conceptually relate to the target concept (Lb). In this way, we obtain a semantically coherent group of images I^{Lb} the majority of which is expected to depict the target concept.

Segmentation & Visual Descriptors: Segmentation is applied $\forall I \in I^{Lb}$ in order to identify visually coherent regions. An algorithm based on k-means with connectivity constraint [8] was used in our work. The generated output is a spatial mask $S = \{s_i, \quad i = 1, \ldots, N\}$, with s_i representing the region of an identified object. Six descriptors proposed by MPEG-7 [9] namely ColorLayout, ColorStructure, ScalableColor, EdgeHistogram, HomogeneousTexture and RegionShape, capturing different elements of color, texture and shape were investigated. Various combinations of these descriptor were tested by concatenating their normalized values on a single vector. Thus, $\forall I_q \in I^{Lb} \quad \& \quad \forall s_i^{I_q} \in S_{I_q}$ we extract a vector $f(s_i^{I_q}) = \{f_1, f_2, \ldots, f_M\}$ where M determines the dimensionality of the feature space.

Feature-based Regions Clustering & Labeling: Working under the assumption that all regions representing the same concept will have a relatively high amount of common visual characteristics, we apply the k-means clustering algorithm on all feature vectors $f(s_i^{I_q})$ extracted $\forall I_q \in I^{Lb} \quad \& \quad \forall s_i^{I_q} \in S_{I_q}$. At this point, two issues need to be resolved a) what will be the number of clusters utilized by k-means and b) which of the formulated clusters is the one containing the regions that we are interest in. The following analysis provides support to our claim that by splitting the set of regions in two clusters and choosing the one with the smallest population, we have more than 50% probability of selecting the appropriate cluster. Table 1 summarizes the necessary notations.

Table 1. Legend of Introduced Terms

let C_1 and C_2 be the 1st & the 2nd of the formulated clusters
let $s_i^{I_q}$ be an un-clustered region
let K be the total number of images in I^{Lb}
let u be the average number of regions identified $\forall I_q \in I^{Lb}$
let l be the set of regions depicting the target concept
let $C_1 < C_2$ be the fact that C_1 is smaller than C_2
let $l \in C_1$ be the fact that C_1 is the appropriate cluster
let $P1$ be the probability that clustering assigns a target concept region $s_i^{I_q} \in l$ in cluster C_1, when $l \in C_1$
let $P2$ be the probability that clustering assigns an irrelevant visual concept region $s_i^{I_q} \notin l$ in cluster C_2, when $l \in C_1$

Ideally, clustering will manage to assemble all regions representing the target concept in one cluster pushing all irrelevant regions to the other. $P1$ and $P2$ are actually an indication of how efficiently k-means manage to perform this task. Given these probabilities in order for the smallest cluster to have more than 50% probability of being the one containing the regions of interest the following inequality should hold:

$$P(C_1 < C_2 \mid l \in C_1) > 0.5 \tag{1}$$

We examine $PC1, PC2$, that express the probability of an un-clustered region $s_i^{I_q}$ to be assigned in C_1 or C_2 respectively, given that $l \in C_1$:

$$\begin{aligned} P(s_i^{I_q} \rhd C_1 \mid l \in C_1) &= \tfrac{\|l\|}{K \cdot u} P1 + \tfrac{K \cdot u - \|l\|}{K \cdot u}(1 - P2) \\ P(s_i^{I_q} \rhd C_2 \mid l \in C_1) &= \tfrac{\|l\|}{K \cdot u}(1 - P1) + \tfrac{K \cdot u - \|l\|}{K \cdot u} P2 \end{aligned} \tag{2}$$

Where $\|l\|$ is the cardinality of set l and the symbol $\rhd$ represents the assignment of an un-clustered region to a cluster. In order for equation (1) to be true one should expect that $PC2 - PC1 > 0$ and from (2) we derive:

$$\tfrac{\|l\|}{K \cdot u}(2 - 2P1 - 2P2) + 2P2 - 1 > 0 \tag{3}$$

As demonstrated in Section 3 where u, $P1$ and $P2$ are estimated experimentally, the probability of expression (3) being true is significantly high.

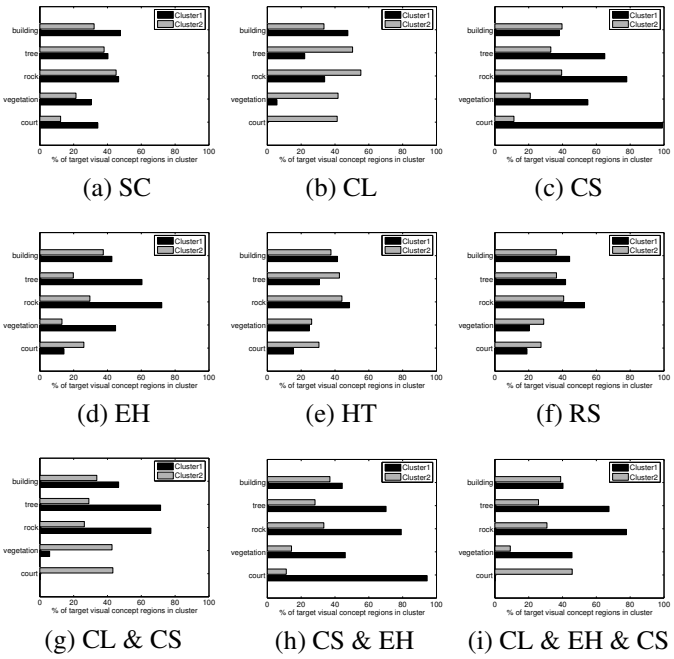
The last operation of the proposed framework is to train a classifier detecting the target concept. Support Vector Machines (SVMs) were chosen for engineering the object detection classifiers. The feature vectors of all regions belonging to the chosen cluster are used as positive examples for training.

3. EXPERIMENTAL EVALUATION

The goal of our study is to experimentally validate our theoretical analysis and check whether the performance of object detectors generated using the proposed framework is comparable to the performance achieved by manually trained object detectors. A lexicon of 25 concepts C_L was used to manually annotate 1800 images at region-detail and produce the fully annotated dataset I^M. On the other hand, 3000 images I^F were crawled from Flickr along with their corresponding tags, so as to depict *cityscape*, *seaside*, *mountain*, *roadside*, *landscape* and *sport-side* locations. For evaluation purposes and after applying the SEMSOC on I^F, we obtained five groups of tagged images having as most frequent tag a member of C_L. Specifically, *building*, *tree*, *rock*, *vegetation* and *court* comprise the set of target concepts C_T that was utilized for evaluation.

In order to validate our theoretical analysis, we examine the clustering output for the five concepts of C_T. Since I^M allows us to explicitly measure the efficiency of clustering, our aim was to have an experimental insight on the probabilities P1 and P2. Moreover, using I_M we had the opportunity to examine how often the smallest cluster is actually the one containing the regions of interest. In order to examine the clustering output we measure the percentage of regions representing the target concept in each of the formulated clusters and organize the bar diagrams in groups of bar couples demonstrating the calculated figures. For inspection purposes all presented bar diagrams follow the convention that the bar located at the lower position of each bar couple, is the one corresponding to the cluster with the smallest population. The first set of experiments depicted in Fig.2 measures the clustering efficiency for all concepts in C_T using each of the MPEG-7 descriptors. By examining Fig.2(a-f) it is clear that CS, EH and CL seems to discriminate the regions better than RS and SC. Furthermore in most of the cases, CS and EH manage to allocate the regions of interest in the smallest of the formulated clusters, as opposed to CL. Clustering was re-applied using different combinations of the most prominent descriptors. Since our intention is to formulate clusters with high percentage of regions representing the target concept and at the same time exhibiting the smallest population, it is clear from Fig.2(g-i) that the combination of CS with EH performs satisfactory for the majority of cases.

Using I_M we can experimentally estimate the probabilities $P1$ and $P2$ for all concepts in C_T. By averaging between all concepts, the values we obtain, using the combination of CS with EH, are $\bar{P1} = 0.508$ and $\bar{P2} = 0.83$. Additionally, since we aim at semantically important concepts, it is reasonable to assume that each image in I^{Lb} would contain at most one region depicting the target concept, thus $\|l\| \leq K$. By substituting the values of $\bar{P1}$ and $\bar{P2}$ in equation (3) we obtain the following inequality that should be valid in order for equation (1) to be satisfied:



(a) SC (b) CL (c) CS

(d) EH (e) HT (f) RS

(g) CL & CS (h) CS & EH (i) CL & EH & CS

Fig. 2. Clustering efficiency using the MPEG-7 descriptors

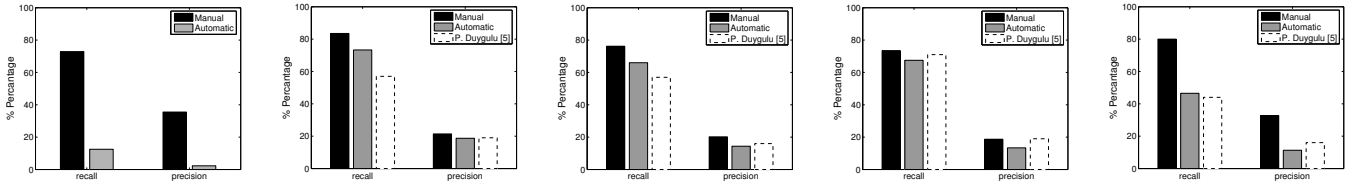
$$-0.676 * \frac{\|l\|}{K \cdot u} + 0.66 > 0 \Rightarrow \|l\| < 0.97 \cdot K \cdot u \qquad (4)$$

By calculating the average number of image regions in I^M to be 4, we obtain from equation (4) that in order for equation (1) to be true it suffices that $\|l\| < 3.8 \cdot K$. This a restriction very likely to be satisfied if we consider that intuitively we have accepted that $\|l\| \leq K$.

In the last experiment we utilize the concepts in C_T to compare the performance of object detectors generated using manual annotations (i.e., I^M), against the performance of the detectors trained using the proposed approach and I^F). The combination of CS with EH was utilized as the feature space. A portion of the manually annotated image set $I^M_{gd} \subset I^M$, not used during training, served as the ground truth for testing.

Fig.3 demonstrates that in some cases the object detectors generated using the proposed approach can perform almost as good as the ones trained manually. For the cases of *Vegetation*, *Rock* and *Tree*, both recall and precision can be considered comparable. This is not the case for *court* where the automatically trained detectors perform considerably poor. This is probably because the clustering algorithm fails to gather all interesting regions in the cluster with the smallest population and as a consequence the wrong set of regions is forwarded to the SVM training scheme. On the other hand, in the case of *building*, although the automatically trained detector is outperformed by the one trained manual, its performance can still be considered satisfactory.

For comparison purposes, we present (whenever possi-



(a) Court (b)Vegetation (c) Rock (d) Tree (e) Building

Fig. 3. Comparative bar diagrams for manually and automatically trained object detection classifiers

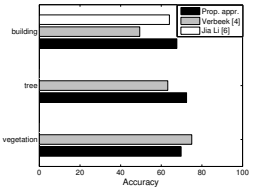


Fig. 4. Comparing with similar systems

ble) the performance figures achieved by works similar to our own, for the object categories we have used in this paper. However, it should be noted that due to differences in the composition of the datasets, these figures are not directly comparable and should not be treated as such. In [5] efficiency is measured using recall and precision and the corresponding values are shown in Fig. 3 (dashed bars). Fig. 4 displays the efficiency achieved by [4], [6] and the proposed approach using accuracy. We can see from both Figs. 3, 4 that our framework manage to achieve comparable performance, even though our models are trained using images annotated with low-quality user-contributed tags. This is in comparison to the Corel (used in [5],[6]) and Microsoft Research Campridge (used in [4]) datesets, where the annotation tags can be considered more consistent from a computer vision perspective.

4. CONCLUSIONS

In this work we have combined techniques exploiting social information with computer vision algorithms to facilitate object detectors training. Although it is difficult for the proposed framework to be effectively applied for every possible object category, the social aspect of user contributed content and its potential to scale in terms of content diversity and size, advocates it's use for the type of objects that appear frequently in social context.

5. ACKNOWLEDGMENT

This work was funded by the X-Media project (www.x-media-project.org) sponsored by the European Commission as part of the Information Society Technologies (IST) programme under EC grant number IST-FP6-026978.

6. REFERENCES

[1] K. Barnard and D. A. Forsyth, "Learning the semantics of words and pictures," in *ICCV*, 2001, pp. 408–415.

[2] R. Fergus, Fei-Fei Li, P. Perona, and A. Zisserman, "Learning object categories from google's image search," in *ICCV*, 2005, pp. 1816–1823.

[3] M. Vasconcelos, N. Vasconcelos, and G. Carneiro, "Weakly supervised top-down image segmentation," in *CVPR (1)*, 2006, pp. 1001–1006.

[4] J. Verbeek and B. Triggs, "Region classification with markov field aspect models," *IEEE Conf. on Computer Vision and Pattern Recognition.*, pp. 1–8, Jun. 2007.

[5] P. Duygulu, K. Barnard, João F. G. de Freitas, and David A. Forsyth, "Object recognition as machine translation: Learning a lexicon for a fixed image vocabulary," in *ECCV (4)*, 2002, pp. 97–112.

[6] J. Li and J. Ze Wang, "Automatic linguistic indexing of pictures by a statistical modeling approach," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 25, no. 9, pp. 1075–1088, 2003.

[7] E. Giannakidou, I. Kompatsiaris, and A. Vakali, "Semsoc: Semantic, social and content-based clustering in multimedia collaborative tagging systems," in *ICSC*, 2008, pp. 128–135.

[8] V. Mezaris, I. Kompatsiaris, and M. G. Strintzis, "Still image segmentation tools for object-based multimedia applications," *IJPRAI*, vol. 18, no. 4, pp. 701–725, 2004.

[9] B. S. Manjunath, J. R. Ohm, V. V. Vinod, and A. Yamada, "Colour and texture descriptors," *IEEE Trans. Circuits and Systems for Video Technology, Special Issue on MPEG-7*, vol. 11, no. 6, pp. 703–715, Jun 2001.