# LEVERAGING SOCIAL MEDIA FOR TRAINING OBJECT DETECTORS

*E. Chatzilari, S. Nikolopoulos, I. Kompatsiaris*

Informatics and Telematics Institute
ITI - CERTH
GR-57001, Greece
email:{ehatzi, nikolopo, ikom}@iti.gr

*E. Giannakidou, A. Vakali*

Department of Informatics
Aristotle University
54124 Thessaloniki, Greece
email:{eirgiann, avakali}@csd.auth.gr

## ABSTRACT

The fact that most users tend to tag images emotionally rather than realistically makes social datasets inherently flawed from a computer vision perspective. On the other hand they can be particularly useful due to their social context and their potential to grow arbitrary big. Our work shows how a combination of techniques operating on both tag and visual information spaces, manages to leverage the associated weak annotations and produce region-detail training samples. In this direction we make some theoretical observations relating the robustness of the resulting models, the accuracy of the analysis algorithms and the amount of processed data. Experimental evaluation performed against manually trained object detectors reveals the strengths and weaknesses of our approach.

***Index Terms***— Social media, object detection, weak annotations, Flickr

## 1. INTRODUCTION

Semantic object detection is one of the most useful operations performed by human visual system and constitute an exciting problem for computer vision scientists. Robust models capable of capturing the diversity of an object's form and appearance, need to be learned from a large number of highly descriptive training examples. However, current literature had showed us that such examples are not existent and therefore very expensive to obtain.

In this perspective, semantic object detection can be viewed as a problem of either supervised [1], [2], [3], [4] or unsupervised learning [5], [6], [7], [8], [9], [10], [11], [12], [13], [14]. In the first case a classifier is trained to recognize an object category e.g., a face [1], [4], a building [2] or a car [3], using a set of hand-labeled training images. The drawback of these schemes is that they require a large amount of strongly annotated images, the generation of which is a laborious and time consuming procedure. To tackle this issue, the methods resorting to unsupervised learning attempt to solve the problem by using weakly annotated training examples. In this case, the idea is to estimate a joint probability distribution on a space of semantic labels and visual characteristics.

A high number of diverse ideas has been proposed in the literature for this purpose. In [15], [13] the problem is viewed as a top-down image segmentation procedure where the recognition of visual objects is incorporated as an intermediate step of segmentation. Aspect models like probabilistic Latent Semantic Analysis (pLSA) [7], [16] and Latent Dirichlet Allocation (LDA) [17], [18] have been used with weakly annotated datasets to estimate the joint probabilities between semantic labels and visual features. In some cases these models are coupled with conditional random fields [12], [19] to incorporate spatial and hierarchical information originating from context, or use Probabilistic Graphical Models (PGM) [11] to consider the role of structure within the detection process. Other techniques, that also rely on observations statistics to estimate these joint probabilities, include [10], [8] where Expectation Maximization is employed, and [9] where stochastic processes are used. Some pioneer work in this direction has been presented in [5] where much information is learned from a handful of images by taking advantage of knowledge coming from previously learned categories, and [6] where the advantages of supervised and un-supervised approaches are combined by solving a multiclass classification problem.

This work concentrates on social media and their potential to serve as the training examples of an object detection scheme. Social sites like flickr, accommodate image corpora that are being populated with hundreds of user tagged images on a daily basis. We are interested on whether such corpora can be leveraged to facilitate the robust estimation of models. By looking at the literature above, we realize that most of the proposed schemes have been tested on purpose specific datasets. For instance [5], [18], [7], [10], are evaluated using the Caltech dataset which is a set of images manually organized in categories, while [16], [20] operates on images collected from the web using key-word based search. Similarly [6], [9], [8], [21], [13] use the Corel dataset, which is a set of images annotated with realistic tags, while [11], [12], [18] operate on Microsoft Research database which is a set of strongly annotated images. Few are the attempts where object detection schemes exploit social data, as in [22], [23],

[14] where photo collections obtained from flickr are used for this purpose. The advantage of using social sites like flickr is that we can obtain a high number of images without spending much effort or time. Consequently, as opposed to supervised approaches, there is no limitation on the types of objects that can be trained, since social sites accommodate images depicting a huge variety of objects.

Our work bears many similarities with [8], where segmentation, visual feature extraction and region clustering are applied on a set of tagged images to facilitate object detectors' training. However, we examine from both theoretical and experimental perspective, the way the robustness of the generated detectors is affected by the relation associating the accuracy of the image analysis algorithms with the size of the processed dataset.

## 2. FRAMEWORK DESCRIPTION

The goal of our framework is to start from a set of user tagged images, obtained from social sites, and automatically extract training examples, suitable for learning an object detection model. Social media processing, segmentation, visual features extraction, clustering and machine learning constitute the analysis components incorporated by our framework, as shown in Fig. 1. We mainly focus on the components of social media processing and clustering, with the intention to tackle the reduced amount of supervision foreseen by our framework and the low quality of tags contributed by the social users. In

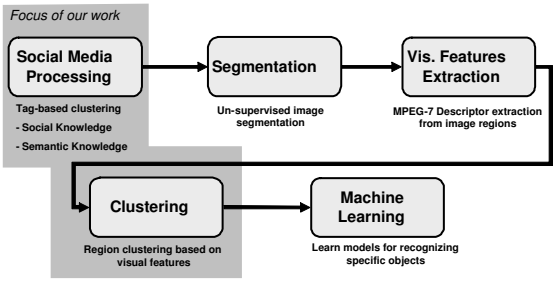



**Fig. 1**. Analysis components incorporate by our framework.

our framework, we identify six analysis steps that are applied consecutively on a set of user tagged images: a) Cluster images using their tags and acquire image groups each one emphasizing on a particular topic. The linguistic description of this topic is usually reflected in the most frequent tag. b) Pick an image group so as its most frequent tag to conceptually relate with the object of interest. c) Segment all images in the selected image group into regions that are likely to represent objects. d) Extract the visual features of these regions with the expectation that all regions representing the same object will share a relative high amount of common characteristics. e) Perform feature-based clustering so as to create groups of similar regions. We anticipate that the majority of regions representing the object of interest will be gathered in one of the clusters, pushing all irrelevant regions to the others. f) Use the visual features extracted from the regions belonging to the cluster representing the object of interest, to train a machine learning-based object detector.

Although, there are issues to be addressed such as a) how to derive image groups with an increased level of semantic coherence, b) how to determine the number of clusters for the feature-based region clustering procedure, and c) how to select the cluster containing the regions depicting the object of interest; our great advantage relies on the social aspect of the analyzed dataset and its potential to grow particularly large. It has been shown [24] that the majority of users tend to contribute similar tags when faced with similar type of visual content. This is attributed to the common background that most users share and is expected to lead the prevailing concepts in tag and visual information space to convergence. Based on this assumption we adopt the following solutions in order to fully automate the aforementioned process. Semantically coherent groups of images are generated using a tag-based clustering approach that incorporates both social and semantic knowledge, detailed in Section 4. The number of clusters for the feature-based region clustering step is determined in an un-supervised manner by employing the Maximin algorithm, tuned using cross validation as described in Section 4. Finally, the most populated of the generated region-clusters is chosen to provide the machine learning algorithm with the necessary training examples, as explained in Section 3.3.

It is evident that selecting the most populated of the generated clusters would certainly constitute the appropriate choice, if all analysis components of computer vision (i.e,. segmentation, discrimination by visual features) worked perfectly. However, since current literature has shown us that this is not true, we examine how the size of the analyzed dataset affects the legitimate error space of the analysis modules, for letting the aforementioned cluster selection to be the appropriate choice. The following section investigates the issue from a theoretical perspective.

## 3. THEORETICAL ANALYSIS

### 3.1. Preliminary Definitions & Conventions

Table 1 summarizes the notations used throughout the presented analysis. Given the diversity characterizing an object's form and appearance, both segmentation and visual feature extraction are likely to introduce errors in the analysis pipeline of Fig. 1. However, if we consider that our final goal is to create clusters of image regions depicting the object of interest, we can accept that all these errors are eventually reflected on the efficiency of the clustering procedure. Thus, we will make the convention that the clustering error incorporates all these sources of error.

**Table 1**. Legend of Introduced Notations

| Symbol | Definition |
|---|---|
| $S$ | The complete social dataset |
| $N$ | The number of images in $S$ |
| $L$ | A particular topic |
| $S^L$ | An image group, subset of $S$ that emphasizes on topic $L$ |
| $n$ | The number of images in $S^L$ |
| $I_q$ | An image from $S$ |
| $R_{I_q} = \{r_i^{I_q}, i = 1, \ldots, m\}$ | Segments identified in image $I_q$ |
| $f_d(r_i^{I_q}) = \{f_i, i = 1, \ldots, z\}$ | Visual descriptor extracted from a region $r_i^{I_q}$ |
| $T_{I_q}$ | Set of tags associated with image $I_q$ |
| $C = \{c_i, i = 1, \ldots, t\}$ | Set of objects that appear in an image group $S^L$ |
| $W = \{w_i, i = 1, \ldots, o\}$ | Set of clusters created by the feature-based clustering algorithm |
| $p_{c_i}$ | probability that social media processing draws from $S$ an image depicting $c_i$ |

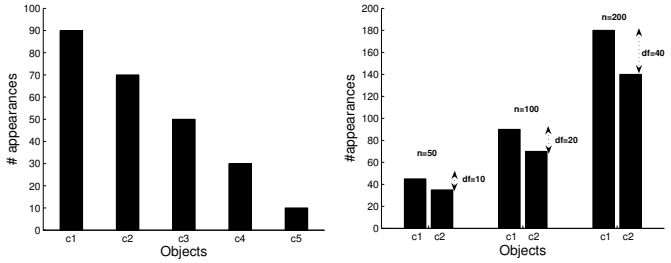
Moreover, we will assume that there is a one-to-one relation between an image and an object (i.e., we do not consider cases where the same object is depicted in two different locations of the image).

### 3.2. Social Media processing

The goal of social media processing is to cluster images into semantically coherent groups, $S^L \subset S$. We are interested in the frequency distribution of objects $c_i \in C$ appearing in $S^L$ based on their frequency rank. If we focus on a single image group $S^L$, we can view this process as the act of populating $S^L$ with images selected from a large dataset $S$ using certain criteria, (see Section 4). In this case, the number of images in $S^L$ that depict the object $c_i$, 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 $p_{c_i}$. Considering that an image depicts more than one concepts we can say that the probabilities $p_{c_i}, \forall c_i \in C$ are independent from each other and they depend on the nature of the dataset. Given that $S$ is sufficiently large, drawing an image from this dataset can be considered as an independent trial. Thus, the number of times an object $c_i \in C$ appears in $S^L$ can be expressed by a random variable $K$ following the binomial distribution with probability $p_{c_i}$. In this way we can use the corresponding probability mass function ($Pr(K = k)$) depicted in eq. (1), to estimate the probability that $S^L$ contains $k$ images depicting $c_i$:

$$Pr(K = k) = \binom{n}{k} p^k (1-p)^{n-k} \quad (1)$$

Moreover, since the social media processing aims at creating groups of images emphasizing on a particular topic, we can assume that there will be an object $c_1$ that is drawn with probability $p_{c_1}$ higher than $p_{c_2}$, which is the probability that $c_2$ is drawn, and so forth for the remaining $c_i \in C$. This assumption is experimentally verified in Section 5.1 where the tag-frequency histograms of different image groups are measured. Given the above, we can use the expected value (E(K)) of a random variable following the binomial distribution (eq.(2)) to estimate the number of times an object $c_i \in C$ will appear in $S^L$, if its drawn from the initial dataset $S$ with probability $p_{c_i}$. This is actually the value of $k$ maximizing the corresponding probability mass function.




(a) (b)

**Fig. 2**. a) Distribution of #appearances of the objects $C$ in $S^L$, 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 populations between $c_1$, $c_2$, using different values of $n$

$$E(K) = np \quad (2)$$

In this way, we are able to estimate how the number of appearances (#appearances) of objects $c_i \in C$ are distributed in $S^L$, based on their frequency rank. Fig. 2(a) show how such a distribution would look like given that $(p_{c_1} > p_{c_2} > \ldots)$. Based on this distribution and given the fact that as $N$ increases $n$ will also increase, we examine how the population of the generated region clusters relates with the clustering error space and $n$.

### 3.3. Clustering

The goal of feature-based region clustering is to group together regions representing the same object. Ideally, 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 $w_1$ contains all regions depicting the most frequently appearing object $c_1$. However, there is very little chance that we will get perfectly solid clusters, each one containing regions representing a single object.

Nevertheless, given the fact that object models can be robustly learned even from rather noisy training sets, we seek to detect the point where $w_1$, which is the cluster containing the majority of regions depicting $c_1$, will stop be the most populated cluster and therefore not selected by our framework to train $c_1$. Clearly, this depends on the clustering error and

the difference in population separating the first two most frequently appearing objects $c_1, c_2 \in C$. This difference depends on $p_{c_1}$, $p_{c_2}$ and increases proportionally to $n$ as derived from eq. (2) and shown in Fig. 2(b). Here, we work under the assumption that it is more likely for the second most highly ranked cluster $w_2$ to become more populated than $w_1$ as the clustering error increases. Thus, we only consider $c_1$ and $c_2$ and examine how their difference in population relates with $n$ and clustering performance.

In order to do this we make an initial assignment of objects to clusters based on their ranks $c_i \rightarrow w_i$, and express clustering error using the notations of Table 2.

**Table 2**. Notations for Clustering

| Symbol | Definition |
|---|---|
| $TC_i$ | Number of regions depicting object $c_i$ |
| $tc_i$ | Number of regions depicting $c_i$, correctly assigned to cluster $w_i$ |
| $Pop_i$ | Population of cluster $w_i$ |
| $FP_i$ | False positives of $w_i$ with respect to $c_i$ |
| $FN_i$ | False negatives of $w_i$ with respect to $c_i$ |
| $DR_i = FP_i - FN_i$ | Displacement rate of $w_i$, with respect to $c_i$ |

Given the above, $FP_i = Pop_i - tc_i$ and $FN_i = TC_i - tc_i$. By substituting $tc_i$ we have $Pop_i = TC_i + FP_i - FN_i$. However, $TC_i$ is actually the number of times the object $c_i$ appears in $S^L$ (#appearances) and according to eq. (2) we have $TC_i = np_i$. Now, $w_1$ will be selected by our framework for learning $c_1$ as long as:

$$\begin{gathered} Pop_1 - Pop_2 > 0 \quad \Rightarrow \\ TC_1 - TC_2 + (FP_1 - FN_1) - (FP_2 - FN_2) > 0 \Rightarrow \\ n > \frac{DR_2 - DR_1}{p_{c_1} - p_{c_2}} \end{gathered} \tag{3}$$

The displacement rate $DR_i$ shows how the $Pop_i$ of cluster $w_i$ modifies according to the clustering error and with respect to the ideal case where this error is zero. Positive values of $DR_i$ indicates leakages in $w_i$ population, while negative values indicate inflows. Using eq. (3) we 3D plot in Fig. 3 the space where $Pop_1 - Pop_2 > 0$. Every horizontal slice of this volume corresponds to the legitimate values of $DR_1$ and $DR_2$ for a certain value of $n$. As n increases, the surface of the corresponding slices increases also and thus the legitimate error space for clustering increases too.

## 4. IMPLEMENTING THE FRAMEWORK

**Social media processing:** For acquiring image groups with an increased amount of semantic coherence we adopted the SEMSOC approach introduced by Giannakidou et. al. in [25]. In this work, an unsupervised model for efficient and scalable mining of multimedia social-related data is presented. The

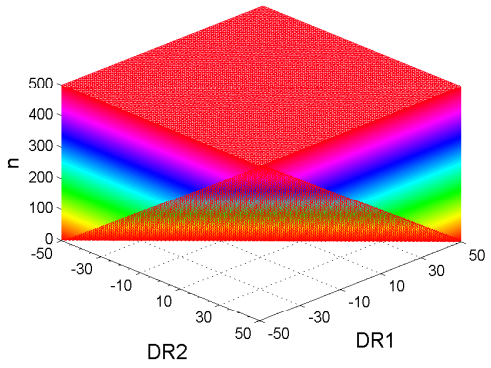



**Fig. 3**. Space in which $w_1$ remains the most populated of the generated clusters, derived from eq. (3)

reason for adopting this approach is to overcome the limitations that characterize collaborative tagging systems such as tag spamming, tag ambiguity, tag synonymy and granularity variation, and increase the semantic coherence of the generated groups. Each group emphasizes on a particular topic and the set of its containing tags reflects the way users perceive it. SEMSOC manages to create meaningful groups by jointly considering social and semantic features. Its outcome is a set of image groups $S^{L_i} \subset S, \quad i = 1, \ldots, m$ where $L_i$ is an indicator of the emphasized topic and $m$ is the number of created clusters. In this case the number of clusters is determined empirically, as described in [25].

Every image $I_q$ has an associated set of tags $T_{I_q}$. We choose the image group $S^{L_i}$ where its most frequent tag conceptually relates with the object that we want to detect. In this way, we obtain a semantically coherent group of images the majority of which is expected to depict the object of interest.

**Segmentation:** Segmentation is applied to all images in $S^L$ 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 [26]. The output of this algorithm is a set of segments $R_{I_q} = \{r_i^{I_q}, i = 1, \ldots, m\}$, which in the ideal case correspond to meaningful objects, $c_i \in C$.

**Visual descriptors:** Seven descriptors proposed by MPEG-7 [27] capturing different aspects of color, texture and shape were used. These descriptors namely $mpeg7$={Dominant Color (DC), Color Layout (CL), Color Structure (CS), Scalable Color (SC), Edge Histogram (EH), Homogeneous Texture (HT), Region Shape (RS)} were extracted $\forall r_i^{I_q} \in R_{I_q}$ and $\forall I_q \in S^L$. Different descriptors' combinations were composed by concatenating their normalized values on a single vector, $f_d(r_i^{I_q}) = \{f_i, i = 1, \ldots, z\}$. In this case, $d \in mpeg7$ determines the descriptors' combination and $z$ the dimensionality of the feature space, see Section 5.3. The concatenation approach was used only for training the object

models using SVMs.

**Clustering:** For performing feature-based region clustering we applied k-means on all extracted feature vectors $f_d(r_i^{I_q}), \quad \forall r_i^{I_q} \in R_{Iq}$ and $\forall I_q \in S^L$. For calculating the distance between two regions we have used the functions presented in [27] by independently measuring the distance in each feature space and summing their normalized values. However, the problem that arises from the use of a parametric clustering algorithm like k-means is that a) the number of the clusters must be known in advance, and b) its performance is sensitive to the initial positions of the cluster centers. In order to overcome these problems, we employed the Maximin algorithm as described in [26], both for selecting the number of clusters and estimating the initial positions of their centers.

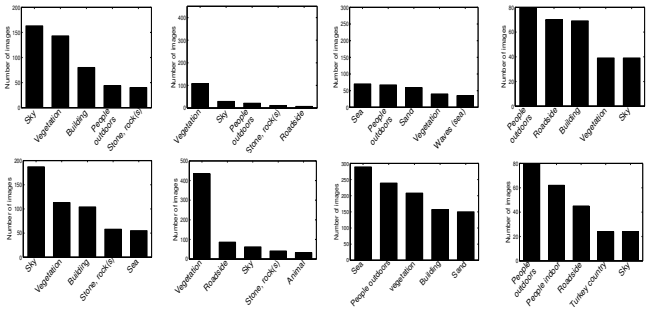
**Learning model parameters:** Support Vector Machines (SVMs) [28] were chosen for generating the object detection models, due to their ability in coping efficiently with high-dimensionality pattern recognition problems. All feature vectors assigned to the most populated of the created clusters are used as positive examples for training a binary classifier. Negative examples are chosen arbitrary from the remaining dataset. Tuning arguments include the selection of Gaussian radial basis kernel and the adoption of a brute force strategy for selecting the kernel parameters.

## 5. EXPERIMENTAL STUDY

The goal of our experimental study is twofold. On the one hand, we wanted to get an experimental insight on the error introduced by the analysis algorithms and check whether our theoretical claims stand. On the other hand, we aimed at comparing the quality of object models trained using the proposed framework, against the ones trained using high quality, manually provided, region-detail annotations. Experiments necessary for tuning some of the employed algorithms are also presented.

To carry out our experiments we utilized three datasets, a strongly annotated dataset constructed manually by asking people to produce region-detail image annotations, and two weakly annotated social datasets obtained from Flickr. For the first dataset $S^M$, a lexicon of 7 objects $C^M$={Vegetation, Rock, Sky, Person, Boat, Sand, Sea}, was used to strongly annotate 536 images at region-detail. The output of this process was to record relations associating an image segment $r_i^{I_q}$, identified automatically by the segmentation algorithm, with an object from $C^M$. On the other hand, two datasets from Flickr were crawled using the wget[1] utility and Flickr API facilities. The first dataset $S^{3K}$ consists of 3000 images depicting among others $C^{3K}$= {cityscape, seaside, mountain, roadside, landscape, sport-side}, while the second one $S^{10K}$ consists of 10000 images, mostly related to $C^{10K}$={jaguar, turkey, apple, bush, sea, city, vegetation, roadside, rock, tennis}.

[1] wget: http://www.gnu.org/software/wget




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

**Fig. 4**. Distribution of objects' appearance in an image group $S^L$, obtained from $S^{3K}$ (upper line) and $S^{10K}$ (lower line)

For the purposes of our experimental study and after applying SEMSOC [25] on both $S^{3K}$ and $S^{10K}$, we ended up with four object categories $C^{bench}$={sky, sea, person, vegetation}, that exhibited significant presence in all three datasets. These object categories served as benchmarks for comparing the quality of different models.

### 5.1. Social media processing

As claimed in Section 3.2, we expect the gap between the number of appearances of the first ($c_1$) and second ($c_2$) most highly ranked objects of $C$, to broaden as the volume of the analyzed dataset increases. In order to verify this experimentally, we plot the distribution of objects' #appearances in an image group $S^L$. Each of the bar diagrams depicted in Fig. 4, describes the distribution of objects' #appearances inside an image group $S^L$, as evaluated by human subjects. The image groups are created by applying SEMSOC on both $S^{3k}$ and $S^{10K}$, and selecting the groups emphasizing in one of the benchmark object categories . It is clear that as we move from $S^{3k}$ to $S^{10K}$ the gap between the number of images depicting $c_1$ and $c_2$, increases in all four cases.

### 5.2. Tuning Maximin

As mentioned before, Maximin is used to decide the number of clusters and generate an initial estimation of the cluster centers, to be used by K-means. However, Maximin largely depends on a parameter called $\gamma$, that specifies the threshold according to which new clusters are created or not. The purpose of this experiment was on the one hand to optimally tune $\gamma$, in order to use it for all subsequent experiments, and on the other hand to check whether this value deviates substantially as the training examples and the object category vary. This is to ensure that the tuned value can be safely used under various contexts. For this purpose we use $S^M$ and apply 10-fold cross validation, for all available objects of $C^M$ and all pos-

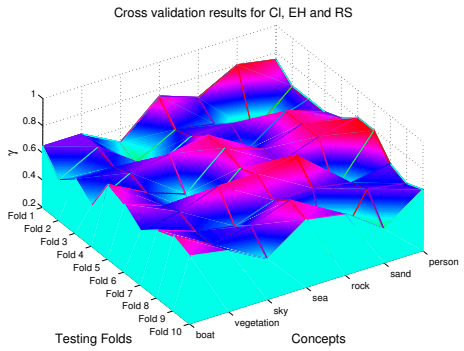



**Fig. 5**. Cross-validation results for descriptor combination CL, EH and RS ($\gamma_{avg} = 0.633$, $\gamma_{min} = 0.4$ ,$\gamma_{max} = 0.8$)

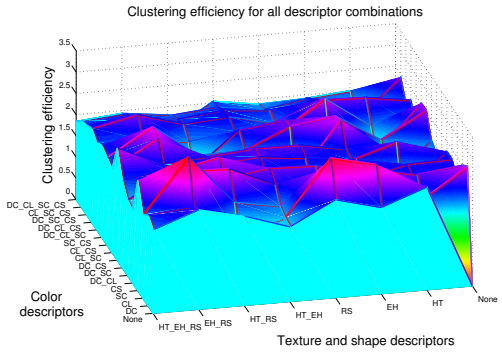



**Fig. 6**. Clustering efficiency for all combinations of MPEG-7 descriptors

sible descriptor combinations $d \in mpeg7$. Given that $S^M$ is strongly annotated, clustering efficiency can be measured explicitly using typical classification metrics (i.e. F-Measure).

For every object $c_i \in C^M$, the subset of images $S^{c_i}$ depicting this object is selected using the manually provided annotations. Images are segmented and visual features are extracted. Subsequently, the regions are divided in 10 folds, using each time one fold for "testing" and 9 for "training". For every run of the experiment we vary the value of parameter $\gamma$ within [0.2 0.96] using steps of 0.05. For each value of $\gamma$, the number of clusters determined by applying Maximin on the "training" folds, is used to perform clustering using k-means in the regions belonging to the "testing" fold. The F-measure of the most populated cluster $w_1$ is calculated with respect to the most frequently appearing object $c_1$ in $S^{c_i}$. Given that for each value of $\gamma$ we can measure the clustering efficiency $F_{i,j,\gamma}$, on the basis of a $S^{c_i}$ and fold $j$, we are able to determine the optimal value of $\gamma$ as $\gamma_{opt} = \arg\max_{\gamma}(F_{i,j,\gamma})$.

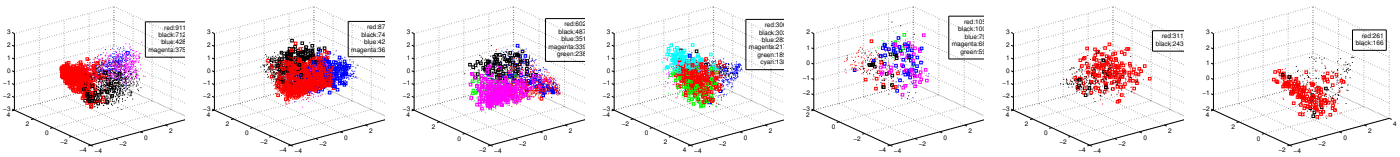
Finally, the average of the optimal values among folds and objects ($\gamma = 0.633$) was used for the remaining of our experiments. Fig. 5 is a 3D plot summarizing the aforementioned results for the feature space derived by combining CL, EH and RS. It is clear that the optimal values of $\gamma$ does not deviate substantially as the object category and the folds vary. Similar observations were made for all other combinations of MPEG-7 descriptors, the results of which are not included in this manuscript due to lack of space.

### 5.3. Optimal Feature Space

Visual descriptors determine the attributes by which a model tries to capture an object's form an appearance. After tuning the Maximin algorithm for all different combinations of MPEG-7 descriptors, we utilized the strongly annotated dataset $S^M$ to determine the optimal feature space, in terms of clustering efficiency. As in the previous case $\forall c_i \in C^M$, a subset $S^{c_i} \subset S^M$ of images depicting $c_i$ was selected to serve as the image group. For each of those image groups, clustering efficiency was measured by calculating the F-Measure of the most populated cluster $w_1$, with respect to the most highly ranked object $c_1$ in $S^{c_i}$. Finally, these values were summed over all different objects $c_i \in C^M$, to form a cumulative f-measure metric assessing the clustering efficiency for a certain combination of visual descriptors (i.e., feature space). Fig. 6 summarizes the results by plotting in the $z$-axis the value of cumulative f-measure obtained for the feature space determined by combining the descriptors indicated by the $x-$ and $y-$axis. We can see that clustering efficiency maximizes when CL, EH and RS are combined. This experimental observation is also compliant with human intuition since color, texture and shape are considered important attributes of visual perception for discriminating between different objects. The feature space determined by $d = \{CL, EH, RS\}$ was utilized for the remaining of our experiments.

### 5.4. Cluster Selection

Having tuned the Maximin algorithm and selected the optimal feature space, the purpose of this experiment was to validate using real data our theoretical claim that the most populated cluster contains the majority of regions depicting the object of interest. In order to do so, $\forall c_i \in C^M$ we obtain $S^{c_i} \subset S^M$ and apply k-means clustering using $\gamma = 0.633$ and $d = \{CL, EH, RS\}$. In Fig. 7 we visualize the way regions are distributed among the clusters by projecting their feature vectors in three dimensions using PCA (Principal Component Analysis). The regions depicting the object of interest $c_i$ are marked in squares, while the other regions are marked in dots. Color code indicating a cluster's rank according to their population (i.e., red: 1st, black: 2nd, blue: 3rd, magenta: 4rth, green: 5th, cyan: 6th) is used. Thus, in the ideal case all squares should be painted red and all dots should be colored differently. Squares being painted in colors other than red, indicate false negatives and dots painted in red indicate false positives. We can see that our claim is validated in 5 (i.e., *sky, sea, person, vegetation and rock*) out of 7 examined cases.



(a) Sky (b) Sea (c) sand (d) Person (e) Boat (f) Vegetation (g) Rock

**Fig. 7**. Regions distribution amongst clusters. This Figure is best viewed in color with magnification.

The visual diversity of objects *boat* and *sand*, causes segmentation and visual feature extraction to introduce significant error, that prevents clustering from gathering the regions of interest into the most populated cluster.

### 5.5. Object models comparison

Assessing the quality of object detection models, generated using both the proposed framework and the manually provided region-detail annotations, is the purpose of this experiment. Additionally, we want to validate our claim that as the scale of the utilized social dataset increases, the error allowed to be introduced by the analysis components increases also and the models produced by the proposed framework are more robust. With this intention, we generated object models using $S^M$, $S^{3K}$ and $S^{10K}$ for the object categories of $C^{bench}$. For each object $c_i \in C^{bench}$ one model was trained in a fully supervised manner using the strong annotations of $S^M$, and two models were trained without supervision using the weak annotations of $S^{3K}$ and $S^{10K}$ and the proposed framework. In order to evaluate the performance of these models, we utilized a portion (i.e., 268 images) of the strongly annotated dataset $S^M_{test} \subset S^M$ as ground truth, not used during training.

By looking at the bar diagram of Fig. 8, we note that models trained in a fully supervised manner perform optimally in all cases. However, the performance achieved by the models trained without supervision, although inferior, is still satisfactory, especially if we take into account the time and effort gained using the proposed framework. Another interesting observation concerns the improvement in performance achieved in all cases, between the models trained using $S^{10K}$ and $S^{3K}$, respectively. This tendency verifies our claim that there is a relation between the size of the utilized social dataset and the robustness of the generated models.

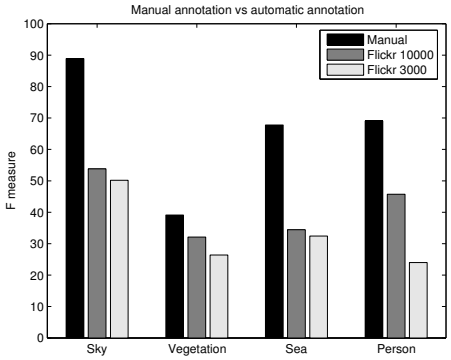



**Fig. 8**. Comparing the quality of different object models

## 6. CONCLUSIONS & FUTURE WORK

Although the quality of the object models trained using the proposed unsupervised technique is still inferior from the one achieved using supervised approaches, we have shown that under certain circumstances social data can be effectively used to learn the parameters modeling an object's form and appearance. Moreover, as it is reasonable to expect that the proposed framework would not graciously scale to every possible object category, the social aspect of user contributed content and its potential to scale in terms of content diversity and size, advocates it's use for the type of objects that appear frequently in social context. Our plans for future work include exploiting more of the user contributed information (e.g., Flickr groups) for obtaining suitable (from a computer vision perspective) datasets, and the employment of outlier detection techniques for training the models using less noisy region-clusters.

## 7. ACKNOWLEDGMENT

This work was funded by the X-Media project (www.x-media-project.org) sponsored by the European Commission as part of the Information Society Technologies (IST) programme under EC grant number IST-FP6-026978 and the European Community's Seventh Framework Programme FP7/2007-2013 under grant agreement n215453 - WeKnowIt.

## 8. REFERENCES

[1] Paul A. Viola and Michael J. Jones, "Rapid object detection using a boosted cascade of simple features," in *CVPR (1)*, 2001, pp. 511–518.

[2] Yi Li and Linda G. Shapiro, "Consistent line clusters for building recognition in cbir," in *ICPR (3)*, 2002, pp. 952–956.

[3] Bastian Leibe, Ales Leonardis, and Bernt Schiele, "An implicit shape model for combined object categorization and segmentation," in *Toward Category-Level Object Recognition*, 2006, pp. 508–524.

[4] Kah Kay Sung and Tomaso Poggio, "Example-based learning for view-based human face detection," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 20, no. 1, pp. 39–51, 1998.

[5] Fei-Fei Li, Robert Fergus, and Pietro Perona, "One-shot learning of object categories," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 28, no. 4, pp. 594–611, 2006.

[6] Gustavo Carneiro, Antoni B. Chan, Pedro J. Moreno, and Nuno Vasconcelos, "Supervised learning of semantic classes for image annotation and retrieval," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 29, no. 3, pp. 394–410, 2007.

[7] Josef Sivic, Bryan C. Russell, Alexei A. Efros, Andrew Zisserman, and William T. Freeman, "Discovering objects and their localization in images," in *ICCV*, 2005, pp. 370–377.

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

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

[10] Robert Fergus, Pietro Perona, and Andrew Zisserman, "Object class recognition by unsupervised scale-invariant learning," in *CVPR (2)*, 2003, pp. 264–271.

[11] Giuseppe Passino, Ioannis Patras, and Ebroul Izquierdo, "On the role of structure in part-based object detection," in *ICIP*, 2008, pp. 65–68.

[12] Jakob J. Verbeek and Bill Triggs, "Region classification with markov field aspect models," in *CVPR*, 2007.

[13] Manuela Vasconcelos, Nuno Vasconcelos, and Gustavo Carneiro, "Weakly supervised top-down image segmentation," in *CVPR (1)*, 2006, pp. 1001–1006.

[14] Till Quack, Bastian Leibe, and Luc J. Van Gool, "World-scale mining of objects and events from community photo collections," in *CIVR*, 2008, pp. 47–56.

[15] Thanos Athanasiadis, Phivos Mylonas, Yannis S. Avrithis, and Stefanos D. Kollias, "Semantic image segmentation and object labeling," *IEEE Trans. Circuits Syst. Video Techn.*, vol. 17, no. 3, pp. 298–312, 2007.

[16] Robert Fergus, Fei-Fei Li, Pietro Perona, and Andrew Zisserman, "Learning object categories from google's image search," in *ICCV*, 2005, pp. 1816–1823.

[17] Fei-Fei Li, Pietro Perona, and California Institute of Technology, "A bayesian hierarchical model for learning natural scene categories," in *CVPR (2)*, 2005, pp. 524–531.

[18] Bryan C. Russell, William T. Freeman, Alexei A. Efros, Josef Sivic, and Andrew Zisserman, "Using multiple segmentations to discover objects and their extent in image collections," in *CVPR (2)*, 2006, pp. 1605–1614.

[19] Antonio B. Torralba, Kevin P. Murphy, and William T. Freeman, "Contextual models for object detection using boosted random fields," in *NIPS*, 2004.

[20] Keiji Yanai, "Generic image classification using visual knowledge on the web," in *ACM Multimedia*, 2003, pp. 167–176.

[21] Kobus Barnard, Pinar Duygulu, David A. Forsyth, Nando de Freitas, David M. Blei, and Michael I. Jordan, "Matching words and pictures," *Journal of Machine Learning Research*, vol. 3, pp. 1107–1135, 2003.

[22] Alexander Jaffe, Mor Naaman, Tamir Tassa, and Marc Davis, "Generating summaries and visualization for large collections of geo-referenced photographs," in *Multimedia Information Retrieval*, 2006, pp. 89–98.

[23] James Philbin, Ondrej Chum, Michael Isard, Josef Sivic, and Andrew Zisserman, "Object retrieval with large vocabularies and fast spatial matching," in *CVPR*, 2007.

[24] Cameron Marlow, Mor Naaman, Danah Boyd, and Marc Davis, "Ht06, tagging paper, taxonomy, flickr, academic article, to read," in *Hypertext*, 2006, pp. 31–40.

[25] Eirini Giannakidou, Ioannis Kompatsiaris, and Athena Vakali, "Semsoc: Semantic, social and content-based clustering in multimedia collaborative tagging systems," in *ICSC*, 2008, pp. 128–135.

[26] Vasileios Mezaris, Ioannis Kompatsiaris, and Michael G. Strintzis, "Still image segmentation tools for object-based multimedia applications," *IJPRAI*, vol. 18, no. 4, pp. 701–725, 2004.

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

[28] B. Scholkopf, A. Smola, R. Williamson, and P. Bartlett, "New support vector algorithms," *Neural Networks*, vol. 22, pp. 1083–1121, 2000.