Leveraging social media for scalable object detection

E. Chatzilari[a,b], S. Nikolopoulos[a,c,*], I. Patras[c], I. Kompatsiaris[a]

[a] *Centre for Research and Technology Hellas - Informatics and Telematics Institute - 6th km Charilaou-Thermi Road, Thermi-Thessaloniki - GR-57001 Thessaloniki - Greece - Tel. +30-2311.257701-3, Fax.+30-2310-474128*

[b] *Centre for Vision, Speech and Signal Processing University of Surrey Guildford, GU2 7XH, UK*

[c] *School of Electronic Engineering and Computer Science - Queen Mary University of London, E1 4NS, London, UK - Tel. +44 20 7882 7523, Fax. +44 20 7882 7997*

Abstract

In this manuscript we present a method that leverages social media for the effortless learning of object detectors. We are motivated by the fact that the increased training cost of methods demanding manual annotation, limits their ability to easily scale in different types of objects and domains. At the same time, the rapidly growing social media applications have made available a tremendous volume of tagged images, which could serve as a solution for this problem. However, the nature of annotations (i.e. global level) and the noise existing in the associated information (due to lack of structure, ambiguity, redundancy, emotional tagging), prevents them from being readily compatible (i.e. accurate region level annotations) with the existing methods for training object detectors. We present a novel approach to overcome this deficiency by using the collective knowledge aggregated in social sites to automatically determine a set of image regions that can be associated with a certain object. We study theoretically and experimentally when the prevailing trends (in terms of appearance frequency) in visual and tag information space converge into the same object, and how this convergence is influenced by the number of utilized images and the accuracy of the visual analysis algorithms. Evaluation results show that although the models trained using leveraged social media are inferior to the ones trained manually, there are cases where the user contributed content can be successfully used to facilitate scalable and effortless learning of object detectors.

Keywords: Social media, collaborative tagging, flickr, object detection, weak annotations, scalable learning, effortless learning

[*] Corresponding author. Tel: +30-2311257752, Fax:+30-2310474128

Email addresses: `ehatzi@iti.gr` (E. Chatzilari), `nikolopo@iti.gr` (S. Nikolopoulos), `i.patras@eecs.qmul.ac.uk` (I. Patras), `ikom@iti.gr` (I. Kompatsiaris)

1. Introduction

Semantic object detection is considered one of the most useful operations performed by the human visual system and constitutes an exciting problem for computer vision scientists. Many researchers in the field have focused on trying to discover a scalable (in terms of the number of concepts) and effortless (in terms of the necessary annotation) way to teach the machine how to recognize visual objects the way a human does. The authors of [1] make the hypothesis that once a few categories have been learned with significant cost, some information may be abstracted from the process to make learning further categories more efficient. Similarly in [2] when images of new concepts are added to the visual analysis model, the computer only needs to learn from the new images, since profiling models are used to store the information learned from previous concepts. In the same lines the need to efficiently handle the huge amounts of data generated on the Web, has prompted many researchers to investigate the use of online learning algorithms [3] for exploiting those data. Motivated by the same need but relying on a non-parametric approach, the authors of [4] claim that with the availability of overwhelming amounts of data many problems can be solved without the need for sophisticated algorithms. The authors present a visual analog to Google's "Did you mean" tool, which corrects errors in search queries by memorizing billions of query-answer pairs and suggesting the one closest to the user query. Additionally, the authors of [5] employ multiple instance learning [6] to learn models from globally annotated images, while in [7] object recognition is viewed as machine translation that uses expectation maximization in order to learn how to map visual objects (blobs) to concept labels. The approaches relying on human computation such as Google Image Labeler [8] and Peekaboom [9] for image global and regional annotation respectively, also belong to the category of methods that aim at scalable and effortless learning. Motivated by the same objective, in this work we investigate whether the knowledge aggregated in social tagging systems by the collaboration of web users, can help in the process of teaching the machine to recognize objects.

Machine learning algorithms for object detection fall in two main categories in terms of the annotation granularity characterizing their learning samples. The algorithms that are designed to learn from strongly annotated samples [10], [11], [12] (i.e. samples in which the exact location of an object within an image is known) and the algorithms that learn from weakly annotated samples [13], [7], [5], [14] (i.e. samples in which it is known that an object is depicted in the image, but its location is unknown). In the first case, the goal is to learn a mapping from visual features f_i to semantic labels c_i (e.g. a face [10], [12] or a car [11]) given a training set made of pairs (f_i, c_i). New images are annotated by using the learned mapping to derive the semantic labels that correspond to the visual features of the new image. On the other hand, in the case of weakly annotated training samples the goal is to estimate the joint probability distribution between the visual features f_i and the semantic labels c_i given a training set made of pairs between sets $\{(f_1, \ldots, f_n), (c_1, \ldots, c_m)\}$. New images are annotated by choosing the semantic labels that maximize the learned joint probability distribution given the visual features of the new image. Some indicative works that fall within the weakly supervised framework include the ones relying on aspect models like probabilistic Latent Semantic Analysis (pLSA) [13], [15] and Latent Dirichlet Allocation (LDA) [16], [17] that are typically used for estimating the necessary joint probability distribution.

While model parameters can be estimated more efficiently from strongly annotated samples, such samples are very expensive to obtain raising scalability problems. On the contrary, weakly annotated samples can be easily obtained in large quantities from social networks. Motivated by this fact our work aims at combining the advantages of both strongly supervised (learn model parameters more efficiently) and weakly supervised (learn from samples obtained at low cost) methods, by allowing the strongly supervised methods to learn from training samples that can be mined from collaborative tagging environments. The problem we consider is essentially a multiple-instance learning problem in noisy context, where we try to exploit the noise reduction properties that characterize massive user contributions, given that they encode the collective knowledge of multiple users. Indeed, Flickr hosts a series of implicit links between images that can be mined using criteria such as geo-location information, temporal proximity between the timestamps of images uploaded by the same user, or images associated with the same event. The goal of this work is to exploit the social aspect of the contributed content at the level of tags. More specifically, given that in collaborative tagging environments the generated annotations may be considered to be the result of the collaboration among individuals, we can reasonably expect that tag assignments are filtered by the collaborative effort of the users, yielding more consistent annotations. In this context, drawing from a large pool of weakly annotated images, our goal is to benefit from the knowledge aggregated in social tagging systems in order to automatically determine a set of image regions that can be associated with a certain object.

In order to achieve this goal, we consider that if the set of weakly annotated images is properly selected, the most populated tag-“term” and the most populated visual-“term” will be two different representations (i.e. textual and visual) of the same object. We define tag-“terms” to be sets of tag instances grouped based on their semantic affinity (e.g. synonyms, derivatives, etc.). Respectively, we define visual-“terms” to be sets of region instances grouped based on their visual similarity (e.g. clustering using the regions’ visual features). The most populated tag-“term” (i.e. the most frequently appearing tag, counting also its synonyms, derivatives, etc.) is used to provide the semantic label of the object that the developed classifier is trained to recognize, while the most populated visual-“term” (i.e. the most populated cluster of image regions) is used to provide the set of positive samples for training the classifier in a strongly supervised manner. Our method relies on the fact that due to the common background that most users share, the majority of them tend to contribute relevant tags when faced with similar type of visual content [18]. Given this fact, it is expected that as the pool of the weakly annotated images grows, the most frequently appearing “term” in both tag and visual information space will converge into the same object.

In this context, the contribution of our work is on studying theoretically and experimentally the conditions under which the most frequently appearing "terms" in tag and visual information space are expected to converge into the same object. This is evident in the ideal case where tags are accurate and free of ambiguity, and no error is introduced by the visual analysis algorithms. However, considering that this is rarely the case, we expect that the use of a large size dataset favors convergence since a statistically significant amount of samples can compensate for the error introduced by noisy tagging. On the contrary, the amount of error introduced by the visual analysis algorithms (i.e. segmentation accuracy and clustering efficiency) hinders convergence since the formulated clusters of image regions may not be consistent in a semantic sense. Our purpose in this work is to examine how these two aforementioned factors influence the convergence level between the most frequently appearing "terms" in visual and tag information space.

Preliminary versions of this work include [19] and [20]. The main difference with [19] is that in this early work we have followed a different methodological approach for selecting the set of regions that can be associated with a certain object. More specifically, the full set of image regions was split in two clusters and the cluster with the smallest population was selected to provide the training samples for the object detection model. Although successful for the objects that appeared frequently in social context, it was observed that our framework performed poorly for a non-negligible number of cases. This was the reason for turning into the methodological approach presented in this work, an early version of which was included in [20]. However, while the focus of [20] has been mostly on experimenting with various feature spaces and tuning the clustering algorithm, in this manuscript we provide a solid theoretical ground for gaining insight into the functionality of the proposed approach and deriving some conclusions about its success or failure. Moreover, we experimentally examine the ability of our method in scaling to various types of objects, allowing us to derive useful conclusions about the learning efficiency of the resulting object detection models.

The rest of the manuscript is organized as follows. Section 2 reviews the related literature. Section 3 describes the general architecture of the framework we propose for leveraging social media and provides technical details for the analysis components that are employed by our framework. Section 4 investigates theoretically the relation between the size of the dataset, the visual analysis error and the convergence level of the most frequently appearing tag and visual "terms". Our experimental study is presented in Section 5, while Section 6 discusses the results and provides some directions for future work.

2. Related Work

Lately there has been considerable interest on weakly labeled data and their potential to serve as the training samples for various computer vision tasks. The common objective of these approaches is to compensate for the loss in learning from weakly annotated and noisy training data, by exploiting the arbitrary large amount of available samples. Web 2.0 and collaborative tagging environments have further boosted this idea by making available plentiful user tagged data.

Our work can be considered to relate with various works in the literature in different aspects. From the perspective of exploring the trade-offs between analysis efficiency and the characteristics of the dataset we find similarities with [21], [22]. In [21] the authors explore the trade-offs in acquiring training data for image classification models through automated web search as opposed to human annotation. The authors try to learn a model that operates on prediction features (i.e. cross-domain similarity, model generalization, concept frequency, within-training-set model quality) and provide quantitative measures in order to estimate when the cheaply obtained data is of sufficient quality for training robust object detectors. In [22] the authors investigate both theoretically and empirically when effective learning is possible from ambiguously labeled images. They formulate the learning problem as partially-supervised multiclass classification and provide intuitive assumptions under which they expect learning to succeed. This is done by using convex formulation and showing how to extend a general multiclass loss function to handle ambiguity.

There are also works [23], [24], [25] that rely on the same principle assumption with our work, stating that users tend to contribute similar tags when faced with similar type of visual content. In [23] the authors are based on social data to introduce the concept of flickr distance. Flickr distance is a measure of the semantic relation between two concepts using their visual characteristics. The authors rely on the assumption that images about the same concept share similar appearance features and use images obtained from flickr to represent a concept. Although different in purpose from our approach the authors present some very interesting results demonstrating that social media like flickr can be used to facilitate various computer vision tasks. In [24] the authors make the assumption that semantically related images usually include one or several common regions (objects) with similar visual features. Based on this assumption they build classifiers using as positive examples the regions clustered in a cluster that is decided to be representative of the concept. They use multiple region-clusters per concept and eventually they construct an ensemble of classifiers. They are not concerned with object detection but rather with concept detection modeled as a mixture/constellation of different object detectors. In the same lines, the work presented in [25] investigates inexpensive ways to generate annotated training samples for building concept classifiers. The authors utilize clickthrough data logged by retrieval systems that consist of the queries submitted by the users, together with the images from the retrieved results, that these users selected to click on in response to their queries. The method is evaluated using global concept detectors and the conclusion that can be drawn from the experimental study is that although the automatically generated data cannot surpass the performance of the manually produced ones, combining both automatically and manually generated data consistently gives the best results.

The employment of unsupervised methods (e.g. clustering) for mining images depicting certain objects, is the attribute that relates our work with [26], [14]. In [26] the authors make use of community contributed collections and demonstrate a location-tag-vision-based approach for retrieving images of geography-related landmarks. They use clustering for detecting representative tags for landmarks, based on their location and time information. Subsequently, they combine this information with vision-assisted process for presenting the user with a representative set of images. Eventually, the goal is to sample the formulated clusters with the most representative images for the selected landmark. In [14] the authors are concerned with images that are found in community photo collections and depict objects (such as touristic sights). The presented approach is based on geotagged photos and the task is to mine images containing objects in a fully unsupervised manner. The retrieved photos are clustered according to different modalities (including visual content and text labels) and Frequent Itemset Mining is applied on the tags associated with each cluster in order to assign cluster labels. Eventually, the formulated clusters are used to automatically label and geo-locate new photos.

Finally our work bares also similarities with works like, [27], [28] that operate on segmented images with associated text and perform annotation using the joint distribution of image regions and words. In [27] the problem of object recognition is viewed as a process of translating image regions to words, much as one might translate from one language to another. The authors develop a number of models for the joint distribution of image regions and words, using weak annotations. In [28] the authors propose a fully automatic learning framework that learns models from noisy data such as images and user tags from flickr. Specifically, using a hierarchical generative model the proposed framework learns the joint distribution of a scene class, objects, regions, image patches, annotation tags as well as all the latent variables. Based on this distribution the authors support the task of image classification, annotation and semantic segmentation by integrating out of the joint distribution the corresponding variables.

3. Framework Description

3.1. General Framework Architecture

The framework we propose for leveraging social media to train object detection models is depicted in Fig. 1. The analysis components that can be identified in our framework are: a) construction of an appropriate image set, b) image segmentation, c) extraction of visual features from image regions, d) clustering of regions using their visual features and e) supervised learning of object recognition models using strongly annotated samples.

More specifically, given an object c_k that we wish to train a detector for (e.g. *sky* in Fig. 1), our method starts from a large collection of user tagged images and performs the following actions. Images are appropriately selected so as to formulate a set of images that emphasizes on object c_k. By emphasizing we refer to the case where the majority of the images within the image set depict a certain object and that the linguistic description of that object can be obtained from the most frequently appearing tag (see Section 3.2.1 for more details). Subsequently, clustering is performed on all regions extracted from the images of the image set, that have been pre-segmented using an automatic segmentation algorithm. During region clustering the image regions are represented by their visual features and each of the generated clusters typically contains visually similar regions. Since the majority of the images within the selected image set depicts instances of the desired object c_k, we anticipate that the majority of regions representing the object of interest will be gathered in the most populated cluster, pushing all irrelevant regions to the other clusters. Eventually, we use as positive samples the visual features extracted from the regions belonging to the most populated cluster, to train in a supervised manner an SVM-based binary classifier for recognizing instances of c_k. After training the classifier, object detection is performed on unseen images by using the automatic segmentation algorithm to extract their regions, and then apply the classifier to decide whether these regions depict c_k.

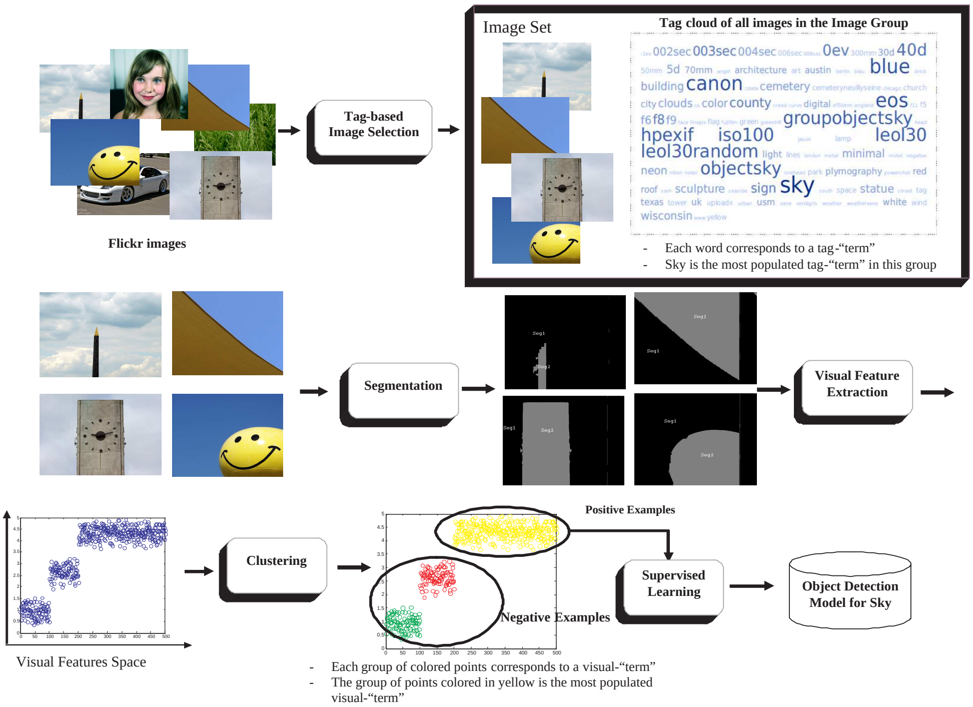

Flickr images
Tag-based Image Selection
Image Set
Tag cloud of all images in the Image Group
-1ev 002sec 003sec 004sec 006sec 008sec 0ev 300mm 30d 40d 50mm 5d 70mm angel architecture art austin berlin blau blue brick building canon cdata cemetery cemeteryneuillyseine chicago church city clouds co color county creed curve digital ef50mm england eos f11 f5 f6 f8 f9 face finepix flag fujifilm green greenhill groupobjectsky head hpexif iso100 jason lamp leol30 leol30random light lines london metal minimal motel negative neon nikon nolan objectsky onehead park plymography powershot red roof sam sculpture seaside sign sky south space statue street tag texas tower uk uploadx urban usm vane verdigris weather weathervane white wind wisconsin www yellow
- Each word corresponds to a tag-“term”
- Sky is the most populated tag-“term” in this group
Segmentation
Seg1
Seg2
Visual Feature Extraction
Visual Features Space
Clustering
Positive Examples
Negative Examples
Supervised Learning
Object Detection Model for Sky
- Each group of colored points corresponds to a visual-“term”
- The group of points colored in yellow is the most populated visual-“term”

Figure 1: Proposed framework for leveraging a set of user tagged images to train a model for detecting the object *sky*.

3.2. Analysis Components

We use the notation of Table 1 to provide technical details, formalize the functionality and describe the links between the components employed by our framework.

Table 1: Legend of used notation

Symbol	**Definition**
S	The complete social media dataset
N	The number of images in S
S^{c_k}	An image set, subset of S that emphasizes on object c_k
n	The number of images in S^{c_k}
I	An image from S
$R = \{r_i, i = 1, \ldots, m\}$	Complete set of regions identified in all images of S^{c_k} by an automatic segmentation algorithm
$T = \{t_i, i = 1, \ldots, n\}$	Complete set of tags contributed for all images of S^{c_k} by web users
$F = \{f_(r_i), i = 1, \ldots, m\}$	Complete set of visual features extracted from all regions in R
$C = \{c_i, i = 1, \ldots, t\}$	Set of distinct objects that appear in the image set S^{c_k}
$\mathbf{R} = \{\mathbf{r_i}, i = 1, \ldots, o\}$	Set of clusters created by performing clustering on the regions extracted from all images of S^{c_k} based on their visual similarity (i.e. visual-terms)
$\mathbf{T} = \{\mathbf{t_j}, j = 1, \ldots, d\}$	Set of clusters created by clustering together the tags contributed for all images in S^{c_k}, based on their semantic affinity (i.e. tag-terms)
p_{c_i}	Probability that tag-based image selection draws from S an image depicting c_i
TC_i	Number of regions depicting object c_i in S^{c_k}

*we use normal letters (e.g. z) to indicate individuals of some population and bold face letters (e.g. **z**) to indicate clusters of individuals of the same population

3.2.1. Construction of an appropriate image set

In this section we refer to the techniques that we use in order construct a set of images emphasizing on object c_k, based on the associated textual information (i.e. annotations). If we define $ling(c_k)$ to be the linguistic description of c_k (e.g. the words "sky", "heaven", "atmosphere" for the object sky), a function describing the functionality of this component takes as input a large set of images and $ling(c_k)$, and returns a set of images S^{c_k}, subset of the initial set, that emphasizes on object c_k.

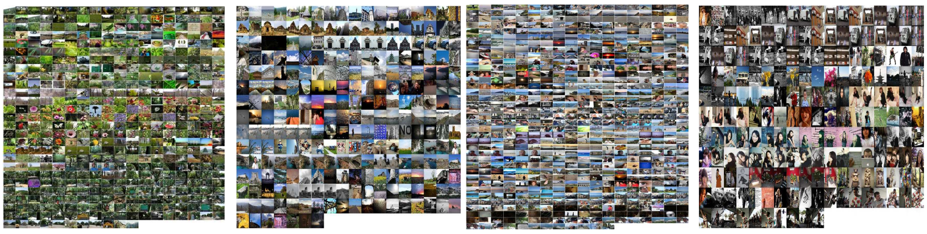
$$imageSet(S, ling(c_k)) = S^{c_k} \subset S \tag{1}$$

For the purposes of our work we use three different implementations of this function based on the type of associated annotations.

Keyword-based selection. This approach is used for selecting images from strongly annotated datasets. These datasets are hand-labeled at region detail and the labels provided by the annotators can be considered to be mostly accurate and free of ambiguity. Thus, in order to create S^{c_k} we only need to select the images where at least one of its regions is labeled with $ling(c_k)$.

Flickr groups. *Flickr groups* (http://www.flickr.com/groups/) are virtual places hosted in collaborative tagging environments that allow social users to share content on a certain topic which can be also an object. Although managing *flickr groups* still involves some type of human annotation (i.e. a human assigns an image to a specific *flickr group*) it can be considered weaker than the previous case since this type of annotation does not provide any information about the boundaries of the object depicted in the image. From here on we will refer to the images obtained from *flickr groups* as roughly-annotated images. In this case, S^{c_k} is created by taking a predefined number of images from a *flickr group* that is titled with $ling(c_k)$. Here, the tags of the images are not used as selection criteria. One drawback of *flickr groups* derives from the fact that since they are essentially virtual places they are not guaranteed to constantly increase their size and therefore cater for datasets of arbitrary scale. Indeed, the total number of positive samples that can be extracted from the images of a *flickr group* has an upper limit on the total number of images that have been included in this group by the users, which is typically much smaller than the total number of flickr images that actually depict this object. This is the reason that we also investigate the following selection technique that operates on image tags, and is therefore capable of producing considerably larger sets of images emphasizing on a certain object.

SEMSOC. SEMSOC stands for SEmantic, SOcial and Content-based clustering and is applied by our framework on weakly annotated images in order to create sets of images emphasizing on different topics. SEMSOC was introduced by Giannakidou et. al. in [29] and is an un-supervised model for the efficient and scalable mining of multimedia social-related data that jointly considers social and semantic features. Given the tendency of social tagging systems to formulate knowledge patterns that reflect the way content is perceived by the web users [18], SEMSOC aims at identifying these patterns and create an image set emphasizing on c_k. The reason for adopting this approach in our framework is to overcome the limitations that characterize collaborative tagging systems such as tag spamming, tag ambiguity, tag synonymy and granularity variation (i.e. different description level). The outcome of applying SEMSOC on a large set of images S, is a number of image sets $S^{c_i} \subset S, \quad i = 1, \ldots, m$, where m is the number of created sets. This number is determined empirically, as described in [29]. Then in order to obtain the image set S^{c_k} that emphasizes on object c_k, we select the SEMSOC-generated set S^{c_i} where its most frequent tag closely relates with $ling(c_k)$. Although the image sets generated by SEMSOC are not of the same quality as those obtained from *flickr groups*, they can be significantly larger favoring the convergence between the most populated visual-



(a) Vegetation (b) Sky (c) Sea (d) Person

Figure 2: Examples of image sets generated using SEMSOC (in caption the corresponding most frequent tag). It is clear that the majority of images in each set include instances of the object that is linguistically described by the most frequent tag. The image is best view in color and with magnification.

and tag-"term". In this case, the total number of positive samples that can be obtained is only limited by the total number of images that have been uploaded on the entire flickr repository and depict the object of interest. Moreover, since SEMSOC considers also the social and semantic features of tags when creating the sets of images, the resulting sets are expected to be of higher semantic coherence than the sets created using for instance, a straightforward tag-based search. Fig. 2 shows four examples of image clusters generated by SEMSOC along with the corresponding most frequent tag.

3.2.2. Segmentation

Segmentation is applied on all images in S^{c_k} with the aim to extract the spatial masks of visually meaningful regions. In our work, we have used a K-means with connectivity constraint algorithm as described in [30]. The output of this algorithm, when applied to a single image, is a set of segments which roughly correspond to meaningful objects, as shown in Fig. 1. Thus, the segmentation analysis component takes as input the full set of images that are included in S^{c_k} and generates an extensive set of independent image regions:

$$segm(S^{c_k}) = \{r_i \in R : \forall I \in S^{c_k}\} \quad (2)$$

3.2.3. Visual Descriptors

In order to visually describe the segmented regions we have employed an approach similar to the one described in [31], with the important difference that in our case descriptors are extracted to represent each of the identified image regions, rather than the whole image. More specifically, for detecting interest points we have applied the Harris-Laplace point detector on intensity channel, which has shown good performance for object recognition [32]. In addition, we have also applied a dense-sampling approach where interest points are taken every 6^{th} pixel in the image. For each interest point (identified both using the Harris-Laplace and dense sampling) the 128-dimensional SIFT descriptor is computed using the version described by Lowe [33]. Then, a Visual Word

Vocabulary (Codebook) is created by using the K-Means algorithm to cluster in 300 clusters, approximately 1 million SIFT descriptors that were sub-sampled from a total amount of 28 million SIFT descriptors extracted from 5 thousand training images. The Codebook allows the SIFT descriptors of all interest points enclosed by an image region, to be vector quantized against the set of Visual Words and create a histogram. Thus, a 300-dimensional feature vector $f(r_i)$ is extracted $\forall r_i \in R$, which contains information about the presence or absence of the Visual Words included in the Codebook. Then, all feature vectors are normalized so as the sum of all elements of each feature vector to be equal to 1. Thus, the visual descriptors component takes as input the full set of independent image regions R extracted from all images in S^{c_k} and generates an equivalent number of feature vectors.

$$vis(R) = \{f(r_i) \in F : \forall r_i \in R\} \tag{3}$$

3.2.4. Clustering

For performing feature-based region clustering we applied the affinity propagation clustering algorithm on all extracted feature vectors F. Affinity propagation was proposed by Frey and Dueck [34] and selected for our work due to the following reasons:

a) The requirements of our framework imply that in order to learn an efficient object detection model, clustering will have to be performed on a considerably large number of regions, making computational efficiency an important issue. In contrast to common clustering algorithms that start with an initial set of randomly selected centers and iteratively refine this set so as to decrease the sum of squared errors, affinity propagation simultaneously considers all data points as potential centers. By viewing each data point as a node in a network, affinity propagation recursively transmits real-valued messages along the edges of the network until a good set of centers and corresponding clusters emerges. In this way, it removes the need to re-run the algorithm with different initializations, which is very beneficiary in terms of computational efficiency.

b) The fact that the number of objects depicted in the full set of images can not be known in advance, poses the requirement for the clustering procedure to automatically determine the appropriate number of clusters based on the analyzed data. Affinity propagation, rather than requiring that the number of clusters is pre-specified, takes as input a real number for each data point, called "preference". These "preference" values influence the number of identified clusters, which also emerges from the message-passing procedure. If a priori, all data points are equally suitable as centers (as in our case) the preferences should be set to a common value. This value can be varied to produce different numbers of clusters and taken for example to be the median of the input similarities (resulting in a moderate number of clusters) or their minimum (resulting in a small number of clusters). The minimum value has been used in our experiments.

Thus, the clustering component takes as input the full set of feature vectors extracted by the visual descriptors component and generates clusters of feature vectors based on a similarity distance between those vectors. These clusters of feature vectors can be directly translated to clusters of regions since there is one to one correspondence between regions and feature vectors. Thus, the functionality of the clustering component can be described as follows:

$$clust(F) = \{\mathbf{r}_i \in \mathbf{R}\} \tag{4}$$

Out of the generated clusters of regions we select the most populated $\mathbf{r}_v$, as described in detail in Section 4, and we use the regions included in this cluster to learn the parameters of a model recognizing c_k.

3.2.5. Learning Model Parameters

Support Vector Machines (SVMs) [35] were chosen for generating the object detection models due to their ability in smoothly generalizing and coping efficiently with high-dimensionality pattern recognition problems. All feature vectors corresponding to the regions assigned to the most populated $\mathbf{r}_v$ of the generated clusters, are used as positive samples for training a binary classifier. Negative examples are chosen arbitrarily from the remaining dataset. Tuning arguments include the selection of Gaussian radial basis kernel and the use of cross validation for selecting the kernel parameters. Thus, the functionality of the model learning component can be described by the following function:

$$svm(vis(\mathbf{r}_v), c_k) = m_{c_k} \tag{5}$$

4. Rationale of our approach

4.1. Problem Formulation

The goal of our framework is to train an SVM-based binary classifier in order to recognize whether a region r_i of an un-seen image I depicts a certain object c_k. In order to do that, we need to provide the classifier with a set of positive and a set of negative samples (i.e. image regions) for c_k. Given that negative samples can be chosen arbitrarily from a random population, our main problem is to find a set of image regions depicting the object c_k, $(\mathbf{r}^+, c_k)$. However, the annotations found in social networks are in the form of tagged images $\{I, (t_1, t_2, \ldots, t_n)\}$, which can be transformed to $\{(r_1, r_2, \ldots, r_m)^I, (t_1, t_2, \ldots, t_n)^I\}$ after segmenting I into regions. Ideally, the tagged images could be used to extract the positive samples for c_k if we could perfectly cluster the visual and tag information space. More specifically, If we take R and T to be the total set of regions and tags extracted from all images in S respectively, by performing clustering based on the *similarity* between the individuals of the same population (i.e. visual similarity for image regions and semantic affinity for contributed tags), we are able to generate clusters of individuals in each population as shown below:

$$\begin{array}{rll} visualCluster(R) = \mathbf{r}_i, & \mathbf{r}_i \subseteq R & \text{visual-terms} \\ tagCluster(T) = \mathbf{t}_j, & \mathbf{t}_j \subseteq T & \text{tag-terms} \end{array} \tag{6}$$

Now, given a large set of tagged images $I \in S$ this process would produce for each object c_l depicted by the images of S, a triplet of the form $(\mathbf{r}_i, \mathbf{t}_j, c_l)$.

Ideally in each triplet, $\mathbf{r}_i$ is the set of regions extracted from all images in S that depict c_l, and $\mathbf{t}_j$ is the set of tags from all images in S that were contributed to linguistically describe c_l. We consider that an object c_l may have many different instantiations in both visual (e.g. different angle, illumination, etc.) and tag (e.g. synonyms or derivatives of the words expressing the object; for instance the object sea can be linguistically described using many different words such as "sea", "seaside", "ocean", etc.) information space. Thus, $\mathbf{r}_i$ can be used to provide the positive samples needed to train the SVM-based classifier, while $\mathbf{t}_i$ can be used to provide the linguistic description of the object that the classifier is trained to recognize. However, the aforementioned process can only be made feasible in the ideal case where the image analysis works perfect and there is no noise in the contributed tags. This is highly unlikely due to the following reasons. From the perspective of visual analysis, in case of over or under segmentation, or in case the visual descriptors are inadequate to perfectly discriminate between different semantic objects, it is very likely that the clustering algorithm will create a different number of clusters than the actual number of semantic objects depicted by the images of S, or even mix regions depicting different objects into the same cluster. From the perspective of tag-analysis the well known problems of social networks (i.e. lack of structure, ambiguity, redundancy, emotional tagging) hinders the process of clustering together the tags contributed to refer to the same object.

For this reason, in our work, we relax the constraints of the aforementioned problem and instead of requiring that one triplet is extracted for every object c_l depicted by the images of S, we only aim at extracting the triplet corresponding to the object c_k, which is the object emphasized by the processed image set. Thus, the first step is to create an appropriate set of images S^{c_k} that emphasizes on object c_k. Then, based on the assumption that there will be a connection between what is depicted by the majority of the images in S^{c_k} and what is described by the majority of the contributed tags, we investigate the level of semantic consistency (i.e. the level of which the majority of regions included in $\mathbf{r}_v$ depict c_k and the majority of tags included in $\mathbf{t}_g$ are linguistically related with c_k) of the triplet $(\mathbf{r}_v, \mathbf{t}_g, c_k)$, if v and g are selected as follows. Since both $\mathbf{r}_i$ and $\mathbf{t}_j$ are clusters (of images regions and tags, respectively), we can apply the $Pop(\cdot)$ function on them, that calculates the population of a cluster (i.e. the number of instances included in the cluster). Then v and g are selected such as the corresponding clusters are the most populated from all clusters generated by the clustering functions of eq. (6), that is $v = \arg\max_i(Pop(\mathbf{r}_i))$ and $g = \arg\max_j(Pop(\mathbf{t}_j))$.

Although the errors generated from imperfect visual analysis may have different causes (e.g. segmentation error, imperfect discrimination between objects), they all hinder the creation of semantically consistent region clusters. Therefore, in our work, we consider that the error generated from the inaccurate clustering of image regions with respect to the existing objects ($error_{cl-obj}$), incorporates all other types of visual analysis error. Similarly, although the contributed tags may incorporate different types of noise (i.e. ambiguity, redundancy, granularity variation, etc.) they all hinder the process of associating a tag with the objects that are depicted in the image, and thus is reflected on the level of emphasis that is given on object c_k when collecting S^{c_k}. Eventually, the problem addressed in this work is what should be the characteristics of S^{c_k} and $error_{cl-obj}$ so as the triplet $(\mathbf{r}_v, \mathbf{t}_g, c_k)$ determined as described above, to satisfy our objective (i.e. that the majority of regions included in $\mathbf{r}_v$ depicts c_k and the majority of tags included in $\mathbf{t}_g$ are linguistically related with c_k).

4.2. Image set construction

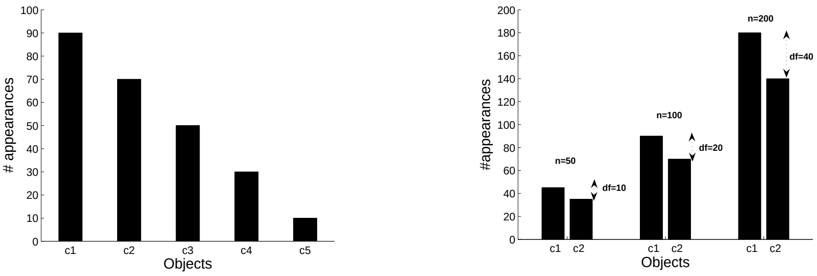
In order to investigate how the characteristics of the constructed image set S^c impact the success probability of our approach, we need to analytically express the association between the number of images included in S^c with the expected number of appearances of any object depicted by those images. Using image tag information to construct an image set that emphasizes on a certain object (e.g. c_1), can be viewed as the process of selecting images from a large pool of weakly annotated images using as argument $ling(c_1)$ (along with possible synonyms, derivatives, etc.). Although misleading and ambiguous tags will hinder this process, the expectation is that as the number of selected images grows, there will be a connection between what is depicted in the majority of the selected images and what is described by the majority of the contributed tags. This can be formalized as follows. When one picks an image from a pool of weakly annotated images using $ling(c_1)$ as an argument, the probability that the selected image depicts c_1 is greater than the probability that the image depicts any other object.

Let us assume that we construct an image set $S^{c_1} \subset S$ that emphasizes on object c_1. What we are interested in is the frequency distribution of objects $c_i \in C$ appearing in S^{c_1} based on their frequency rank. We can view the process of constructing S^{c_1} as the act of populating an image set with images selected from a large dataset S using certain criteria. In this case, the number of times an image depicting object c_i appears in S^{c_1}, can be considered to be equal with the number of successes in a sequence of n independent success/failure trials, each one yielding success with probability p_{c_i}. Given that S is sufficiently large, drawing an image from this dataset can be considered as an independent trial. Thus, the number of images in S^{c_1} that depict object $c_i \in C$ can be expressed by a random variable K following the binomial distribution with probability p_{c_i}. Eq. (7) shows the probability mass function of a random variable following the binomial distribution:

$$Pr_{c_i}(K = k) = \binom{n}{k} p_{c_i}^k (1 - p_{c_i})^{n-k} \tag{7}$$

Given the above, we can use the expected value $E(K)$ of a random variable following the binomial distribution to estimate the expected number of images in S^{c_1} that depict object $c_i \in C$, if they are drawn from the initial dataset S with probability p_{c_i}. This is actually the value of k maximizing the corresponding probability mass function, which is:

$$E_{c_i}(K) = np_{c_i} \tag{8}$$



(a)

(b)

Figure 3: a) Distribution of *#appearances* $\forall c_i \in C$ based on their frequency rank, for n=100 and p_{c_1}=0.9, $p_{c_2} = 0.7$, $p_{c_3} = 0.5$, $p_{c_4} = 0.3$, $p_{c_5} = 0.1$. b) Difference of *#appearances* between c_1, c_2, using fixed values for $p_{c_1} = 0.8$ and $p_{c_2} = 0.6$ and different values for n.

If we consider γ to be the average number of times an object appears in an image, then the number of appearances (*#appearances*) of an object in S^{c_1} is:

$$TC_i = \gamma n p_{c_i} \tag{9}$$

Moreover, based on the assumption mentioned earlier in this section, we accept that there will be an object c_1 that is drawn (i.e. appears in the selected image) with probability p_{c_1} higher than p_{c_2}, which is the probability that an image depicting c_2 is drawn, and so forth for the remaining $c_i \in C$. This assumption is experimentally verified in Section 5.1 where the frequency distribution of objects for different image sets are measured in a manually annotated dataset. Finally, using eq. (9) we can estimate the expected number of appearances (*#appearances*) of an object in S^{c_1}, $\forall c_i \in C$. Fig. 3(a) shows the *#appearances* $\forall c_i \in C$ against their frequency rank, given some example values for p_{c_i} with $p_{c_1} > p_{c_2} > \ldots$. It is clear from eq. (9) that if we consider the probabilities p_{c_i} to be fixed the expected difference, in absolute terms, on the *#appearances* between the first and the second most highly ranked objects c_1 and c_2, increases as a linear function of n (see Fig. 3(b) for some examples). Based on this observation and given the fact that as N increases n will also increase, we examine how the population of the generated region clusters relates with $error_{cl-obj}$ and n.

4.3. Clustering

The purpose of this section is to help the reader derive some intuitive conclusions about the impact of the dataset size and the error introduced by the visual analysis algorithms $error_{cl-obj}$, on the success probability of our approach. In order to do this we examine clustering from the perspective of how much a possible solution deviates from the perfect case. This allows us to approximate $error_{cl-obj}$ with a measurable quantity and derive an analytical form of the association between the visual analysis error, the size of the dataset and an indicator of the success probability of our approach.

Given an image set S^{c_1} that emphasizes on object c_1 the goal of region clustering is to group together regions representing the same object. If perfect grouping is accomplished in a semantic sense, the distribution of clusters' population based on their population rank, coincides with the distribution of objects' *#appearances* based on their frequency rank. In this case, the most populated cluster contains all regions depicting the most frequently appearing object. However, as the visual analysis techniques are expected to introduce error, we are interested on the connection between the $error_{cl-obj}$ and the population of the resulting clusters. Since there is no way to explicitly measure the $error_{cl-obj}$, we use the notation of Table 2 to approximate its effect on the population of the generated clusters.

Table 2: Notations for Clustering

Symbol	**Definition**
Pop_j	Population of cluster $\mathbf{r}_j$
$FP_{i,j}$	False positives of $\mathbf{r}_j$ with respect to c_i
$FN_{i,j}$	False negatives of $\mathbf{r}_j$ with respect to c_i
$DR_{i,j} =$ $FP_{i,j} - FN_{i,j}$	Displacement of $\mathbf{r}_j$, with respect to c_i

Without loss of generality we work under the assumption that due to the $error_{cl-obj}$ it is more likely for the cluster corresponding to the second most frequently appearing object, to become more populated than the cluster corresponding to the first most frequently appearing object, than any other cluster. A cluster that corresponds to an object c_i is considered to be the cluster that exhibits the highest F-measure (F_1) score, with respect to that object, among all generated clusters. Thus, the cluster corresponding to object c_i is found using function Z, which is defined as:

$$Z(c_i, \mathbf{R}) = \mathbf{r}_\kappa, \quad \kappa = \arg\max_j (F_1(c_i, \mathbf{r}_j)) \tag{10}$$

where F_1 is the harmonic mean of precision (prec) and recall (rec) and is calculated using the following equation:

$$\begin{gathered} F_1(c_i, w_j) = \tfrac{2prec_{i,j}rec_{i,j}}{prec_{i,j}+rec_{i,j}} \quad with \\ rec_{i,j} = \tfrac{TC_i - FN_{i,j}}{TC_i}, \quad prec_{i,j} = \tfrac{TC_i - FN_{i,j}}{Pop_j} \end{gathered} \tag{11}$$

Then, given that $\mathbf{r}_\kappa$ has been decided to be the corresponding cluster of c_i, the population Pop_κ of the cluster $\mathbf{r}_\kappa$ is equal to the number of regions TC_i depicting c_i, adding the number of false positives $FP_{i,\kappa}$ and removing the number of false negatives $FN_{i,\kappa}$ that have been generated from the $error_{cl-obj}$. Thus, we have:

$$\begin{gathered} Pop_\kappa = TC_i + FP_{i,\kappa} - FN_{i,\kappa} \Rightarrow \\ Pop_\kappa = TC_i + DR_{i,\kappa} \end{gathered} \tag{12}$$

$DR_{i,\kappa}$ is defined to be the displacement of $\mathbf{r}_k$ with respect to c_i and is an indicator of how much the content of $\mathbf{r}_k$ deviates from the perfect solution. $DR_{i,\kappa}$ shows how the Pop_κ of cluster $\mathbf{r}_\kappa$ is modified according to the $error_{cl-obj}$ introduced by the visual analysis algorithms. Positive values of $DR_{i,\kappa}$ indicates inflows in $\mathbf{r}_\kappa$ population, while negative values indicate leakages. In the typical case where the clustering result does not exhibit high values for $FP_{i,\kappa}$ and $FN_{i,\kappa}$ simultaneously (see Section 5.2), $DR_{i,\kappa}$ is also an indicator of result's quality since it shows how much the content of a cluster has been changed with respect to the perfect case. Let us denote $\mathbf{r}_\alpha = Z(c_1, \mathbf{R})$ and $\mathbf{r}_\beta = Z(c_2, \mathbf{R})$ the clusters corresponding to c_1 (i.e. the most frequently appearing object in S^{c_1}) and c_2 (i.e. the second most frequently appearing object in S^{c_1}), respectively. We are interested in the relation connecting Pop_α and Pop_β given $DR_{1,\alpha}$, $DR_{2,\beta}$. Thus we have:

$$\begin{aligned} Pop_\alpha - Pop_\beta &= TC_1 + DR_{1,\alpha} - TC_2 - DR_{2,\beta} \Rightarrow \\ Pop_\alpha - Pop_\beta &= (TC_1 - TC_2) + (DR_{1,\alpha} - DR_{2,\beta}) \end{aligned} \tag{13}$$

We know about the first parenthesis on the right hand side of the equation that since S^{c_1} emphasizes on c_1 this object will appear more frequently than any other object in S^{c_1}, thus $TC_1 - TC_2 > 0$. In the case where the second parenthesis on the right hand side of the equation is also positive (i.e. $DR_{1,\alpha} - DR_{2,\beta} > 0$), the value $Pop_\alpha - Pop_\beta$ will be greater than zero since it is the sum of two positive numbers. This indicates that despite the $error_{cl-obj}$, cluster $\mathbf{r}_\alpha$ remains the most populated of the generated clusters and continues to be the most appropriate (i.e. in terms of the maximum F_1 criterion) cluster for training a model detecting object c_1. When $DR_{1,\alpha} - DR_{2,\beta} > 0$ we can distinguish between the three qualitative cases for clustering that are described in Table 3. The superscripts are used to indicate the sign (i.e. positive or negative) of the corresponding displacement in each case.

If $DR_{1,\alpha} - DR_{2,\beta} < 0$, the two parentheses of the right hand side of the eq. (13) have different signs and the sign of the value $Pop_\alpha - Pop_\beta$ depends on the difference between the absolute values of $|TC_1 - TC_2|$ and $|DR_{1,\alpha} - DR_{2,\beta}|$. In this case one of the factors controlling whether the most populated cluster $\mathbf{r}_\alpha$ will be the most appropriate cluster for training a model detecting c_1, is the absolute difference between TC_1 and TC_2, which according to our analysis in Section 4.2 depends largely on the number of images n in S^{c_1}. The three qualitative cases for clustering that we can identify when $DR_{1,\alpha} - DR_{2,\beta} < 0$ are shown in Table 3.

In order to get an intuitive view of the relation between n and the probability of selecting the most appropriate cluster when $DR_{1,\alpha} - DR_{2,\beta} < 0$, we approximate the effect of $error_{cl-obj}$ on the distribution of the generated clusters' population by measuring how much a certain clustering solution deviates from the perfect solution. In order to do this, we view clustering as a recursive process with starting point the perfect solution. Then, the deviation of some clustering solution $t+1$ from the perfect solution depends on the deviation of

Table 3: Qualitative cases for clustering

$DR_{1,\alpha} - DR_{2,\beta} > 0$	$DR^{+}_{1,\alpha} > DR^{+}_{2,\beta}$	Both w_α and w_β increase their population but the inflows of w_α are greater than the inflows of w_β.
	$DR^{+}_{1,\alpha} \quad DR^{-}_{2,\beta}$	w_α increases its population while w_β reduces its own.
	$DR^{-}_{1,\alpha} > DR^{-}_{2,\beta}$	Both w_α and w_β reduce their population but the leakages of w_α are lesser than the leakages of w_β.
$DR_{1,\alpha} - DR_{2,\beta} < 0$	$DR^{+}_{1,\alpha} < DR^{+}_{2,\beta}$	Both w_α and w_β increase their population but the inflows of w_α are lesser than the inflows of w_β.
	$DR^{-}_{1,\alpha} \quad DR^{+}_{2,\beta}$	w_α reduces its population while w_β increases its own.
	$DR^{-}_{1,\alpha} < DR^{-}_{2,\beta}$	Both w_α and w_β reduce their population but the leakages of w_α are greater than the leakages of w_β.

*the superscripts indicate the sign (i.e. positive or negative) of the corresponding displacement

the previous solution t from the perfect solution. Respectively, the population of a cluster in solution $t+1$ is equal to the population of this cluster in the previous solution t, adding the number of false positives and removing the number of false negatives that have been generated from the transition $t \rightarrow t+1$. This can be expressed using the following recursive equation:

$$\begin{gathered} Pop_k^{t+1} = Pop_k^t + FP_{i,k}^{t\rightarrow t+1} - FN_{i,k}^{t\rightarrow t+1} \Rightarrow \\ Pop_k^{t+1} = Pop_k^t + DR_{i,k}^{t\rightarrow t+1} \end{gathered} \tag{14}$$

If we take as starting point the perfect solution, we have $Pop_k^0 = TC_i$. If we also consider $DR_{i,k}^{dt}$ to be constant for all transitions, we can find a closed-form solution for the recursive equation:

$$Pop_k^{t+q} = TC_i + qDR_{i,k}^{dt} \tag{15}$$

Where q is the number of transitions that have taken place and provides and intuitive measure of how much distance there is between current clustering solution and the perfect solution. However, TC_i is the number of times the object c_i appears in S^c (*#appearances*) and according to eq. (9) we have $TC_i = \gamma n p_{ci}$. By substituting TC_i in eq. (15) we have:

$$Pop_{\alpha}^{t+q} = \gamma n p_{c_i} + qDR_{i,k}^{dt} \tag{16}$$

Given that $DR_{1,\alpha} - DR_{2,\beta} < 0$, the population of cluster $\mathbf{r}_\alpha$ is increasing/decreasing with a rate lower/higher from the rate that $\mathbf{r}_\beta$ increases/decreases. So, we are interested in the number of transitions that are needed for causing the population of $\mathbf{r}_\alpha$ to become equal or less to the population of $\mathbf{r}_\beta$. The equality corresponds to the minimum number of transitions.

$$\begin{gathered} Pop_{\alpha}^{t+q\mathbf{r}} - Pop_{\beta}^{t+q} \leq 0 \\ \gamma n p_{c_1} + qDR_{1,\alpha}^{dt} - \gamma n p_{c_2} - qDR_{2,\beta}^{dt} \leq 0 \\ q \geq \frac{\gamma n (p_{c_1} - p_{c_2})}{(DR_{2,\beta}^{dt} - DR_{1,\alpha}^{dt})} \end{gathered} \tag{17}$$

In order to derive some conclusions from this equation we need to make the following remarks. Given our basic assumption we have $p_{c_1} > p_{c_2}$. Moreover, given that $DR_{1,\alpha} - DR_{2,\beta} < 0$ we can also accept that $DR_{1,\alpha}^{dt} - DR_{2,\beta}^{dt} < 0$. Thus, all terms on the right hand side of eq. (17) are positive. It is clear from eq. (17) that the number of transitions q required for causing $\mathbf{r}_\alpha$ not to be the most populated of the generated clusters, increases proportionally to the dataset size n and the difference of probabilities $(p_{c_1} - p_{c_2})$. It is important to note that q does not correspond to any physical value since clustering is not a recursive process, it is just an elegant way to help us derive the intuitive conclusion that as n increases, there is higher probability in $\mathbf{r}_\alpha$ being the most appropriate cluster for learning c_1, due to the increased amount of deviation from the perfect solution that can be tolerated.

5. Experimental study

The goal of our study is to use real social data for experimentally validating our expectations on the size of the processed dataset and the error introduced by the visual analysis algorithms. We examine the conditions under which the most populated visual- and tag-“term” converge into the same object and evaluate the efficiency of the object detection models generated by our framework. To this end, in Section 5.1 we experimentally verify that the absolute difference between the first and second most frequently appearing objects in a dataset constructed to emphasize on the former, increases as the size of the dataset grows. Section 5.2 provides an experimental insight on the $error_{cl-obj}$ introduced by the visual analysis algorithms and examines whether our expectation on the most populated cluster holds. In Section 5.3 we compare the quality of object models trained using flickr images leveraged by the proposed framework, against the models trained using manually provided, strongly annotated samples. Moreover, we also examine how the volume of the initial dataset affects the efficiency of the resulting models. In addition to the above, in Section 5.4 we examine the ability of our framework to scale in various types of objects. We close our experimental study in Section 5.5 where we compare our work with other existing methods in the literature.

To carry out our experiments we have relied on three different types of datasets. The first type includes the strongly annotated datasets constructed by asking people to provide region detail annotations of images pre-segmented with the automatic segmentation algorithm of Section 3.2.2. For this case we have used a collection of 536 images S^B from the *Seaside* domain annotated in our lab (http://mklab.iti.gr/project/scef) and the publicly available MSRC dataset (http://research.microsoft.com/vision/cambridge/recognition) S^M consisting of 591 images. The second type refers to the roughly-annotated datasets like the ones obtained from *flickr groups*. In order to create a dataset of this type S^G, for each object of interest, we have downloaded 500 member images from a *flickr group* that is titled with a name related to the name of the object, resulting in 25 groups of 500 images each (12500 in total). The third type refers to the weakly annotated datasets like the ones that can be collected freely from collaborative tagging environments. For this case, we have crawled 3000 S^{F3K} and 10000 S^{F10K} images from flickr, in order to investigate the impact of the dataset size on the efficiency of the generated models. Depending on the annotation type we use the tag-based selection approaches presented in Section 3.2.1 to construct the necessary image sets S^c. Table 4 summarizes the information of the datasets used in our experimental study. Note that since our approach is working on the level of regions rather than the level of images, the number of media objects handled by our framework (i.e. feature extraction, clustering, SVM-learning) is much larger than the number of images depicted in Table 4, approximately multiplied by 7.

5.1. Objects' distribution based on the size of the image set

As claimed in Section 4.2, we expect the absolute difference between the number of appearances (*#appearances*) of the first (c_1) and second (c_2) most highly ranked objects within an image set S^{c_1}, to increase as the volume of the dataset increases. This is evident in the case of keyword-based selection since, due to the fact that the annotations are strong, the probability that the selected image depicts the intended object is equal to 1, much greater than the probability of depicting the second most frequently appearing object. Similarly, in the case of *flickr groups*, since a user has decided to assign an image to the *flickr group* titled with the name of the object, the probability of this image depicting the intended object should be close to 1. On the contrary, for the case of SEMSOC that operates on ambiguous and misleading tags this claim is not evident. For this reason and in order to verify our claim experimentally, we plot the distribution of objects' *#appearances* in four image sets that were constructed to emphasize on objects *sky*, *sea*, *vegetation*, *person*, respectively.

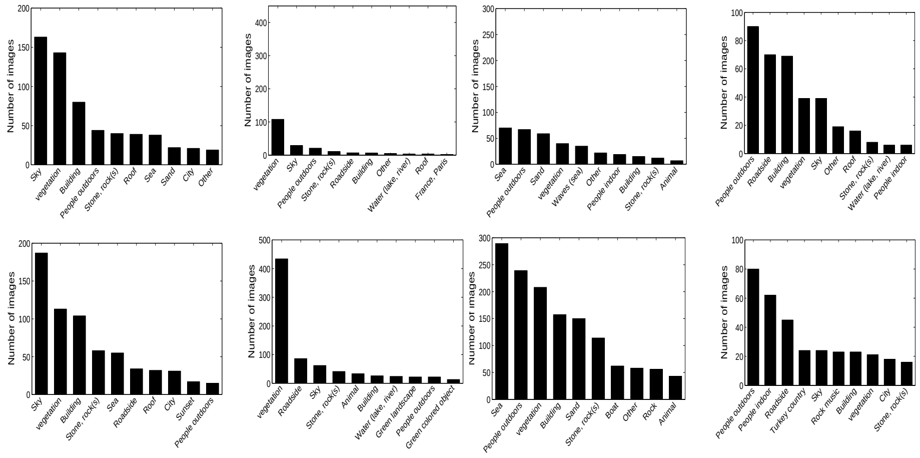
Table 4: Datasets Information

Symbol	**Source**	**Annotation Type**	**No. of Images**	**objects**	**Selection approach**
S^B	internal dataset	strongly annotated	536	sky, sea, vegetation, person, sand, rock, boat	keyword based
S^M	MSRC	strongly annotated	591	aeroplane, bicycle, bird, boat, body, book, cat, chair, cow, dog, face, flower, road, sheep, sing, water, car, grass, tree, building, sky	keyword based
S^G	*flickr groups*	roughly-annotated	12500 (500 for each object)	sky, sea, vegetation, person and the 21 MSRC objects	*flickr groups*
S^{F3K}	flickr	weakly annotated	3000	cityscape, seaside, mountain, roadside, landscape, sport-side	SEMSOC
S^{F10K}	flickr	weakly annotated	10000	jaguar, turkey, apple, bush, sea, city, vegetation, roadside, rock, tennis	SEMSOC

These image sets were generated from both S^{F3K} and S^{F10K} using SEMSOC. Each of the bar diagrams depicted in Fig. 4, describes the distribution of objects' *#appearances* inside an image set S^c, as evaluated by humans. This annotation effort was carried out in our lab and its goal was to provide weak but noise-free annotations in the form of labels for the content of the images included in both S^{F3K} and S^{F10K}. It is clear that as we move from S^{F3K} to S^{F10K} the difference, in absolute terms, between the number of images depicting c_1 and c_2 increases in all four cases, advocating our claim about the impact of the dataset size on the distribution of objects' *#appearances*, when using SEMSOC.

5.2. Clustering assessment

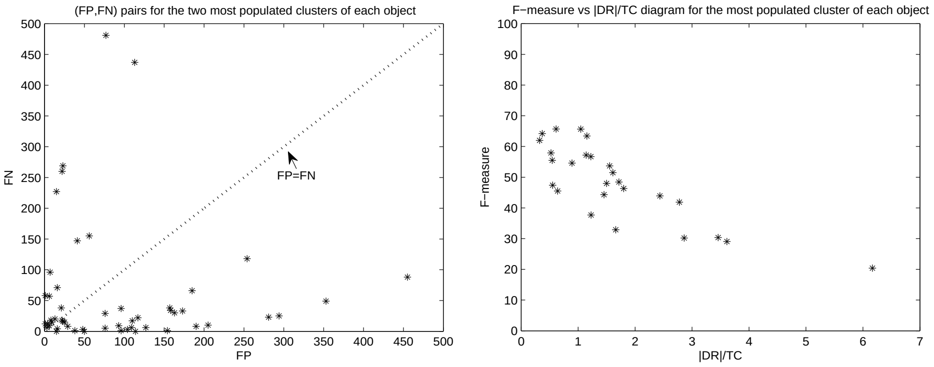
The purpose of this experiment is to provide insight on the validity of our approach in selecting the most populated cluster, in order to train a model recognizing the most frequently appearing object. In order to do so we evaluate the content of each of the formulated clusters using the strongly annotated datasets S^B and S^M. More specifically, $\forall c_i$ depicted in S^B or S^M we obtain $S^{c_i} \subset S^B$ or $S^{c_i} \subset S^M$ using keyword based search and apply clustering on



(a) Sky (b) Vegetation (c) Sea (d) Person

Figure 4: Distribution of objects' #appearance in an image set S^c, generated from S^{F3K} (upper line) and S^{F10K} (bottom line) using SEMSOC

the extracted regions. Then, for each S^{c_i} we calculate the values TC_1, $DR_{1,\alpha}$ and Pop_α for the most frequently appearing object c_1 and its corresponding cluster $\mathbf{r}_a$; and TC_2, $DR_{2,\beta}$ and Pop_β for the second most frequently appearing object c_2 and its corresponding cluster $\mathbf{r}_\beta$. Both $\mathbf{r}_\alpha$ and $\mathbf{r}_\beta$ are determined based on eq. (10) of Section 4.3. Subsequently, we examine whether $\mathbf{r}_\alpha$ is the most populated among all the clusters generated by the clustering algorithm, not only among $\mathbf{r}_\alpha$ and $\mathbf{r}_\beta$ (i.e. we examine if $Pop_\alpha = \max Pop_i$ for all generated clusters). If this is the case we consider that our framework has succeeded in selecting the most appropriate cluster for training a model to recognize c_1 (a $\surd$ is inserted in the corresponding entry of the *Suc* column of Table 5). If $\mathbf{r}_\alpha$ is not the most populated cluster, we consider that our framework has failed in selecting the appropriate cluster (a X is inserted in the corresponding entry of the *Suc.* column). Table 5 summarizes the results for the 7 objects of S^B and the 19 objects of S^M (the objects bicycle and cat were omitted since there was only one cluster generated). We notice that the appropriate cluster is selected in 21 out of 26 cases advocating our expectation that the $error_{cl-obj}$ introduced by the visual analysis process is usually limited and allows our framework to work efficiently. By examining the figures of Table 5 more thoroughly we realize that $DR_{1,\alpha} - DR_{2,\beta} > 0$ for all success cases, with the only exception of object *sky* for S^B. This is in accordance with our analysis in Section 4.3 which showed that if the relative inflow from $\mathbf{r}_\alpha$ to $\mathbf{r}_\beta$ is positive our framework will succeed in selecting the appropriate cluster. In the case of object *sky* our analysis does not hold due to the excessive level of over-segmentation. Indeed, by examining the content of the images belonging to the image set $S^{sky} \subset S^B$ we realize that despite the fact that *sky* is the most frequently appearing object in the image set, after segmenting all images in S^{sky} and manually annotating the



(a) (b)

Figure 5: a) Diagram showing (FP,FN) scatter plot for $\mathbf{r}_\alpha$ and $\mathbf{r}_\beta$ clusters of all objects. It is evident that the (FP,FN) pairs produced by the clustering algorithm lay close to the diagonal ($FP = FN$) only when they are close to (0,0). b) Diagram showing the F-Measure scores exhibited for the $\mathbf{r}_\alpha$ cluster of each object, against the observed $|DR_{i,j}|$ value of this cluster normalized with the total number of true positives TC_i. The qualitative aspect of $|DR_{i,j}|$ is advocated by the observation that the F-measure tends to decrease as the ratio $|DR_{i,j}|/TC_i$ increases.

extracted regions, the number of regions depicting *sky* $TC_1 = 470$ is less than the number of regions depicting *sea* $TC_2 = 663$. This is a clear indication that the effect of over-segmentation has inverted the objects' distribution making *sea* the most frequently appearing object in S^{sky}. In accordance with our analysis are also the fail cases where the relative inflow from $\mathbf{r}_\alpha$ to $\mathbf{r}_\beta$ is negative (i.e. $DR_{1,\alpha} - DR_{2,\beta} < 0$). In none of this 5 cases the difference between $(TC_1 - TC_2)$ was high enough to compensate for the error introduced by the visual analysis process.

Additionally, we have used the experimental observations of Table 5 in order to verify the qualitative aspect of $|DR_{i,j}|$ mentioned in Section 4.3. More specifically, by producing the (FP,FN) scatter plot for the $\mathbf{r}_\alpha$ and $\mathbf{r}_\beta$ clusters of the 7 *Seaside* and 19 *MSRC* objects (Fig. 5(a)), we verify that no (FP,FN) pairs lay close to the diagonal ($FP = FN$) unless they are close to (0,0). Thus, given that $DR_{i,j} = FP_{i,j} - FN_{i,j}$, there are no cases exhibiting high values for both FP and FN and low values for $|DR_{i,j}|$. This renders $|DR_{i,j}|$ a valid indicator for the quality of the result since a poor quality cluster exhibiting high values for either FP or FN, exhibit also high values for $|DR_{i,j}|$. This qualitative aspect of $|DR_{i,j}|$ is also verified by the diagram of Fig. 5(b). In this diagram we plot the F-measure scores for the $\mathbf{r}_\alpha$ cluster of each object (see Section 5.3), against the observed $|DR_{i,j}|$ value of this cluster normalized by the total number of true positives TC_i. It is evident from the diagram that the F-Measure tends to decrease as the ratio $|DR_{i,j}|/TC_i$ increases, showing a clear connection between the $|DR_{i,j}|$ quantity used in our analysis and the quality of the result.

Table 5: Clustering Output Insights

S^{c_i}	n	c_1	TC_1	$DR_{1,\alpha}$	Pop_α	c_2	TC_2	$DR_{2,\beta}$	Pop_β	Suc.	$sign(DR_{1,\alpha} - DR_{2,\beta})$
S^B $(Seaside)$											
S^{sea} ∗	395	sea	732	-404	328	sky	395	-212	183	X	-
S^{sand}	359	sand	422	136	558	sky	337	-103	234	√	+
S^{rock}	53	rock	155	95	250	sea	86	47	133	√	+
S^{boat}	68	boat	96	120	216	sky	69	-57	12	√	+
S^{person}	215	person	435	-238	198	sea	406	-99	307	X	-
$S^{vegetation}$	80	vegetation	157	140	297	sea	114	59	173	√	+
S^{sky}	418	sky	470	-246	224	sea	663	-324	339	X	+
S^M $(MSRC)$											
S^{sign}	27	sign	65	101	166	building	19	-10	9	√	+
S^{sky}	129	sky	139	-89	50	building	115	119	234	X	-
$S^{building}$	88	building	209	304	513	sky	52	-17	35	√	+
S^{car}	6	car	6	37	43	road	7	-3	4	√	+
S^{road}	74	road	94	269	363	sky	32	93	125	√	+
S^{tree}	100	tree	226	258	484	sky	45	124	169	√	+
S^{body}	32	body	54	195	249	face	19	4	23	√	+
S^{face}	21	face	35	121	156	body	17	10	27	√	+
S^{grass}	154	grass	221	367	588	sky	48	133	181	√	+
S^{bird}	29	bird	58	71	129	grass	15	-6	9	√	+
S^{dog}	27	dog	56	84	140	road	11	21	32	√	+
S^{water}	62	water	113	182	295	sky	19	7	26	√	+
S^{cow}	43	cow	109	114	223	grass	57	-51	6	√	+
S^{sheep}	5	sheep	13	15	28	grass	13	-11	2	√	+
S^{flower}	28	flower	60	103	163	grass	8	12	20	√	+
S^{book}	33	book	149	-55	94	face	5	153	158	X	-
S^{chair}	19	chair	39	95	134	road	9	-3	6	√	+
$S^{aeroplane}$	18	aeroplane	12	50	68	sky	12	-8	4	√	+
S^{boat}	15	boat	25	45	70	water	25	-7	18	√	+

∗ although $Pop_\alpha > Pop_\beta$ in this case, the population Pop_γ of the cluster corresponding to the third most frequently appearing object was found to be the highest, which is why we consider this case as a failure

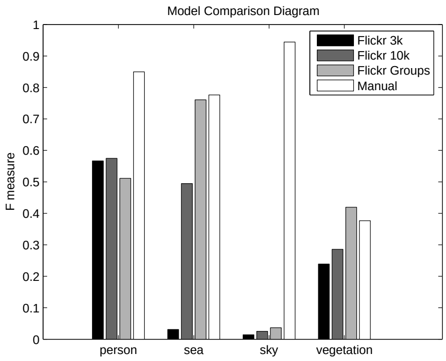


Figure 6: Performance comparison between four object recognition models that are learned using images of different annotation quality (i.e. strongly, roughly and weakly)

5.3. Comparing object detection models

In order to compare the efficiency of the models generated using training samples with different annotation type (i.e. strongly, roughly, weakly), we need a set of objects that are common in all three types of datasets. For this reason after examining the contents of S^B, reviewing the availability of groups in flickr and applying SEMSOC on S^{F3K} and S^{F10K}, we determined 4 object categories C^{bench}={sky, sea, vegetation, person}. These objects exhibited significant presence in all different datasets and served as benchmarks for comparing the quality of the different models. For each object $c_i \in C^{bench}$ one model was trained using the strong annotations of S^B, one model was trained using the roughly-annotated images contained in S^G, and two models were trained using the weak annotations of S^{F3K} and S^{F10K}, respectively. In order to evaluate the performance of these models, we test them using a subset (i.e. 268 images) of the strongly annotated dataset $S^B_{test} \subset S^B$, not used during training. The F_1 metric was used for measuring the efficiency of the models.

By looking at the bar diagram of Fig. 6, we derive the following conclusions: a) Model parameters are estimated more efficiently when trained with strongly annotated samples, since in 3 out of 4 cases they outperform the other models and sometimes by a significant amount (e.g. sky, person). b) *Flickr groups* can serve as a less costly alternative for learning the model parameters, since using the roughly-annotated samples we get comparable and sometimes even better (e.g. vegetation) performance than manually trained models, while requiring considerable less effort to obtain the training samples. c) The models learned from weakly annotated samples are usually inferior from the other cases, especially in cases where the proposed approach for leveraging the data has failed in selecting the appropriate cluster (e.g. *sea* and *sky* for the S^{F3K} dataset). However, the efficiency of the models trained using weakly annotated samples improves when the size of the dataset increases. From the bar diagram of Fig. 6, it is clear that when using the S^{F10K} the incorporation of a larger number of positive samples into the training set improves the generalization ability of the generated models in all four cases. Moreover, in the case of object *sea* we note also a drastic improvement of the model's efficiency. This is attributed to the fact that the increment of the dataset size compensates, as explained in Section 4, for the $error_{cl-obj}$ and allows the proposed method to select the appropriate cluster. On the other hand, in the case of object *sky* it seems that the correct cluster is still missed despite the use of a larger dataset. The correct cluster is also missed for the object *sky* when the weakly annotated samples are obtained from *flickr groups*. This shows that $error_{cl-obj}$ is considerably high for this object and does not allow our framework to select the correct cluster.

5.4. Scaling in various types of objects

In order to test the ability of our approach in scaling to various types of objects we have performed experiments using the MSRC dataset. MSRC (S^M) is a publicly available dataset that has been widely used to evaluate the performance of many object detection methods. The reason for choosing MSRC over other publicly available benchmarking datasets, such as the the PASCAL VOC challenge [36], was its widespread adoption by many works in the literature allowing us to compare our work with state of the art methods (see Section 5.5). MSRC consists of 591 hand-segmented images annotated at region detail for 23 objects. Due to their particular small number of samples *horse* and *mountain* objects were ignored in our study. In order to test our approach for these objects we have relied on *flickr groups* to obtain 21 image groups, with 500 members each, suitable for training models for the 21 objects of S^M. All images of S^M were segmented by the segmentation algorithm described in Section 3.2.2 and the ground truth label of each segment was taken to be the label of the hand-labeled region that overlapped with the segment by more than the 2/3 of the segment's area. In any other case the segment was labeled as void. The S^M was split randomly in 295 training S^M_{train} and 296 testing S^M_{test} images, ensuring approximately proportional presence of each object in both sets.

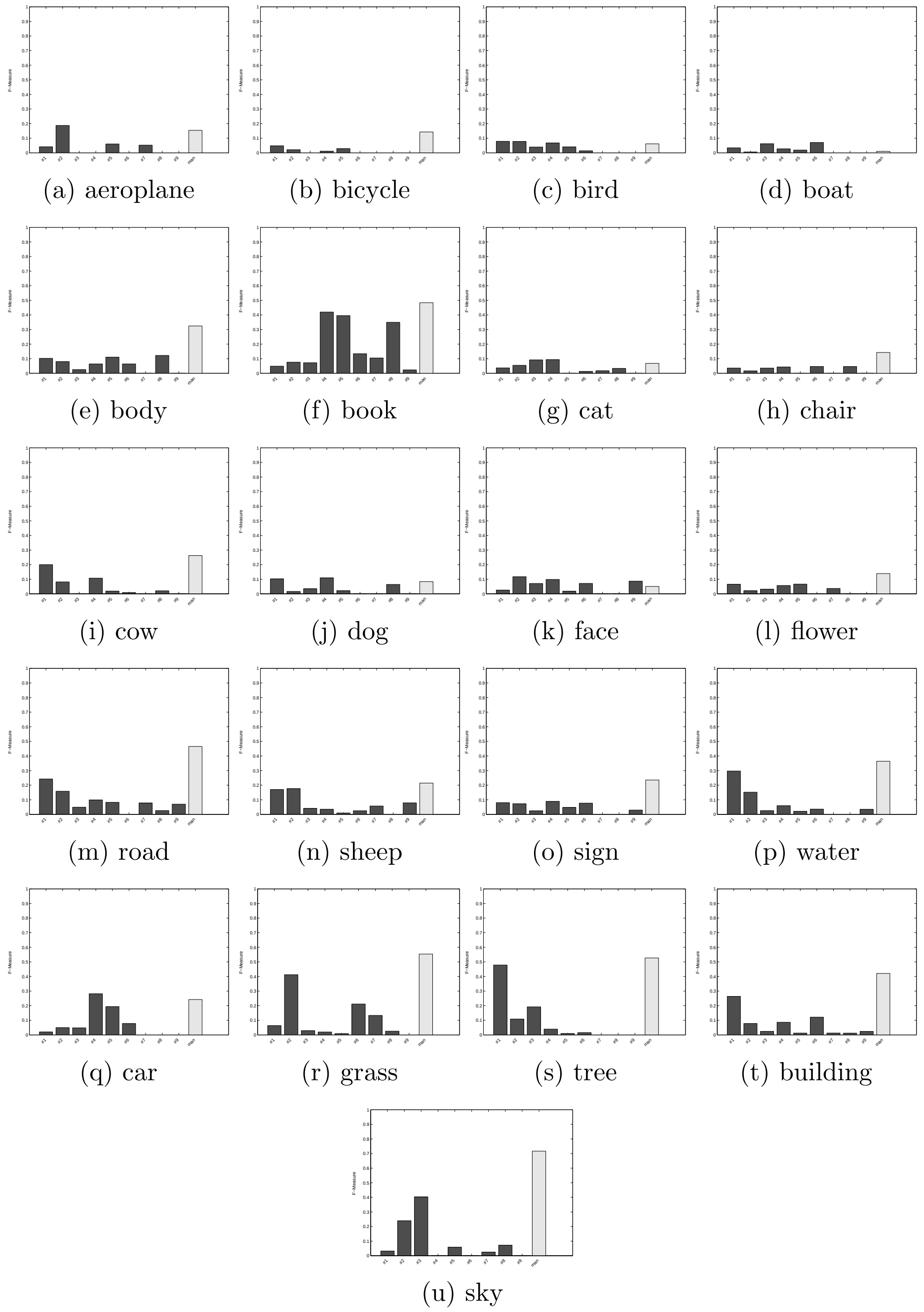
In an attempt not only to evaluate the efficiency of the developed models but also to discover whether the root cause for learning a bad model is the selection of an inappropriate set of training samples, or the deficiency of the employed visual feature space to discriminate the examined object, we perform the following. Since we don't have strong annotations for the images obtained from *flickr groups* and is impossible to assess the quality of the generated clusters as performed in Section 5.2, we train as many models as the number of generated clusters (not only using the most populated) and test them using S^M_{test}. Our aim is to assess the quality of the generated clusters indirectly, by looking at the recognition rates of the models trained with the member regions of each cluster. The bar diagrams of Fig. 7 show the object recognition rates (measured using the F_1 metric) for the models trained using as positive samples the members of each of the nine most populated (in descending order) clusters. The last bar in each diagram corresponds to the performance of the model trained using the strong annotations of S^M_{train} and tested using S^M_{test}. Moreover, in order to visually inspect the content of the generated clusters we have implemented a viewer that is able to read the clustering output and simultaneously display all regions included in the same cluster. By having an overall view of the regions classified in each cluster we can better understand the distribution of clusters to objects and derive some conclusions on the reasons that make the proposed approach to succeed or fail. By looking at the bar diagrams of Fig. 7 we can distinguish between four cases.

In the first case we classify the objects *bird*, *boat*, *cat*, *dog* and *face* that are too diversiform with respect to the employed visual feature space and as a consequence, none of the developed models (not even the one trained using the manual annotations) manage to achieve good recognition rates. In addition to that, the particular small number of relevant regions in the testing set renders most of these objects inappropriate for deriving useful conclusions.

In the second case we classify the objects *bicycle*, *body*, *chair*, *flower* and *sign* that although seem to be adequately discriminated in the visual feature space (i.e. the model trained using the manually annotated samples performs relatively well), none of the models trained using the formulated clusters manages to deliver significantly better recognition rates from the other clusters. Thus, none of the generated clusters contains good training samples which indicates that the images included in the selected *flickr group* are not representative of the examined object, as perceived by the MSRC annotators.

Aeroplane, *book*, *car*, *grass*, *sky*, *sheep* are classified in the third case including the objects that are effectively discriminated in the visual feature space (i.e. the model trained using the manually annotated samples performs relatively well) and there is at least one cluster that delivers performance comparable with the manually trained model. However, the increased $error_{cl-obj}$ has prevented this cluster to be the most populated, since the regions representing the examined object are split in two or more clusters. Indeed, if we take for instance the object *sky* and use the viewer to visually inspect the content of the formulated clusters, we realize that clustering has generated many different clusters containing regions depicting sky. As a result the cluster containing the regions of textured objects has become the most populated. Fig. 8 shows indicative images for some of the generated clusters for object *sky*. The clusters' rank (#) refers to their population. We can see that the clusters ranked #2, #3, #6 and #7 contain *sky* regions while the most populated cluster #1 contains the regions primarily depicting statues and buildings. Consistently, we can see in Fig. 7 that the performance of the models trained using clusters #2, #3 is much better than the performance of the model trained using cluster #1.

Finally, in the last case we classify the objects *cow*, *road*, *water*, *tree*, *building*, where our proposed approach succeeds in selecting the appropriate cluster and allows the classifier to learn an efficient model. Fig. 9 presents some indicative regions for 6 out of the 9 clusters, generated by applying the proposed approach for the object *tree*. For each cluster we present five indicative images in order to show the tendency, in a semantic sense, of the regions aggregated in each cluster. It is interesting to see that most of the formulated clusters tend to include regions of a certain semantic object such as *tree* (#1), *grass* (#2), *sky* (#5), *water* (#9) or noise regions. In these cases where the $error_{cl-obj}$ is



(a) aeroplane (b) bicycle (c) bird (d) boat

(e) body (f) book (g) cat (h) chair

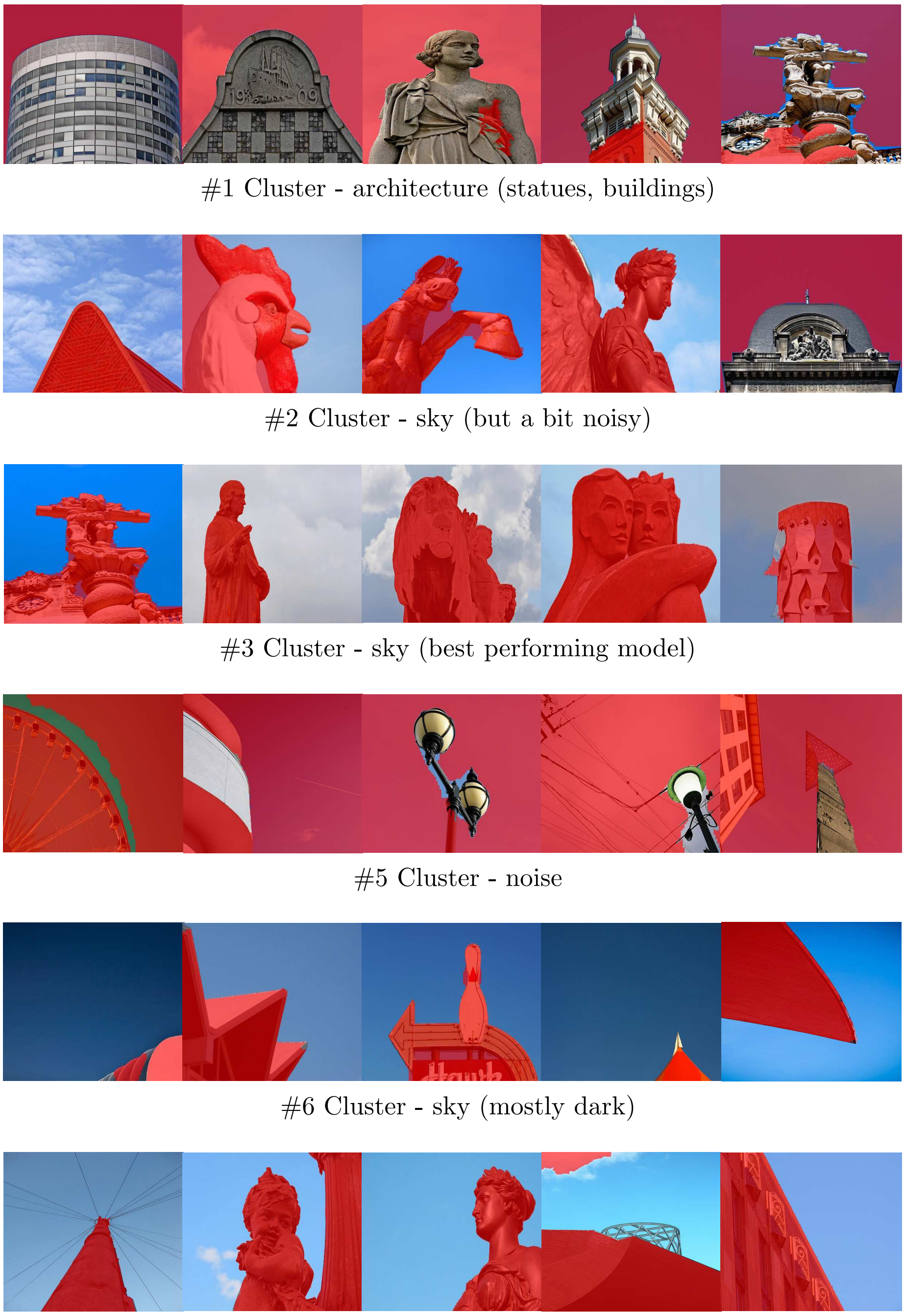
(i) cow (j) dog (k) face (l) flower

(m) road (n) sheep (o) sign (p) water

(q) car (r) grass (s) tree (t) building

(u) sky

Figure 7: Experiments on the 21 objects of MSRC dataset. In each bar diagram the nine first bars (colored in black) show the object recognition rates (measured using F_1 metric) for the models trained using as positive samples the members of each of the nine most populated (in descending order) clusters. The last bar (colored in gray) in each diagram correspond to the performance of the model trained using strongly annotated samples.



#1 Cluster - architecture (statues, buildings)

#2 Cluster - sky (but a bit noisy)

#3 Cluster - sky (best performing model)

#5 Cluster - noise

#6 Cluster - sky (mostly dark)

#7 Cluster - sky (mostly light)

Figure 8: Indicative regions from the clusters generated by applying our approach for the object *sky*. The regions that are not covered in red are the ones that have been assigned to the corresponding cluster.

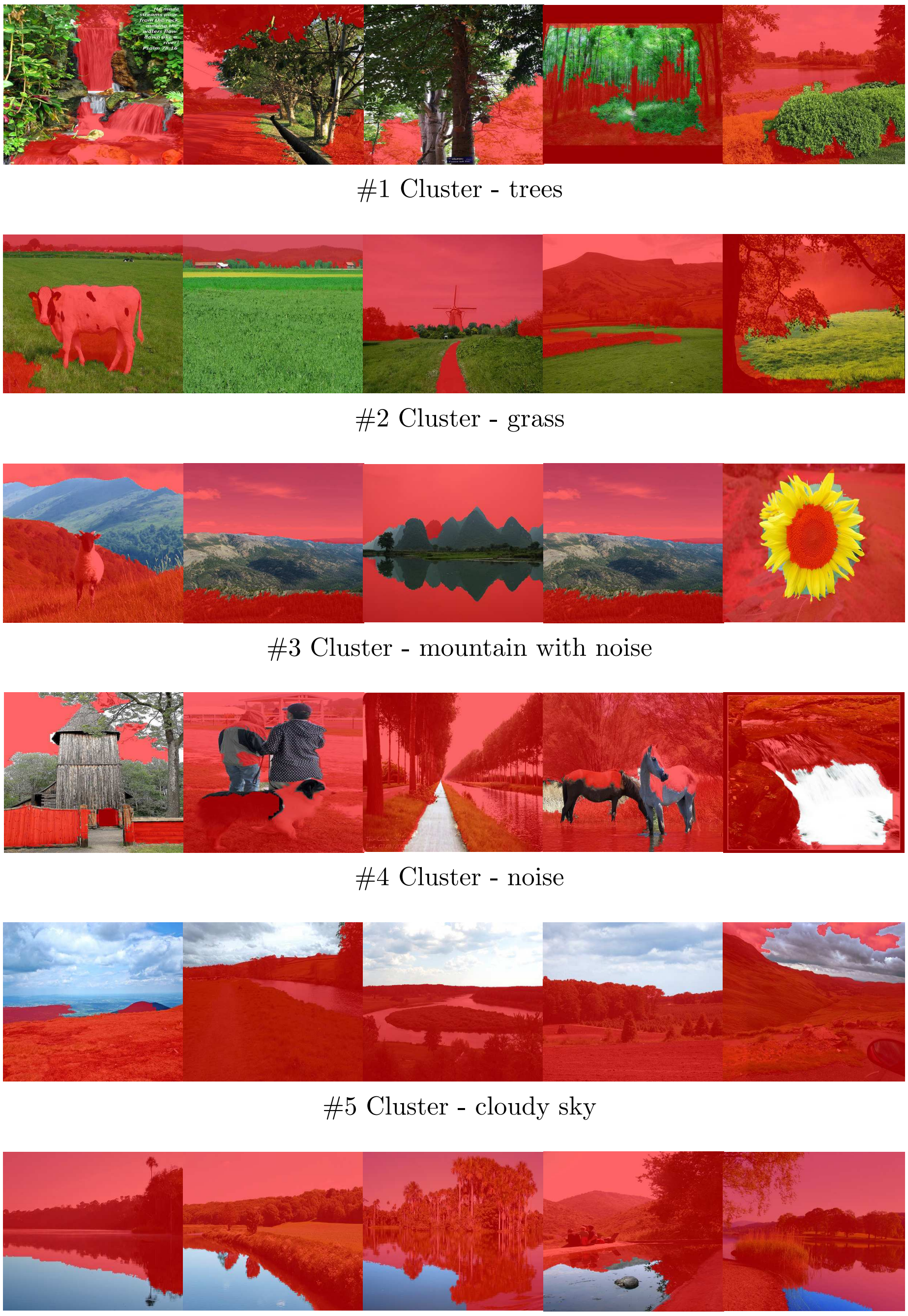
limited, it is clear that the regions of the object that appears more frequently in the dataset (*tree* in this case) are gathered in the most populated cluster.

5.5. Comparison with existing methods

Our goal in the previous experiments was to highlight the potential of social media to serve as the source of training samples for object recognition models. Thus, we have focused on the relative loss in performance that results from the use of leveraged rather than manually annotated training samples, and not on the absolute performance values of the developed models. However, in order to provide an indicative measure of the loss in performance that we suffer when compared with other existing works in the literature, we calculate the classification rate (i.e. number of correctly classified cases divided by the total number of correct cases) of our framework for the 21 objects of MSRC. Then, we compare the results with two methods [37], [38] that are known to deliver state of the art performance on this dataset. Textonboost [37] uses conditional random fields to obtain accurate image segmentation and is based on textons, which jointly model shape and texture. The combination of Markov Random Fields (MRF) and aspect models is the approach followed in [38] in order to produce aspect-based spatial field models for object detection. Note that the reported classification rates are not directly comparable since the methods are not relying on the same set of visual features, the training/test split is likely to be different and the results are reported at different level (in [37] at pixel level, in [38] at the level of 20x20 image patches, and in our case at the level of arbitrary shaped segments which are extracted by an automatic segmentation algorithm). However, the comparison of these methods allows us to make some useful conclusions about the trade-off between the annotation cost for training and the efficiency of the developed models. Table 6 summarizes the classification rates per object for each method.

On average, the accuracy obtained from our approach (45%) is inferior to the one obtained from PLSA-MRF/I (50%) which is again inferior to the accuracy obtained from Textonboost (58%). The performance scores obtained by the three methods are ranked proportionally to the amount of annotation effort required to train their models. Indeed, Textonboost [37] requires strongly annotated images that can only be produced manually, the PLSA-MRF/I algorithmic version of [38], requires weakly but noise-free annotated images the generation of which typically involves light human effort, and our framework operates on weakly but noisy annotated images that can be automatically collected from social sites at no cost.

The costless nature of our approach motivated the execution of two additional experiments that are essentially variations of our original approach, mixing manually labeled data from MSRC and noisy data from flickr. More specifically, the first variation Prop.Fram./M-F/W mixes MSRC and flickr data at the level of images. Initially, the strong region-to-label associations provided by MSRC are relaxed to become weak associations of the form image-to-label(s). Then, these weakly annotated MSRC images are mixed with images from flickr and the proposed framework is applied on the mixed set of images. Finally, the



#1 Cluster - trees

#2 Cluster - grass

#3 Cluster - mountain with noise

#4 Cluster - noise

#5 Cluster - cloudy sky

#9 Cluster - water

Figure 9: Indicative regions from the clusters generated by applying our approach for the object *tree*. The regions that are not covered in red are the ones that have been assigned to the corresponding cluster.

Table 6: Comparing with existing methods in object detection. The reported scores are the classification rates (i.e. number of correctly classified cases divided by the total number of correct cases) per object for each method.

	Building	Grass	Tree	Cow	Sheep	Sky	Aeroplane	Water	Face	Car	Bicycle	Flower	Sign	Bird	Book	Chair	Road	Cat	Dog	Body	Boat	Average
Prop. Framework	87	9	65	45	45	14	29	53	56	12	75	88	27	30	25	50	44	59	71	29	41	45
PLSA-MRF/I [38]	45	64	**71**	75	**74**	**86**	81	47	1	**73**	55	88	6	6	**63**	18	**80**	27	26	**55**	8	50
Prop.Fram./M-F/W	**83**	**72**	69	**91**	70	1	**87**	**53**	**33**	12	**87**	**100**	**47**	**79**	53	**47**	55	**33**	**67**	11	**61**	**57**
Textonboost [37]	62	**98**	**86**	58	50	**83**	**60**	**53**	**74**	**63**	75	63	35	19	**92**	15	**86**	54	19	**62**	7	58
Prop.Fram./M-F/S	**63**	67	76	**73**	**70**	51	27	47	67	17	**94**	**100**	**53**	**47**	59	**47**	68	**92**	**73**	59	**55**	**62**

samples used for training the object recognition models consist of the regions belonging to the most populated of the clusters generated from the mixed set. The Prop.Fram./M-F/W variation is directly compared with PLSA-MRF/I [38] since they use the MSRC annotations in the same way. The second variation Prop.Fram./M-F/S mixes MSRC and flickr data at the level of regions. The samples used for training the object recognition models consist of the strongly annotated regions from MSRC plus the regions belonging to the most populated of the clusters generated from flickr data. The Prop.Fram./M-F/S variation is directly compared with Textonboost [37] since they use the MSRC annotations in the same way. Table 6 shows that both variations of our approach, mixing MSRC and flickr data, outperform their directly comparable state-of-the art approaches. In the case of Prop.Fram./M-F/W the obtained average accuracy (57%) outperforms PLSA-MRF/I by 7%, while in the case of Prop.Fram./M-F/S the obtained average accuracy (62%) outperforms Textonboost by 4%.

6. Discussion of the results & Future Work

In this manuscript we have shown that the collective knowledge encoded in social media can be successfully used to remove the need for close human supervision when training object detectors. The experimental results have demonstrated that although the performance of the detectors trained using leveraged social media is inferior to the one achieved by manually trained detectors, there are cases where the gain in effort compensates for the small loss in performance. In addition, we have seen that by increasing the number of utilized images we manage to improve the performance of the generated detectors, advocating the potential of social media to facilitate the creation of reliable and effective object detectors. The value of social media was also advocated by the experiments showing that when mixing manually labeled and effortlessly obtained flickr data, we manage to outperform the state-of-the-art approaches relying solely on manually labeled samples. Finally, despite the fact that there will always be a strong dependence between the discriminative power of the employed feature space and the efficiency of the proposed approach in selecting the appropriate set of training samples, our analysis has shown that we can maximize the probability of success by using large volumes of user contributed content. Our plans for future work include the investigation of techniques that will allow us to make better use of the tag information space when selecting the image set emphasizing on a particular object. Alternative methods for selecting the appropriate cluster or even merging some of the clusters to create a more suitable training set, also fall within our intentions for improving the proposed method.

Acknowledgment

This work was supported by the WeKnowIt and Glocal projects, partially funded by the EC, under contract numbers FP7-215453 and FP7-248984 respectively.

References

[1] F.-F. Li, R. Fergus, P. Perona, One-shot learning of object categories, IEEE Trans. Pattern Anal. Mach. Intell. 28 (4) (2006) 594–611.

[2] J. Li, J. Z. Wang, Real-time computerized annotation of pictures, IEEE Trans. Pattern Anal. Mach. Intell. 30 (6) (2008) 985–1002.

[3] A. Bordes, New algorithms for large-scale support vector machines, Ph.D. thesis, pour obtenir le Grade de Docteur en Sciences de lUniversite Paris VI – Pierre et Marie Curie (2010).

[4] A. Torralba, R. Fergus, W. T. Freeman, 80 million tiny images: A large data set for nonparametric object and scene recognition, IEEE Transactions on Pattern Analysis and Machine Intelligence 30 (2008) 1958–1970.

[5] G. Carneiro, A. B. Chan, P. J. Moreno, N. Vasconcelos, Supervised learning of semantic classes for image annotation and retrieval, IEEE Trans. Pattern Anal. Mach. Intell. 29 (3) (2007) 394–410.

[6] T. G. Dietterich, R. H. Lathrop, T. Lozano-Pérez, Solving the multiple instance problem with axis-parallel rectangles, Artif. Intell. 89 (1-2) (1997) 31–71.

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

[8] L. von Ahn, L. Dabbish, Labeling images with a computer game, in: SIGCHI, ACM, 2004, pp. 319–326.

[9] L. von Ahn, R. Liu, M. Blum, Peekaboom: a game for locating objects in images, in: SIGCHI, ACM, 2006, pp. 55–64.

[10] P. A. Viola, M. J. Jones, Rapid object detection using a boosted cascade of simple features, in: CVPR (1), 2001, pp. 511–518.

[11] B. Leibe, A. Leonardis, B. Schiele, An implicit shape model for combined object categorization and segmentation, in: Toward Category-Level Object Recognition, 2006, pp. 508–524.

[12] K. K. Sung, T. Poggio, Example-based learning for view-based human face detection, IEEE Trans. Pattern Anal. Mach. Intell. 20 (1) (1998) 39–51.

[13] J. Sivic, B. C. Russell, A. A. Efros, A. Zisserman, W. T. Freeman, Discovering objects and their localization in images, in: ICCV, 2005, pp. 370–377.

[14] T. Quack, B. Leibe, L. J. V. Gool, World-scale mining of objects and events from community photo collections, in: CIVR, 2008, pp. 47–56.

[15] R. Fergus, F.-F. Li, P. Perona, A. Zisserman, Learning object categories from google's image search, in: ICCV, 2005, pp. 1816–1823.

[16] F.-F. Li, P. Perona, C. I. of Technology, A bayesian hierarchical model for learning natural scene categories, in: CVPR (2), 2005, pp. 524–531.

[17] B. C. Russell, W. T. Freeman, A. A. Efros, J. Sivic, A. Zisserman, Using multiple segmentations to discover objects and their extent in image collections, in: CVPR (2), 2006, pp. 1605–1614.

[18] C. Marlow, M. Naaman, D. Boyd, M. Davis, Ht06, tagging paper, taxonomy, flickr, academic article, to read, in: Hypertext, 2006, pp. 31–40.

[19] S. Nikolopoulos, E. Chatzilari, E. Giannakidou, I. Kompatsiaris, Towards fully un-supervised methods for generating object detection classifiers using social data, in: 10th Workshop on Image Analysis for Multimedia Interactive Services, WIAMIS '09., 2009, pp. 230 –233.

[20] E. Chatzilari, S. Nikolopoulos, I. Kompatsiaris, E. Giannakidou, A. Vakali, Leveraging social media for training object detectors, in: 16th International Conference on Digital Signal Processing, DSP '09., 2009, pp. 1 –8.

[21] L. S. Kennedy, S.-F. Chang, I. Kozintsev, To search or to label?: predicting the performance of search-based automatic image classifiers, in: Multimedia Information Retrieval, 2006, pp. 249–258.

[22] T. Cour, B. Sapp, C. Jordan, B. Taskar, Learning from ambiguously labeled images, in: IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'09), 2009.

[23] L. Wu, X.-S. Hua, N. Yu, W.-Y. Ma, S. Li, Flickr distance, in: ACM Multimedia, 2008, pp. 31–40.

[24] Y. Sun, S. Shimada, Y. Taniguchi, A. Kojima, A novel region-based approach to visual concept modeling using web images, in: ACM Multimedia, 2008, pp. 635–638.

[25] T. Tsikrika, C. Diou, A. P. de Vries, A. Delopoulos, Image annotation using clickthrough data, in: 8th ACM International Conference on Image and Video Retrieval, Santorini, Greece, 2009.

[26] L. S. Kennedy, M. Naaman, S. Ahern, R. Nair, T. Rattenbury, How flickr helps us make sense of the world: context and content in community-contributed media collections, in: ACM Multimedia, 2007, pp. 631–640.

[27] K. Barnard, P. Duygulu, D. A. Forsyth, N. de Freitas, D. M. Blei, M. I. Jordan, Matching words and pictures, Journal of Machine Learning Research 3 (2003) 1107–1135.

[28] L.-J. Li, R. Socher, L. Fei-Fei, Towards total scene understanding: Classification, annotation and segmentation in an automatic framework, in: IEEE Conference on Computer Vision and Pattern Recognition, 2009.

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

[30] V. Mezaris, I. Kompatsiaris, M. G. Strintzis, Still image segmentation tools for object-based multimedia applications, IJPRAI 18 (4) (2004) 701–725.

[31] K. van de Sande, T. Gevers, C. Snoek, Evaluating color descriptors for object and scene recognition, IEEE Transactions on Pattern Analysis and Machine Intelligence 32 (9) (2010) 1582 –1596.

[32] J. Zhang, M. Marszalek, S. Lazebnik, C. Schmid, Local features and kernels for classification of texture and object categories: A comprehensive study, Int. J. Comput. Vision 73 (2) (2007) 213–238.

[33] D. G. Lowe, Distinctive image features from scale-invariant keypoints, Int. J. Comput. Vision 60 (2) (2004) 91–110.

[34] B. J. Frey, D. Dueck, Clustering by passing messages between data points, Science 315 (2007) 972–976.

[35] B. Scholkopf, A. Smola, R. Williamson, P. Bartlett, New support vector algorithms, Neural Networks 22 (2000) 1083–1121.

[36] M. Everingham, L. Van Gool, C. K. I. Williams, J. Winn, A. Zisserman, The PASCAL VOC2009 Results.

[37] J. Shotton, J. M. Winn, C. Rother, A. Criminisi, *TextonBoost*: Joint appearance, shape and context modeling for multi-class object recognition and segmentation, in: ECCV (1), 2006, pp. 1–15.

[38] J. Verbeek, B. Triggs, Region classification with markov field aspect models, in: Computer Vision and Pattern Recognition, 2007. CVPR '07. IEEE Conference on, 2007, pp. 1 –8.