Chapter 9
Enhancing Computer Vision Using the Collective Intelligence of Social Media

Elisavet Chatzilari[1,3], Spiros Nikolopoulos[1,2],
Ioannis Patras[2], and Ioannis Kompatsiaris[1]

[1] Centre for Research & 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
ehatzi@iti.gr, nikolopo@iti.gr, ikom@iti.gr
[2] 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
i.patras@eecs.qmul.ac.uk
[3] Centre for Vision, Speech and Signal Processing University of Surrey Guildford,
GU2 7XH, UK
e.chatzilari@surrey.ac.uk

Abstract. Teaching the machine has been a great challenge for computer vision scientists since the very first steps of artificial intelligence. Throughout the decades there have been remarkable achievements that drastically enhanced the capabilities of the machines both from the perspective of infrastructure (i.e., computer networks, processing power, storage capabilities), as well as from the perspective of processing and understanding of the data. Nevertheless, computer vision scientists are still confronted with the problem of designing techniques and frameworks that will be able to facilitate effortless learning and allow analysis methods to easily scale in many different domains and disciplines. It is true that state of the art approaches cannot produce highly effective models, unless there is dedicated, and thus costly, human supervision in the process of learning that dictates the relation between the content and its meaning (i.e., annotation). Recently, we have been witnessing the rapid growth of Social Media that emerged as the result of users' willingness to communicate, socialize, collaborate and share content. The outcome of this massive activity was the generation of a tremendous volume of user contributed data that have been made available on the Web, usually along with an indication of their meaning (i.e., tags). This has motivated the research objective of investigating whether the Collective Intelligence that emerges from the users' contributions inside a Web 2.0 application, can be used to remove the need for dedicated human supervision during the process of learning. In this chapter we deal with a very demanding learning problem in computer vision that consists of detecting and localizing an object within the image content. We present a method that exploits the Collective Intelligence that is fostered inside an image Social Tagging System in order to facilitate the automatic generation of training data and therefore object detection models.

A. Vakali & L.C. Jain (Eds.): New Directions in Web Data Management 1, SCI 331, pp. 235–271.
springerlink.com © Springer-Verlag Berlin Heidelberg 2011

The experimental results shows that although there are still many issues to be addressed, computer vision technology can definitely benefit from Social Media.

1 Introduction

The recent advances of Web technologies have effectively turned ordinary people into active members of the Web, that generate, share, contribute and exchange various types of information. Web users act as co-developers and their actions and collaborations with one another have added a new social dimension on Web data. This social dimension of information was fostered by the next generation of the Web, namely Web 2.0, the applications of which have generated (and still generate) a remarkable volume of multimedia content. Based on this huge repository of content, various services have evolved [55], ranging from the field of eCommerce, to emergency response [56] and consumer collective applications such as realtravel.com [14]. The intelligence provided by single users organized in communities, takes a radical new shape in the context of Web 2.0, that of Collective Intelligence. Collective Intelligence emerges from the collaboration, communication and sharing among the users of social networks.

Although Collective Intelligence is at least as old as humans and appears in a wide variety of forms e.g., bacteria, animals, computer networks, it is now occurring in dramatically new forms. For example, Google[1] uses the knowledge millions of people have stored in the World Wide Web to provide useful answers to users' queries and Wikipedia[2] motivates thousands of volunteers around the world to create the world's largest encyclopedia. With new communication technologies and using the Internet as host, a large number of people all over the planet can now work together in ways that were never before possible in the history of humanity. But what exactly is Collective Intelligence and how can we benefit from it; The MIT Center for Collective Intelligence[3] frames the research question as *"How can people and computers be connected so that-collectively-they act more intelligently than any individuals, groups, or computers have ever done before?"*. It is now more important than ever for us to understand Collective Intelligence at a deep level so as to take advantage of these new possibilities.

In the field of multimedia data management, Collective Intelligence provides added value to the shared content and enables the accomplishment of tasks that are not possible otherwise. The acquisition of valuable knowledge is a big departure from traditional methods for information sharing, since managing Collective Intelligence poses new requirements. For example, semantic analysis has to fuse information coming both from the content itself, the social context and the emergent social dynamics. This fact has motivated increasing interest in discovering the different layers of Collective Intelligence, as well as in using these layers to empower new forms of

[1] http://www.google.com
[2] http://en.wikipedia.org
[3] http://cci.mit.edu

Web Data Management. Important progress towards this objective has been achieved in the context of the WeKnowIt[4] project were Collective Intelligence is considered to be the synthesis of 5 different layers namely, Personal Intelligence, Media Intelligence, Mass Intelligence, Social Intelligence and Organizational Intelligence.

In this chapter we investigate whether the Collective Intelligence derived from the user contributed content can be used to guide a learning process that will teach the machine how to recognize objects from visual content, the way a human does. We examine the problem both from the perspective of the teacher, which consists of knowledge that is build incrementally in an evolutionary and decentralized manner and therefore is characterized by questionable reliability, lack of structure, ambiguity and redundancy; as well as from the perspective of the learner that consists of models that apply learning algorithms on training data to capture the diversity of an object's form and appearance, and therefore demand for close supervision.

The rest of the chapter is structured as follows. Section 2 elaborates on the role of learning in computer vision and provides a description of Social Tagging Systems in the context of Web Multimedia Data. Section 3 emphasizes on the key aspect of multimedia analysis and provides an overview of the basic mechanisms that are used for learning. Section 4 presents an approach for training object detection models using data from collaborative tagging environments, that exploits the Collective Intelligence derived from the massive users' contribution. Concluding remarks are drawn in Section 6.

2 Learning and Web 2.0 Multimedia

2.1 Learning in Computer Vision

Learning has always been of primary importance for computer vision scientists. If we wish to construct a visual system that is able to scale on an arbitrary large number of concepts, effortless learning is crucial. Humans learn to recognize materials, objects and scenes from very few examples and without much effort. A 3-year old child is capable of building models for a substantial number of concepts and recognizing them using these models. By age of six humans recognize more than 10^4 categories of objects [7] and keep learning more throughout their life. Can a computer program learn how to recognize semantic concepts from images? This is the general question addressed by the computer vision scientists. But what is the process of learning; what is the mechanism that allows humans to initially require many examples to learn, as performed by little babies, and after they have learned how to learn, they can learn from just a few examples; and most importantly what is the role of the teacher in this process and what is the minimum amount of supervision that is absolutely necessary for facilitating efficient learning?

In [38] the authors make the hypothesis that, once a few categories have been learned with significant effort, some information may be abstracted from the process to make learning further categories more efficient. Similarly in [41] when images

[4] http://www.weknowit.eu/

of new concepts are added to the visual analysis model, the computer only needs to learn from the new images. What has been learned about previous concepts is stored in the form of profiling models, and the computer needs no re-training. On the other hand in [67] the authors claim that with the availability of overwhelming amounts of data, many problems can be solved without the need for sophisticated algorithms. The authors mention the example of 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. In their paper the authors present a visual analog to this tool using a large dataset of 79 million images and a non-parametric approach for image annotation that is based on nearest neighbor matching.

However, the need for effortless learning coupled with the fact that the images archived on the Internet are growing at a phenomenal rate, has motivated other researchers to turn their interest in weakly (i.e. image level) annotated instead of strongly (i.e. region or pixel level) annotated images. Fig. 1 shows an example image with both strong and weak annotations. Photo sharing through the Internet has become a common practice and according to the reports released in 2007, flickr.com has 40 million monthly visitors and hosts two billion photos, with new photos in the order of millions being added on a daily basis. In this context, the authors of [11] use multiple instance learning to learn models from images labeled as containing the semantic concept of interest, but without indication of which image regions are observations of that concept. Similarly in [18] object recognition is viewed as machine translation by learning how to map visual objects (blobs) to concept labels. In [15] models are learned from ambiguously labeled examples, where each example is supplied with multiple potential labels, only one of which is correct. Approaches that learn an object category from just its name include [21], where the authors obtain images from the web using Google or Yahoo search engines and takes the returned results to be pseudo-positively labeled training images.

The key trade-off between the annotation-based models (which use labels provided by human annotators) and search-based models (which use models automatically obtained from the Web), is from the one side the amount of human effort that is required in annotation-based models and on the other side the expected decrease

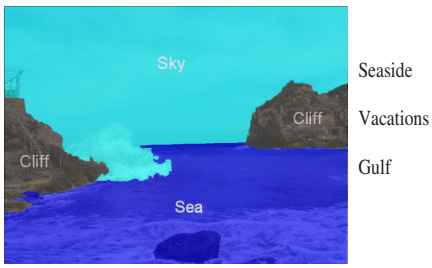


Fig. 1. An example image annotated both strongly (i.e., each of the identified image regions is assigned with a label) and weakly (i.e., a set of labels is provided to describe the image content)

in classification performance that will result from search-based methods [34]. Recently, and driven by the widespread appeal of social sites, we have been witnessing an increasing research interest in using social sites and in particular Social Tagging Systems (STS), instead of the search engines, to obtain the necessary labels.

2.2 Social Tagging Systems and Web 2.0 Multimedia

An STS is a web-based application, where users, either as individuals or more commonly as members of a community (i.e., social networks), assign labels (i.e., arbitrary textual descriptions) to digital resources. Their motivation for tagging is information organization and sharing. Social tagging systems tend to form rich knowledge repositories that enable the extraction of patterns reflecting the way content sementics is perceived by the web users. In [27] the authors show that the tag proportions each resource receives crystallizes after about 100 annotations attributing this behavior to the users' common background and their tendency for imitation on other users' tagging habits. The availability of such content in the Web is high and the exploitation of the Collective Intelligence that is fostered by this type of content still remains a challenge.

In order to extract the knowledge that is stored and often "hidden" in social data, researches have employed various approaches: a) Clustering techniques that are based on tagging information and tag co-occurrence to derive semantically-related groups of tags and resources [4], [28], [30], b) Ontology driven tagging organization and mining, by combining Web 2.0 and Semantic Web, [29], [60], [54], c) content-based analysis of tagging-related sources that explore both tags and visual features (in a supplementary manner) for browsing and retrieving semantically related images [2], [57], [25]. Despite the active research efforts in this area, the full potential of Web 2.0 data has not been exploited yet. Few approaches exploit the fact that the tag and visual information space are highly correlated and the Collective Intelligence that will emerge from the massive participation of users in contributing and tagging multimedia content can be used to facilitate the learning process of computer vision systems.

However, while the Collective Intelligence derived from social data seems a very promising source of information, it has some serious limitations that mainly derive from the unconstrained nature of Web 2.0 applications. Users are prone to make mistakes and they often suggest invalid metadata (tag spamming). The lack of (hierarchical) structure of information results in tag ambiguity (a tag may have many senses), tag synonymy (two different tags may have the same meaning) and granularity variation (users do not use the same description level, when they refer to a concept).

There is a growing number of research efforts that attempt to overcome the aforementioned limitations and exploit the dynamics of social tagging systems to facilitate different types of multimedia applications. In [2], the authors claim that the intrinsic shortcomings of collaborative tagging can be tackled by employing content-based image retrieval technique. The user is facilitated in image database browsing and retrieval by exploiting both the tag and visual features in a supplementary way. In [25] a number of clustering techniques were employed in order to couple tagging

information with content-based features. The clustering was tag-oriented and occurred in two steps. In the first step the resources were assigned to clusters, depending on the similarity of their accompanying tags. In the second step, visual features were employed, in an effort to increase the purity of already created clusters. The second step of the process could be regarded as a "misleading tags tracking phase". Another work that combines user data with feature-based approaches is presented in [24], that is used to rank the results of a video retrieval system. The authors use this knowledge, along with a multimedia ontology to build a learning personalized environment.

There are also works that address the problem of identifying photos from social tagging systems that depict a certain object, location or event [35], [57]. In [35] 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. In [57] 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 geo-tagged 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.

In all cases the authors are trying to benefit from the Collective Intelligence that emerges from the content contributed to STSs and improve the efficiency of certain tasks. However the correlations between the tag and visual information space that are established when the users suggest tags for the uploaded visual content, are mostly treated as complementary sources of information that both contribute to the semantic description of the resources. In contrast to the above this chapter investigates whether the aforementioned correlations can be used to facilitate the learning process of multimedia analysis models. For this reason in the following Section we provide a short introduction to some of the techniques used for multimedia analysis.

3 Multimedia Analysis and Management

3.1 The Need for Semantics

The efficient management of multimedia data poses many technological challenges in terms of indexing, querying and retrieving, that require a deep understanding of the information at a semantic level. Driven by this need and given that machines' perception is limited to numbers and strings, there have been many research efforts that try to map semantic concepts or events to low level features, an issue addressed as bridging the "semantic gap".

The very first attempts for image retrieval were based on keyword search [50] applied either on the associated annotations (assuming that annotations existed) or on the images' file names. However, these approaches apart from requiring textual annotations of the multimedia data, they are barely as descriptive as the multimedia content itself. To overcome these limitations, the use of the image visual characteristics has been proposed. In this case, the visual content is utilized by extracting

a set of visual features from each image or image region. By comparing the visual features an algorithm can decide whether the two images/regions represent the same semantic concept. Then, image retrieval is performed by comparing the visual features of an example image/region that is associated with a semantic concept by the user, to the visual features of all images in a given collection [20] (known as Query By Image Content systems).

Subsequently, more sophisticated methods were proposed that aimed at simulating the functionality of human visual system by allowing the machines to mimic the procedure followed by a human when identifying semantics in visual content. In this direction pattern classification has been brought to the core of most image analysis techniques in order to render a kind of meaning on visual patterns. A typical pattern classification problem can be consider to include a series of sub-problems the most important of which are: a) determining the optimal feature space, b) removing the noisy data that can be misleading, c) avoid overfitting on training data, d) use the most appropriate distribution for the model, e) make good use of any prior knowledge that may help you in making the correct choices, f) perform meaningful segmentation when the related task requires to do so, h) exploit the analysis context, etc. All the above are crucial for initiating a learning process that aims at using the available training samples to estimate the parameters of a model representing a semantic concept. In the following section we discuss and provide related references for some of the aforementioned sub-problems, giving special emphasis on the mechanisms of learning.

3.2 Visual Features Extraction and Regions Identification

Many problems derive from the fact that it is very difficult to describe visual content effectively in a form that can be handled by machines. In general, feature extraction is a domain dependant problem and it is unlikely that a good feature extractor for a specific domain will work as good for another domain. The extraction of features for efficient image representation has attracted a lot of interest in the scientific community of image analysis. Motivated by the principles of human perception, most researchers have tried to describe images/regions using color, shape and texture characteristics. Some of the most widely adopted techniques for representing images/regions include the descriptors proposed by the MPEG-7 standard [1] that capture different aspects of color, texture and shape. Other approaches rely on the corners and edges that can be found inside an image in order to describe the image using a set of interest points. The Scale-Invariant Feature Transform (SIFT) proposed in [44] and its modifications (i.e., color SIFT, opponent SIFT, etc [59]) are considered some of the most representative algorithms of this category. Particularly important is also considered the vector quantization approach initially proposed in [64] where in analogy to text, images/regions are represented as bags-of-visual words that have been learned through an extensive training process using representative data.

Additionally, many problems derive from the fact that images tend to include more than one objects in their content, which decreases the descriptiveness of the feature space and raises the need for segmentation. The segmentation of images into

regions and the use of a separate set of features for each region was introduced to address the aforementioned issue. Segmentation techniques seek to detect groups of pixels sharing similar visual characteristics and identify in this way meaningful objects (similar to the ones identified by human visual system). In the field of segmentation one of the most commonly used methods is Normalized Cuts [62] which is a graph partitioning algorithm using a global criterion, the normalized cut, for segmenting the graph. Other approaches include [52] that segments color images by applying a variance of K-means on intensity, position and texture features, as well as [12] that is based on the Expectation-Maximization (EM) algorithm. Both segmentation and feature extraction are two very important techniques for identifying patterns in visual content. However, in order to bridge the semantic gap these patterns will have to be classified into meaningful concepts. This is where the role of learning takes place since it is used to estimate the parameters of a model that will be sub-sequently used to classify new, unseen images or regions.

3.3 Learning Mechanisms

Humans can classify images through models that are built using examples for every single semantic concept. Based on this assumption, researchers have been trying to simulate human visual system by using machine learning algorithms to classify the visual content. A set of training samples plays the role of the examples that a person uses to learn a concept. Based on the prior knowledge that we have on the training samples during the learning process, we can distinguish between the following basic categories; unsupervised, strongly supervised, semi-supervised and weakly supervised learning.

3.3.1 Un-supervised Learning

Unsupervised learning is a class of problems in which one seeks to determine how the data are organized. It is distinguished from supervised learning in that the learner is given only unlabelled examples. One of the most known forms of unsupervised learning is clustering. The clustering output can be hard (a partition of the data into groups) or fuzzy (where each data point has a variable degree of membership in each output cluster) [6]. Clustering algorithms can be divided in two major categories, hierarchical [32] and partitional [47]. Hierarchical methods produce a nested series of partitions whereas partitional methods produce only one partition. Many clustering algorithms require the specification of the number of clusters to produce in the input data set, prior to the execution of the algorithm. Barring knowledge of the proper value beforehand, the appropriate value must be automatically determined, a problem for which a number of techniques have been developed [13], [23]. Another important step in any clustering scheme is the selection of a distance measure, which determines how the similarity of two elements is calculated. The distance measure influences the shape of the clusters, as some elements may be close to one another according to one distance and farther away according to another.

3.3.2 Strongly-Supervised Learning

In strongly-supervised learning there is prior knowledge about the labels of the training samples and there is one-to-one relation between a sample and its label (e.g., each region of the image depicted in Fig. 2 is associated with a label). The aim of strongly-supervised learning is to generate a global model that maps input objects to the desired outputs and generalize from the presented data to unseen situations in a "reasonable" way. Some of the most widely used types of classifiers that typically rely on strongly annotated samples are the Neural Network (Multilayer perceptron) [19], Support Vector Machines [51], naive Bayes [17], decision tree [8] and radial basis function classifiers [46]. A known issue in supervised learning is overfitting (i.e. the model describes random error or noise instead of the underlying relationship) which is more probable to occur when the training samples are rare and the dimensionality of the feature space is high. In order to avoid overfitting, it is necessary to use additional techniques (e.g. cross-validation, regularization, early stopping, Bayesian priors on parameters or model comparison), that can indicate when further training is not resulting in better generalization.

3.3.3 Semi-supervised Learning

Semi-supervised learning algorithms try to exploit unlabeled data, which are usually of low cost and can be obtained in high quantities, in conjunction with some supervision information. In this case, only a small portion of the data is labeled and the algorithm aims at propagating the labels to the unlabeled data. The earliest idea about using unlabeled data when learning a classification model is self-learning. In self-learning, the classification model is initially trained using only the labeled data and at each step a part of the unlabeled data is labeled according to the output of the current model. Then, a new classification model is trained using both the labeled as well as the data that were labeled as positive from the previous step.

Another category of semi-supervised learning algorithms is based on the cluster assumption, according to which the points that are in the same cluster belong to

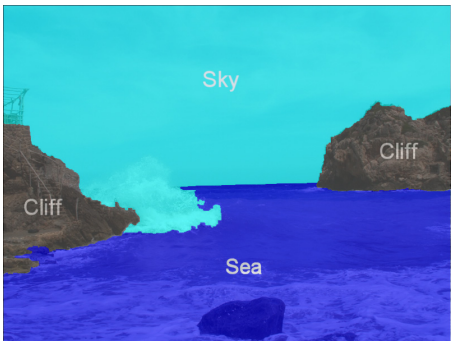


Fig. 2. An image depicting the object sea that is manually annotated at region level

the same class. So there should be regions with high density of points (which formulate the clusters) and low-density regions where the decision boundary lies in. Most of the recent semi-supervised classification approaches aim at creating new specialized learning algorithms that are able to combine labeled and unlabeled data. More specifically, in order to have the ability to choose a learning algorithm with the required attributes most state-of-the-art methods combine semi-supervised learning with boosting techniques. Boosting is part of the ensemble learning family algorithms and aims at building an ensemble by training each new model instance to emphasize the training instances that previous models mis-classified [5]. The most popular algorithm that utilizes the boosting method is Adaboost (short for Adaptive Boosting) presented in [22]. Algorithms that adopted the SemiBoost approach, which employs the boosting method in order to improve any existing supervised learning algorithm with unlabeled data, include [48], [37], [9].

3.3.4 Weakly-Supervised Learning

By weakly-supervised we refer to the process of learning using weakly labeled data (i.e., samples labeled as containing the semantic concept of interest, but without indication of which segments/parts of the sample are observations of that concept, (as shown in Fig. 3). In this case, the basic idea is to introduce a set of latent variables that encode hidden states of the world, where each state induces a joint distribution on the space of semantic labels and image visual features. New images are annotated by maximizing the joint density of semantic labels, given the visual features of the new image [11]. The most indicative weakly-supervised learning algorithms are the ones that are based on aspect models like probabilistic Latent Semantic Analysis (pLSA) [63], [21] and Latent Dirichlet Allocation (LDA) [39], [58]. These models are typically applied on weakly annotated datasets to estimate the joint distribution of semantic labels and visual features.

3.4 Annotation Cost for Learning

Object detection schemes always employ some form of supervision as it is practically impossible to detect and recognize an object without using any semantic information during training. However, semantic labels may be provided at different levels of granularity (global or region level) and preciseness (one-to-one or many-to-many relation between objects and labels), imposing different requirements on the effort required to generate them. Indeed, there is a clear distinction between the strong and accurate annotations that are usually generated manually and constitute a laborious and time consuming task, and the weak and noisy annotations that are usually generated by web users for their personal interest and can be obtained in large quantities from the Web or collaborative tagging environments like flickr[5]. Due to the fact that the annotation cost is a critical factor when designing an object detection scheme with the intention to scale in many different objects and domains, in the

[5] www.flickr.com

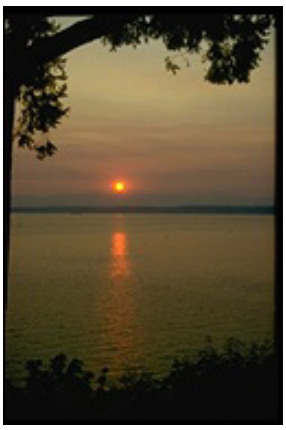


Fig. 3. An image depicting the object sea that is manually annotated at global level

following we distinguish the object detection methods based on the characteristics of the dataset that they employ and the effort required for its annotation. Our goal is to highlight the tradeoff between the annotation cost for preparing the necessary training samples and the quality of the resulting models.

In the first category we classify the methods that use manually annotated images at region level as the one depicted in Fig. 2. These methods rely on strongly supervised learning and are usually developed to recognize certain types of objects with very high accuracy. In [70] and [66] manual annotations of faces are used in order to train the classifiers. In [43] a method for the recognition of buildings is proposed and in [36] an implicit shape model for the detection of cars is presented. In [71] manual annotations at region level are used to train a probabilistic model integrating both visual features and spatial context. Annotating images at region level is probably the task with the highest annotation cost.

Image annotations at global level, even manual ones, are easier to obtain than region level annotations. This fact has motivated many researchers in developing algorithms that rely on weakly-supervised and semi-supervised learning, and are able to exploit global annotations for performing object detection. The Corel database is probably the most widely used set of images annotated manually at global level (as shown in Fig. 3) and has been used in numerous works. Jia Li and James Z. Wang [40] used the Corel dataset to train models for each concept separately, while in [69] it was used to evaluate the performance of an algorithm that considers the recognition of visual concepts to be part of the segmentation process. The Corel dataset has been also used by Duygulu et. al. that presented a methodology for mapping words to image regions using an algorithm based on EM [18], as well as in [10] where a label propagation algorithm that incorporates time, location and visual similarity for event and scene detection has been proposed. The widespread use of Corel and other datasets of similar type can be mainly attributed to the fact that the global annotations associated with the images was noise free and accurate. This allowed the

researchers to derive some probabilistic relations between objects and labels and use these relations to perform object detection on new images. However, labels accuracy comes with the cost of manual annotation which is something that limits the scaling potentials of the schemes relying on such labels. This was the reason that researchers turned their interest on the Web and started to investigate whether it could be used to obtain globally annotated images.

Using the Web as a source many approaches have been proposed that obtain globally annotated images through search engines, using the name of the object as argument (see Fig. 4 for some example images obtained using the query word "sea"). Keiji Yanai uses visual content from the Web as training images for a generic classification system [73] and in [21] the authors learn object categories from Google's image search. However, since search engines in their current form rely primarily on the image filename or the surrounding text to decide whether to return an image or not, the quality of the obtained annotations is very low. Thus, although these type of global annotations can be obtained at practically no cost, the high level of noise renders particularly difficult the extraction of reliable probabilistic relations between objects and labels.

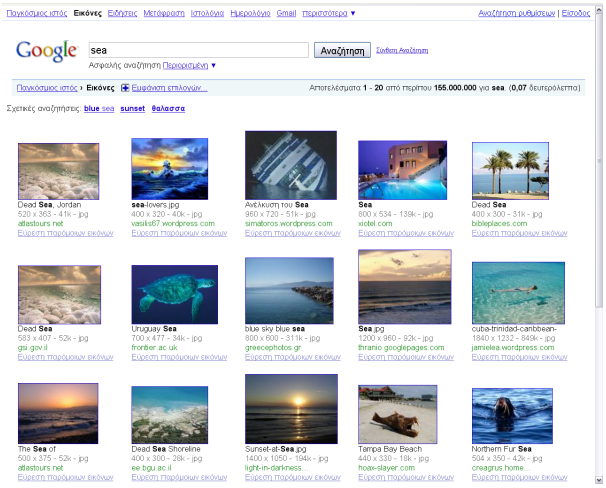


Fig. 4. Images depicting the object sea, obtained automatically from the Google Image Search engine using "sea" as the query word

For this reason, the most recent research efforts are focusing on the content that is being massively contributed by Web users in the context of Web 2.0 applications. In [57] object and event detection is performed by clustering images downloaded from flickr based on textual, visual and spatial information and verified through Wikipedia[6] content. Similarly a framework that probabilistically models geographical information for event and activity detection using geo-tagged images from flickr

[6] www.wikipedia.com

is presented in [33]. Although the tag annotations that accompany the images contributed by social users are less noisy from the ones obtained via the search engines, they are still considered to be rather noisy for directly extracting the necessary probabilistic relations between objects and labels, see Fig. 5 for an example image obtained from Flickr along with the associated tags.

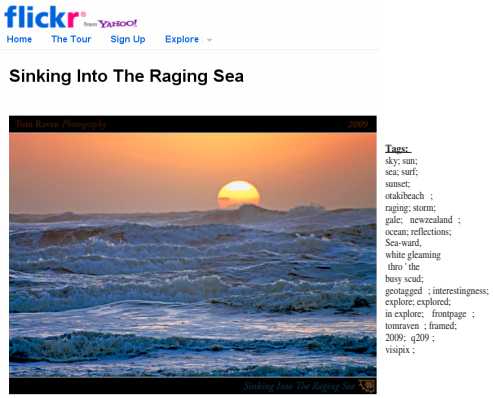


Fig. 5. Image depicting the object sea, obtained automatically from Flickr along with the associated tags

In Table 1 we summarize the pros and cons for each of the aforementioned types of annotation. As a general conclusion we can say that manual image annotation (either at region or global level) is a time consuming task and as such it is particularly difficult to be performed on the desired volumes of content that are needed for building robust and scalable classifiers. On the other hand, the STSs and the Web provides cost free annotations that are very noisy to be used directly for extracting the necessary probabilistic relations between objects and labels. The Collective Intelligence that emerges from the tagged images aggregated in STSs would have to be exploited towards removing the existing obstacles. In this direction we present a method that transforms global image tags into region level annotations, in a form suitable to be used by a strongly-supervised learning algorithm for object detection.

4 Leveraging Social Media for Training Object Detectors

As already described, machine learning algorithms fail in two main categories in terms of the annotation granularity, the algorithms that are designed to learn from strongly annotated samples (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 (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 given a training set made of pairs (f_i, c_i). On the

Table 1. Pros & Cons for the different types of annotation

Annotation Type	Automated Annotation	Scaling Capability	Training Efficiency	Learning Mechanism	Related Techniques
Region-level (manual) (Fig. 2)	Poor	Poor	Excellent	strongly-supervised	Viola & Jones [70], Sung & Poggio [66], Li et al. [43], Leibe et al. [36], Wand et al. [71]
Global-level (manual) (Fig. 3)	Fair	Fair	Good	weakly-supervised	Li & Wang [40], Vascooncelos et al. [69], Duygulu et al. [18], Cao et al. [10]
Global-level (automatically via Search Engines) (Fig. 4)	Excellent	Excellent	Poor	weakly-supervised	Yanai [73], Fergus et al. [21]
Global-level (automatically via Social Networks) (Fig. 5)	Excellent	Excellent	Fair	weakly-supervised	Quack et al. [57], Dhiraj & Lue [33]

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 $\{(f_1,\ldots,f_n),(c_1,\ldots,c_m)\}$.

While model parameters can be estimated more efficiently from strongly annotated samples, such samples are very expensive to obtain. On the contrary, weakly annotated samples can be found in large quantities especially from sources related to 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 object detection models from training samples that are found in collaborative tagging environments.

4.1 Problem Formulation

The problem can be formulated as follows. Drawing from a large pool of weakly annotated images, our goal is to benefit from the knowledge that can be extracted from social tagging systems, in order to automatically transform some of the weakly annotated images into strongly annotated ones. In order to do this, we consider that if the set of weakly annotated images is properly selected from the repository of a collaborative tagging environment, the most populated tag-"term" and the most populated visual-"term" will be two different representations/expressions (i.e., textual and visual) of the same object. We define tag-"terms" to be sets of tags that are provided by social users to describe an image and are grouped based on their semantic

affinity (e.g., synonyms, derivatives, etc). Respectively, we define visual-"terms" to be sets of image regions that are identified by an automatic segmentation algorithm and are grouped based on visual similarity. 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 identify, while the most populated visual-"term" (i.e., the cluster of image regions containing the most instances) is used to provide the set of strongly annotated samples for training the classifier. It is clear that the process of leveraging weakly annotated images to become the strongly annotated training samples of a supervised learning scheme, is primarily achieved through the semantic clustering of image regions to objects (i.e., each cluster consists of regions that depict only one object). Using the notation of Table 2 semantic clustering can be formulated as follows. Given a large set of images $I_q \in S^c$ with annotation information of the type $\{(f_d(r_1^{I_q}),\ldots, f_d(r_m^{I_q})), (c_1,\ldots, c_t)\}$, semantic clustering would produce pairs (w_i, c_i) where each w_i is a set of regions extracted from all images in S^c that depict only c_i. Semantic clustering can only be made feasible in the ideal case where the image analysis techniques employed by our framework works perfect. However, this is highly unlikely due to the following reasons. In case of over or under segmentation we will have more or fewer regions from the actual objects in image, making perfect semantic

Table 2. Legend of Introduced Notations

Symbol	Definition
S	The complete social dataset
N	The number of images in S
S^c	An image group, subset of S that emphasizes on object c
n	The number of images in S^c
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 features extracted from a region $r_i^{I_q}$
$C = \{c_i, i = 1,\ldots,t\}$	Set of objects that appear in the images of group S^c
$W = \{w_i, i = 1,\ldots,o\}$	Set of clusters created by the region-based clustering algorithm
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

clustering impossible. Similarly, the inadequacy of visual descriptors to perfectly discriminate between different semantic objects is likely to lead the clustering algorithm in creating a different number of clusters than the number of actual semantic objects, or even mix regions depicting different objects into the same cluster. Thus, instead of requiring that each w_i is mapped with a c_i, we only search for a single pair (w_k, c_z) where the majority of regions in cluster w_k depicts c_z. Given that both w_i (i.e., visual-"term') and c_i (i.e., tag-"term") are sets (of images regions and user contributed tags, respectively), we can apply the $Pop(\cdot)$ function on them, that calculates the population of a set (i.e., number of members). Eventually, the problem addressed in our approach is what should be the characteristics of S^c so as the pair (w_k, c_z) determined using $k = \arg\max_i(Pop(w_i))$ and $z = \arg\max_i(Pop(c_i))$ satisfies our objective i.e., that the majority of regions included in w_k depicts c_z. Our approach in using user contributed content to create S^c is motivated by the fact that due to the common background that most users share, the majority of them tend to contribute similar tags when faced with similar type of visual content [49]. This is the point where our approach benefits from the Collective Intelligence that emerges from an STS, in the sense that it would be over-ambitious to rely on such an assumption if tags were to be contributed by just one or a few users. However, since the tags in an STS originate from a significantly large amount of users, it is statically safe to conclude that the majority of tag assignments will conform to the aforementioned rule. Then, given this assumption 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.

4.2 Framework Description

The framework we propose for leveraging the weakly annotated data in order to train object detection models, is depicted in Fig. 6. The analysis components that can be identified in our framework are, tag-based image selection, image segmentation, extraction of visual features from image regions, region-based clustering using their visual features and learning of object detection models using strongly annotated samples.

More specifically, given an object c that we wish to train a detector for, our method starts from a large collection of user tagged images and performs the following actions. Images are selected based on their tag information in order to formulate image group(s) that correspond to thematic entities. Given the tendency of social tagging systems to formulate knowledge patterns that reflect the way content is perceived by the web users [49], [27], tag-based processing is expected to identify these patterns and create image group(s) each one emphasizing on a certain object. By emphasizing we refer to the case where the majority of the images within a group depict different instances of a certain object and that the linguistic description of that object can be obtained from the most frequently appearing tag (see Section 4.3.1 for more details). Subsequently, region-based clustering is performed on all images belonging to the image group that emphasizes on object c, that have been pre-segmented by an automatic segmentation algorithm. During

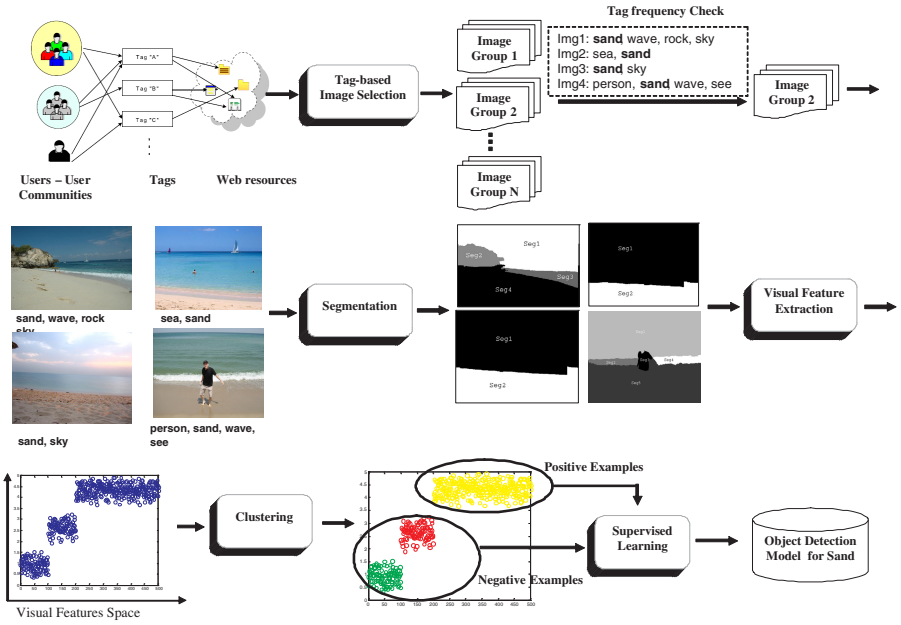


Fig. 6. Actions performed by our framework in order to train a model for detecting the object sand

region-based clustering the image regions are represented by their visual features and each of the generated clusters contains visually similar regions. Since the majority of the images within the selected group depicts instances of the desired object c, 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 strongly supervised manner) a model detecting the object c. Although noisy tags and inaccurate segmentation are likely to prevent the most populated cluster from gathering all regions depicting object c, the fact that the collection of user tagged images can be arbitrary large (due to its "social" origin) can compensate for the loss in accuracy.

We can view the process of using image tag information to create an image group S^c that emphasizes on object c, as the process of selecting images from a large pool of weakly annotated images using as argument a query tag t_q. t_q is the linguistic description of the object c. The selection criteria can be keyword-based search (in the trivial case), pre-annotated groups, or more sophisticated approaches (see Section 4.3.1). Although misleading and ambiguous tags always hinders this process, the expectation is that as the number of selected images grows large and the tag-proportions for each image crystallizes [27], 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.

Let us assume that using tag-based selection we construct an image group $S^c \subset S$ emphasizing on object c. What we are interested in is the frequency distribution of objects $c_i \in C$ appearing in S^c based on their frequency rank. We can view the process of constructing S^c as the act of populating an image group 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, 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 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. (1) shows the probability mass function of a random variable following the binomial distribution.

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

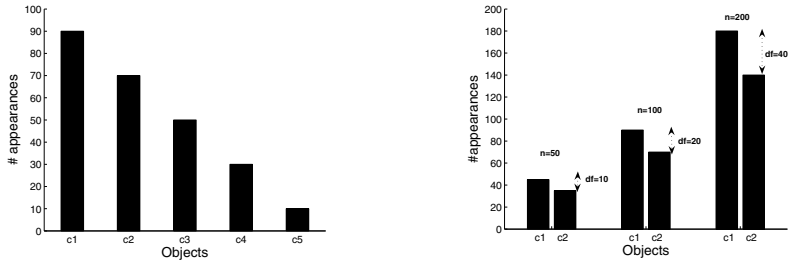
Given the above, we can use the expected value $E(K)$ of a random variable following the binomial distribution to estimate the expected number of images in S^c that depict object $c_i \in C$, if they are drawn from the initial dataset S with probability p_{c_i}. This is actually the value of k maximizing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

$$E_{c_i}(K) = np_{c_i} \tag{2}$$

If we consider α 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 is:

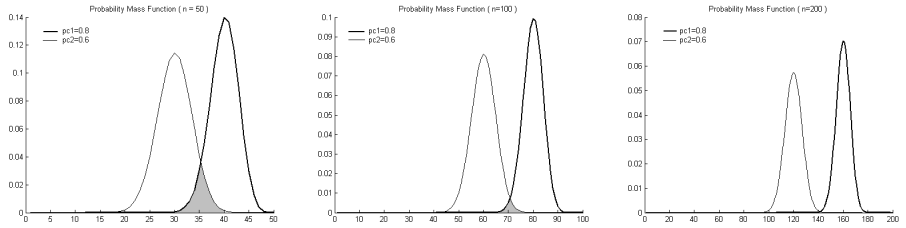
$$TC_i = \alpha np_{c_i} \tag{3}$$

Moreover, 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 c_2 is drawn, and so forth for the remaining $c_i \in C$. This assumption is experimentally verified in Section 4.4.2 where the frequency distribution of objects for different image groups are measured in a manually annotated dataset. Finally, using eq. (3) we can estimate the expected number of appearances (#*appearances*) of an object in S^c, $\forall c_i \in C$. Fig. 7(a) shows the #*appearances* $\forall c_i \in C$ against their frequency rank, given some examples values of p_{c_i} with $p_{c_1} > p_{c_2} > \ldots$. It is clear from eq. (3) 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. 7(b) for some examples. Additionally, apart from increasing the expected absolute difference on the #*appearances* between the two most frequently appearing objects, the high values of n also minimize the probability of the case where c_2 although drawn with probability smaller that c_1 appears more times in the generated image group. In Fig. 8 we draw the probability mass function of two random variables that correspond to objects c_1, c_2 of Fig. 7(b) (i.e., $p_{c_1} = 0.8$, $p_{c_1} = 0.6$) for three different values of n (i.e., $n = 50$, $n = 100$ and $n = 200$). The probability of experiencing the case where c_2, although drawn with smaller probability, appears more times than c_1



(a) (b)

Fig. 7. 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.



(a) (b) (b)

Fig. 8. a) Probability mass function for $n = 50$ trials and $p_{c_1} = 0.8$, $p_{c_2} = 0.6$. b) Probability mass function for $n = 100$ trials and $p_{c_1} = 0.8$, $p_{c_2} = 0.6$. c) Probability mass function for $n = 200$ trials and $p_{c_1} = 0.8$, $p_{c_2} = 0.6$.

in S^c, is proportional to the surface where the two curves overlap. It is clear from Fig. 8 that as n increases, the variance of the two random variables decrease, forcing the surface of the overlapping region to also decrease (e.g., for $n = 200$ the surface of the overlapping region is almost zero). Based on these observations, we reach the theoretical expectation that there is higher probability in w_k being a set of regions the majority of which depict c_z as n increases.

4.3 Implementing the Framework

In this section we provide details for the analysis components that are used by the proposed framework. Due to the fact that a necessary pre-requisite for our framework to work efficiently is operating on a large number of images, a discussion about the complexity of each analysis component is also included.

4.3.1 Tag-Based Image Selection

In this section we specify the approaches that are used to select images from a large dataset of arbitrary content, based on their tag information. We employ one of the following three approaches based on the associated annotations:

Keyword-based search: This approach is used for selecting images from strongly annotated datasets. These datasets are usually hand-labeled and the tags provided by the annotators can be considered to be mostly accurate and free of ambiguity. Thus, in order to create S^c we need only to select the images that are tagged with the linguistic expression of the object c.

Flickr groups: are virtual places hosted in collaborative tagging environments that allow social users to share content on a certain topic. Although managing flickr groups still involves some type of human annotation (i.e., a human assigns an image to a specific group) it can be considered weaker than the previous case since this type of annotation does not provide a full description of the objects depicted in the image. In this case, S^c is created by taking the images contained in a flickr group titled with the name of the object c. From here on we will refer to those images as roughly-annotated images.

SEMSOC: stands for SEmantic, SOcial and Content-based clustering and is applied in our framework on weakly annotated images (i.e., images that have been tagged by humans in the context of a collaborative tagging environment, but no rigid annotations have been provided) in order to create semantically consistent groups of images. SEMSOC was introduced by Giannakidou et. al. in [25], [26] and is an un-supervised model for the efficient and scalable mining of multimedia social-related data that jointly considers social and semantic features. 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 groups $S^{c_i} \subset S, \quad i = 1,\ldots,m$, where m is the number of created groups. This number is determined empirically, as described in [25]. Then in order to obtain the image group S^c that contains the images depicting the desired object c, we select the SEMSOC-generated group S^{c_i} where its most frequent tag relates with c. Fig. 9 shows four examples of image clusters generated by SEMSOC, along with the corresponding most frequent tag.

4.3.2 Segmentation

Segmentation is applied on all images in S^c 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 [53]. The output of this algorithm is a set of segments $R_{I_q} = \{r_i^{I_q}, i = 1,\ldots,m\}$, which roughly correspond to meaningful objects, $c_i \in C$. The time efficiency of the segmentation process depends mainly on the size of the image. The segmentation of low-resolution images is performed

(a) Vegetation (b) Sky

(c) Sea (d) Person

Fig. 9. Examples of image groups generated using SEMSOC (in caption the corresponding most frequent tag). It is clear that the majority of images in each group include instances of the object that is linguistically described by the most frequent tag. The image is best view in color and with magnification.

considerably fast but the time efficiency of the algorithm degrades quickly as the image size increases. Therefore in order to cope with high-resolution images the authors of [53] make the reasonable assumption that the regions falling below the 0.75% of the total image area are insignificant. Based on this assumption the segmentation algorithm is applied on a reduced version of the image (i.e., by downscaling its original version). This improves the time efficiency of the algorithm but at the expense of the quality of the segmentation result. To alleviate this, the pixels belonging to blocks on edges between regions are reclassified using the Bayes classifier. Applying the segmentation algorithm on reduced images with reclassification using the Bayes classifier, delivers the same segmentation quality with segmentation time significantly reduced.

4.3.3 Visual Descriptors

In order to visually describe the segmented regions we have employed the following: a) the Harris-Laplace detector and a dense sampling approach for determining the interest points, b) the SIFT descriptor as proposed by Lowe [45] in order to describe each interest point using a 128-dimensional feature vector and c) the bag-of-words approach initially proposed in [64] in order to obtain a fixed-length feature vector for each region. The feature extraction process is similar to the one described in [59] with the important difference that in our case descriptors are extracted to represent each of the pre-segmented 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 [74]. 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 approach) the 128-dimensional SIFT descriptor is computed using the version described by Lowe [45]. SIFT descriptors have been been found to be particularly robust against variations in scale, rotation, changes in brightness and contrast, etc. A Visual Word Vocabulary (Codebook) was 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 contained in an image region to be vector quantized against the set of Visual Words and create a histogram. Thus, $\forall r_i^{I_q} \in R_{I_q}$ and $\forall I_q \in S^c$ a 300-dimensional feature vector $f(r_i^{I_q})$ is extracted, that contains information about the presence or absence of the Visual Words included in the Codebook. All feature vectors were normalized so as the sum of all dimensions to be equal with 1.

4.3.4 Clustering

For performing feature-based region clustering we applied the affinity propagation clustering algorithm on all extracted feature vectors $f(r_i^{I_q})$, $\forall r_i^{I_q} \in R_{Iq}$ and $\forall I_q \in S^c$. This is an algorithm that takes as input the measures of similarity between pairs of

data points and exchanges messages between data points, until a high-quality set of centers and corresponding clusters is found. Affinity propagation was proposed by Frey and Dueck [23] and was selected for our work due to the following reasons: a) The requirements of our framework imply that in order to learn a robust object detection model, clustering will need to be performed on a considerably large number of regions, making computational efficiency an important issue. The common approach followed by most clustering algorithms is to determine a set of centers such that the sum of squared errors between data points and their nearest centers is minimized. This is done by starting with an initial set of randomly selected centers and iteratively refining this set so as to decrease the sum of squared errors. However, such approaches are sensitive to the initial selection of centers and work well only when the number of clusters is small and the random initialization is close to a good solution. This is the reason why these algorithms need to re-run many times with different initializations in order to find a good solution. In contrast to this, 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 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 images of an image group 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. This number is called preference and has the meaning that data points with larger preferences are more likely to be chosen as centers. In this way the number of identified centers (number of clusters) is influenced by the values of the input preferences but 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). Given that it is better for our framework to handle noisy rather than inadequate (in terms of indicative examples) training sets, we opt for the minimum value in our experiments.

4.3.5 Learning Model Parameters

Support Vector Machines (SVMs) [61] 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 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 use of cross validation for selecting the kernel parameters. Considering that the size of available samples can grow arbitrary big, training a model could become a

particularly costly procedure. The SVM^{light} implementation of SVMs was used to address the problem of large scale tasks. The algorithmic and computational improvements that were incorporated by the SVM^{light} implementation as well as the complexity issues are analyzed in [31].

4.4 Experimental Study

The goal of our experimental study is twofold. On the one hand, we wanted to get an experimental insight on the cluster-to-object assignment error introduced by the visual analysis algorithms and check whether our expectation on the most populated cluster holds. On the other hand, we aimed at comparing the quality of object models generated by the proposed framework, against the models trained by manually provided strong annotations.

4.4.1 Datasets

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 4.3.2. For this case we have used a collection of 536 images from the *Seaside* domain annotated in our lab, denoted as S^B. The second type refers to roughly-annotated datasets like the ones formed by flickr groups. In order to create a dataset of this type, 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, we refer to this dataset as S^G. The third type refers to the weakly annotated datasets like the ones found in collaborative tagging environments. For this case, we have crawled 3000 S^{F3K} and 10000 S^{F10K} images from flickr using the wget[7] utility and the flickr API facilities, in order to investigate the impact of the dataset size on the robustness of the generated models. Depending on the annotation type we use the selection approaches presented in Section 4.3.1 to construct the necessary image groups S^c. Table 3 summarizes the information of the datasets used in our experimental study.

4.4.2 Tag-Based Image Selection

As a result of our assumption on the tagging habits of social users, 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 group S^c, to increase as the volume of the initial dataset S increases. This is evident in the case of keyword-based search 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 appearing object. Similarly, in the case of flickr groups, since a user has decided to assign an image to a

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

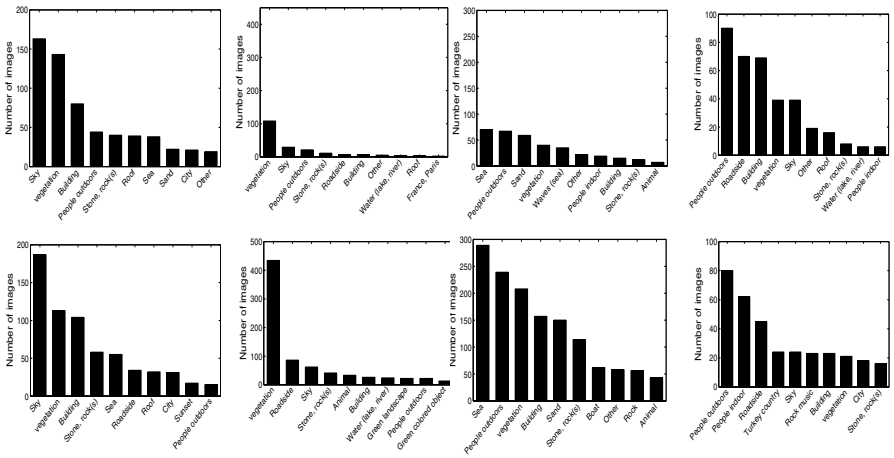
Table 3. Datasets Information

Symbol	Annotation Type	No. of Images	objects	Selection approach
S^B	strongly annotated	536	sky, sea, vegetation, person, sand, rock, boat	keyword based
S^G	roughly-annotated	4000 (500 for each object)	sky, sea, vegetation, person, car, grass, tree, building	flickr groups
S^{F3K}	weakly annotated	3000	cityscape, seaside, mountain, roadside, landscape, sport-side	SEMSOC
S^{F10K}	weakly annotated	10000	jaguar, turkey, apple, bush, sea, city, vegetation, roadside, rock, tennis	SEMSOC

group titled with the name of the object, the probability of this image to depict 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 groups created to emphasize on objects *sky*, *sea*, *vegetation*, *person*, respectively. These image groups were generated from both S^{F3K} and S^{F10K} using SEMSOC. Each of the bar diagrams depicted in Fig. 10, describes the distribution of objects' #*appearances* inside an image group 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.

4.4.3 Clustering Assessment

The purpose of this experiment is to provide an insight on the validity of our approach in always selecting the most populated cluster for training a model recognizing an object described by the most frequently appearing tag. In order to do so we evaluate the content of each of the formulated clusters using the strongly annotated dataset S^B. More specifically, $\forall c_i$ depicted in S^B we obtain $S^{c_i} \subset S^B$ and apply



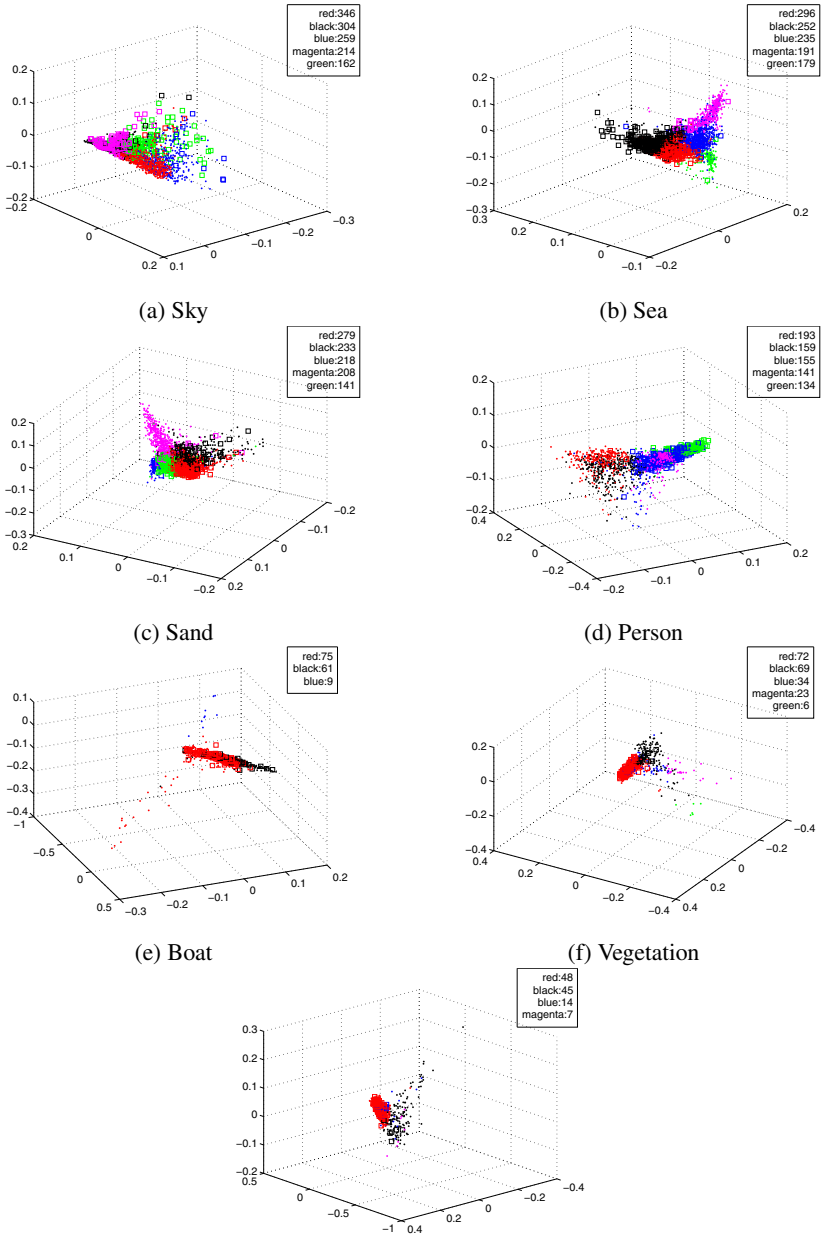
(a) Sky (b) Vegetation (c) Sea (d) Person

Fig. 10. Distribution of objects' #appearance in an image group S^c, generated from S^{F3K} (upper line) and S^{F10K} (lower line) using SEMSOC

clustering on the extracted regions. In Fig. 11 we visualize regions distributions among the generated 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 is used to indicate a cluster's rank according to their population (i.e., red: 1st, black: 2nd, blue: 3rd, magenta: 4rth, green: 5th, cyan: 6th). 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 4 (i.e., *sand, vegetation, rock, boat*) out of 7 examined cases. In the cases of objects *sea*, *sky* and *person*, the error introduced from visual analysis, prevents clustering from assigning the regions of interest into the same cluster.

4.4.4 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 ended up with 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. The factor limiting the number of benchmarking objects is on the one hand the need to have strongly annotated images for these objects and from the other hand the un-supervised nature of SEMSOC that restricts the eligible objects to the ones



(a) Sky (b) Sea

(c) Sand (d) Person

(e) Boat (f) Vegetation

(g) Rock

Fig. 11. Regions distribution amongst clusters. The regions depicting the object of interest are marked in squares, while the other regions are marked in dots. Squares being painted in colors other than red, indicate false negatives and dots painted in red indicate false positives. This Figure is best viewed in color with magnification.

identified by clustering the images in the tag information space. 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. F-Measure was used for measuring the efficiency of the models.

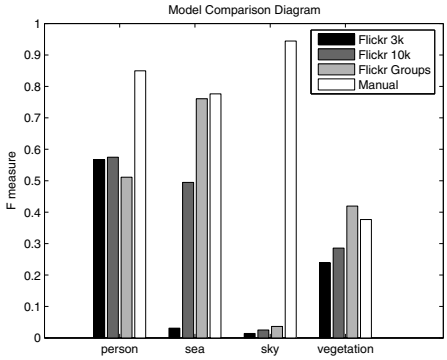


Fig. 12. Performance comparison between four object recognition models that are learned using samples of different annotation quality (i.e., strongly, roughly and weakly)

By looking at the bar diagram of Fig. 12, 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 collect 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*). However, the efficiency of the models trained using weakly annotated samples is likely to be improved if the size of the dataset is increased.

From the bar diagram of Fig. 12 it is clear that when using the S^{F10K} the incorporation of more indicative examples into the training set improves the generalization ability of the generated models in all four cases. However, 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 alleviates the error introduced by visual analysis algorithms and allows the proposed method to select the appropriate cluster for training the model. 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. Using this viewer to inspect the content of the formulated clusters, we realize that the selected cluster is not the one containing the regions depicting sea when using the S^{F3K}, whereas the correct cluster is selected when using the S^{F10K}. Fig. 13 and Fig. 14 show indicative images for some of the generated clusters for the object *sea* obtained using the S^{F3K} and S^{F10K} dataset respectively. The clusters' rank (#) refers to their population. We can see that when using the S^{F3K} dataset the regions depicting *sea* are split in two clusters (ranked #4 and #5), while the most populated cluster #1 consists of regions primarily depicting people. On the other hand, in the case of the S^{F10K} dataset, where the correct cluster is selected (see Fig. 14), it seems that the larger size of the utilized dataset compensates for the error introduced by the visual analysis algorithms.

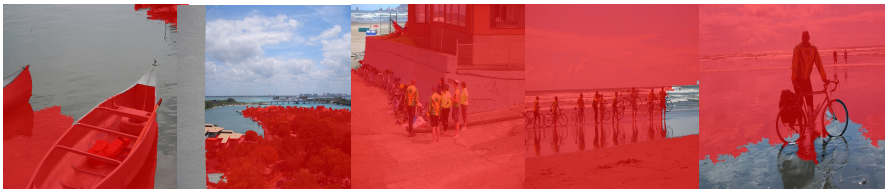
5 Related Methods

The presented method 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 [34], [16]. In [34] 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 set out to determine when and why search-based models manage to perform satisfactory and design a system for predicting the performance trade-off between annotation- and search-based models. Essentially what the authors are trying to do is 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 on when the cheaply obtained data is of sufficient quality for training robust object detectors. In [16] 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 some works [72], [65], [68] that rely on the same principle assumption with our method. In [72] 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. Subsequently, the distance between two concepts is measured using the Jensen-Shannon (JS) divergence between the constructed models. Although different in purpose from our approach the authors present some very interesting results demonstrating that the collaborative tagging environments like flickr can serve as a particular valuable source for mining the necessary information for implementing various computer vision tasks. In [65] the authors make the assumption that semantically related images usually include one or several common regions (objects) with similar visual features. Based on this

#1 Cluster - person

#2 Cluster - noise



#3 Cluster - sea

#4 Cluster - sea + sky

Fig. 13. Indicative regions from the clusters generated by applying our approach for the object *sea* generated by the S^{F3K} dataset. The regions that are not covered in red are the ones that have been assigned to the corresponding cluster.

#1 Cluster - sea

#2 Cluster - person

#4 Cluster - sand

#6 Cluster - sky

#7 Cluster - building

Fig. 14. Indicative regions from the clusters generated by applying our approach for the object *sea* generated by the S^{F10K} dataset. The regions that are not covered in red are the ones that have been assigned to the corresponding cluster.

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 [68] investigate non-expensive ways to generate annotated training samples for building concept classifiers using supervised learning. The authors utilize clickthrough data logged by retrieval systems that consists of the queries submitted by the users, together with the images in the retrieval results, that these users selected to click on in response to their queries. Although the training data collected in this way can be potentially noisy, the authors rely on the fact that clickthrough data exhibit noise reduction properties, given that they encode the collective knowledge of multiple users. The method is evaluated using global concept detectors and the conclusion that can 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.

Finally our work bares also similarities with works like [3] and [42] that operate on segmented images with associated text and perform annotation using the joint distribution of image regions and words. In [3] 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 [42] 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.

6 Conclusions

Although the quality of object detection models trained using the described method is still inferior from the one achieved using manually trained data, we have shown that under certain circumstances Social Media can be effectively used to facilitate effortless learning. Particularly encouraging was the experimental observation concerning the size of the dataset that showed a consistent improvement on all different types of objects. Given that the size of publicly available content is constantly increasing in the context of social networks, we can claim that by using a larger collection of Social Media we will eventually achieve performance similar to the one obtained using manually trained models. As a general conclusion we can say that social networks can provide more semantically enhanced media than search engines, in pretty much the same effort. Although the noise present in the tags hinders

the direct use of these media for training machine learning algorithms, the Collective Intelligence that emerges from the massive participation of users in social networks can be used to remove the need for dedicated human supervision in machine learning.

Another important issue is the computational cost of the proposed framework, especially when the size of the social dataset is large. On a core 2 duo processor running on $3.33GHz$ with $3.25GB$ of RAM, image segmentation takes place in a few seconds and the time needed for extracting the SIFT features and creating the bag-of-words representation of each region is of the same order. Similarly, the clustering of regions and the calculation of the necessary support vectors are also executed within a few seconds, on average. Thus, the time needed for analyzing a single image for the presence of a certain concept is less than a minute, enabling the proposed framework to be used in real life applications.

Acknowledgement. This work was sponsored by the European Commission as part of the Information Society Technologies (IST) programme under EC grant number IST-FP6-026978 - X-Media and the European Community's Seventh Framework Programme FP7/2007-2013 under grant agreement n215453 - WeKnowIt and the contract FP7-248984 GLOCAL.

References

1. MPEG-7 Visual Experimentation Model (XM). Version 10.0, ISO/IEC/JTC1/SC29/WG11, Doc. N4062 (2001)
2. Aurnhammer, M., Hanappe, P., Steels, L.: Augmenting navigation for collaborative tagging with emergent semantics. In: International Semantic Web Conference (2006)
3. Barnard, K., Duygulu, P., Forsyth, D.A., de Freitas, N., Blei, D.M., Jordan, M.I.: Matching words and pictures. Journal of Machine Learning Research 3, 1107–1135 (2003)
4. Begelman, G.: Automated tag clustering: Improving search and exploration in the tag space. In: Proc. of the Collaborative Web Tagging Workshop at WWW 2006 (2006)
5. Bennett, K.P., Demiriz, A., Maclin, R.: Exploiting unlabeled data in ensemble methods. In: KDD 2002: Proceedings of the eighth ACM SIGKDD international conference on Knowledge discovery and data mining, pp. 289–296. ACM, New York (2002), http://doi.acm.org/10.1145/775047.775090
6. Bezdek, J.C.: Pattern Recognition with Fuzzy Objective Function Algorithms. Kluwer Academic Publishers, Norwell (1981)
7. Biederman, I.: Recognition-by-components: A theory of human image understanding. Psychological Review 94, 115–147 (1987)
8. Breiman, L., Friedman, J., Olshen, R., Stone, C.: Classification and Regression Trees. Wadsworth and Brooks, Monterey, CA (1984)
9. d'Alché-Buc, F., Grandvalet, Y., Ambroise, C.: Semi-supervised marginboost. In: NIPS, pp. 553–560 (2001)
10. Cao, L., Luo, J., Huang, T.S.: Annotating photo collections by label propagation according to multiple similarity cues. In: MM 2008: Proceeding of the 16th ACM international conference on Multimedia, pp. 121–130. ACM, New York (2008), http://doi.acm.org/10.1145/1459359.1459376

11. Carneiro, G., Chan, A.B., Moreno, P.J., Vasconcelos, N.: Supervised learning of semantic classes for image annotation and retrieval. IEEE Trans. Pattern Anal. Mach. Intell. 29(3), 394–410 (2007)
12. Carson, C., Belongie, S., Greenspan, H., Malik, J.: Blobworld: Image segmentation using expectation-maximization and its application to image querying. IEEE Transactions on Pattern Analysis and Machine Intelligence 24, 1026–1038 (1999)
13. Comaniciu, D., Meer, P.: Mean shift: a robust approach toward feature space analysis. IEEE Transactions on Pattern Analysis and Machine Intelligence 24(5), 603–619 (2002), doi:10.1109/34.1000236
14. Conrady, R.: Travel technology in the era of Web 2.0. Trends and Issues in Global Tourism 2007. Springer, Heidelberg (2007)
15. Cour, T., Sapp, B., Jordan, C., Taskar, B.: Learning from ambiguously labeled images. In: IEEE Computer Society Conference on Computer Vision and Pattern Recognition, pp. 919–926 (2009), http://doi.ieeecomputersociety.org/10.1109/CVPRW.2009.5206667
16. Cour, T., Sapp, B., Jordan, C., Taskar, B.: Learning from ambiguously labeled images. In: IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR 2009) (2009)
17. Domingos, P., Pazzani, M.J.: On the optimality of the simple bayesian classifier under zero-one loss. Machine Learning 29(2-3), 103–130 (1997), citeseer.ist.psu.edu/domingos97optimality.html
18. Duygulu, P., Barnard, K., de Freitas, J.F.G., Forsyth, D.A.: Object recognition as machine translation: Learning a lexicon for a fixed image vocabulary. In: Heyden, A., Sparr, G., Nielsen, M., Johansen, P. (eds.) ECCV 2002. LNCS, vol. 2353, pp. 97–112. Springer, Heidelberg (2002)
19. Egmont-Petersen, M., de Ridder, D., Handels, H.: Image processing with neural networks–a review. Pattern Recognition 35(10), 2279–2301 (2002), doi:10.1016/S0031-3203(01)00178-9
20. Faloutsos, C., Barber, R., Flickner, M., Hafner, J., Niblack, W., Petkovic, D., Equitz, W.: Efficient and effective querying by image content. J. Intell. Inf. Syst. 3(3-4), 231–262 (1994), http://dx.doi.org/10.1007/BF00962238
21. Fergus, R., Li, F.F., Perona, P., Zisserman, A.: Learning object categories from google's image search. In: ICCV, pp. 1816–1823 (2005)
22. Freund, Y., Schapire, R.E.: A decision-theoretic generalization of on-line learning and an application to boosting. J. Comput. Syst. Sci. 55(1), 119–139 (1997), http://dx.doi.org/10.1006/jcss.1997.1504
23. Frey, B.J., Dueck, D.: Clustering by passing messages between data points. Science 315, 972–976 (2007), www.psi.toronto.edu/affinitypropagation
24. Ghosh, H., Poornachander, P., Mallik, A., Chaudhury, S.: Learning ontology for personalized video retrieval. In: MS 2007: Workshop on multimedia information retrieval on The many faces of multimedia semantics, pp. 39–46. ACM, New York (2007), http://doi.acm.org/10.1145/1290067.1290075
25. Giannakidou, E., Kompatsiaris, I., Vakali, A.: Semsoc: Semantic, social and content-based clustering in multimedia collaborative tagging systems. In: ICSC, pp. 128–135 (2008)
26. Giannakidou, E., Koutsonikola, V.A., Vakali, A., Kompatsiaris, Y.: Co-clustering tags and social data sources. In: WAIM, pp. 317–324 (2008)
27. Golder, S.A., Huberman, B.A.: The structure of collaborative tagging systems. CoRR abs/cs/0508082 (2005)

28. Grahl, M., Hotho, A., Stumme, G.: Conceptual clustering of social bookmarking sites. In: 7th International Conference on Knowledge Management (I-KNOW 2007), Know-Center, Graz, Austria, pp. 356–364 (2007)
29. Gruber, T.: Ontology of folksonomy: A mash-up of apples and oranges (2005), http://tomgruber.org/writing/ontology-of-folksonomy.htm
30. Jaschke, R., Hotho, A., Schmitz, C., Ganter, B., Stumme, G.: Trias–an algorithm for mining iceberg tri-lattices. In: ICDM 2006: Proceedings of the Sixth International Conference on Data Mining, pp. 907–911. IEEE Computer Society, Washington (2006), http://dx.doi.org/10.1109/ICDM.2006.162
31. Joachims, T.: Making large-scale support vector machine learning practical, pp. 169–184 (1999)
32. Johnson, S.: Hierarchical clustering schemes. Psychometrika 32(3), 241–254 (1967)
33. Joshi, D., Luo, J.: Inferring generic activities and events from image content and bags of geo-tags. In: CIVR 2008: Proceedings of the 2008 International Conference on Content-based Image and Video Retrieval, pp. 37–46. ACM, New York (2008), http://doi.acm.org/10.1145/1386352.1386361
34. Kennedy, L.S., Chang, S.-F., Kozintsev, I.: To search or to label?: predicting the performance of search-based automatic image classifiers. In: Multimedia Information Retrieval, pp. 249–258 (2006)
35. Kennedy, L.S., Naaman, M., Ahern, S., Nair, R., Rattenbury, T.: How flickr helps us make sense of the world: context and content in community-contributed media collections. In: ACM Multimedia, pp. 631–640 (2007)
36. Leibe, B., Leonardis, A., Schiele, B.: An implicit shape model for combined object categorization and segmentation. In: Toward Category-Level Object Recognition, pp. 508–524 (2006)
37. Leistner, C., Grabner, H., Bischof, H.: Semi-supervised boosting using visual similarity learning. In: CVPR (2008)
38. Li, F.F., Fergus, R., Perona, P.: One-shot learning of object categories. IEEE Trans. Pattern Anal. Mach. Intell. 28(4), 594–611 (2006)
39. Li, F.F., Perona, P., Technology, C.I: A bayesian hierarchical model for learning natural scene categories. In: CVPR, vol. 2, pp. 524–531 (2005)
40. Li, J., Wang, J.Z.: Real-time computerized annotation of pictures. In: MULTIMEDIA 2006: Proceedings of the 14th Annual ACM International Conference on Multimedia, pp. 911–920. ACM, New York (2006), http://doi.acm.org/10.1145/1180639.1180841
41. Li, J., Wang, J.Z.: Real-time computerized annotation of pictures. IEEE Trans. Pattern Anal. Mach. Intell. 30(6), 985–1002 (2008), http://dx.doi.org/10.1109/TPAMI.2007.70847
42. Li, L.-J., Socher, R., Fei-Fei, L.: Towards total scene understanding: Classification, annotation and segmentation in an automatic framework. In: IEEE Conference on Computer Vision and Pattern Recognition (2009)
43. Li, Y., Shapiro, L.G.: Consistent line clusters for building recognition in cbir. In: ICPR, vol. (3), pp. 952–956 (2002)
44. Lowe, D.: Object recognition from local scale-invariant features. In: The Proceedings of the Seventh IEEE International Conference on Computer Vision, vol. 2, pp. 1150–1157 (1999), doi:10.1109/ICCV.1999.790410
45. Lowe, D.: Distinctive image features from scale-invariant keypoints. Int. J. Comput. Vision 60(2), 91–110 (2004), http://dx.doi.org/10.1023/B:VISI.0000029664.99615.94

46. Lukaszyk, S.: A new concept of probability metric and its applications in approximation of scattered data sets. Computational Mechanics 33, 299–304 (2004), http://www.ingentaconnect.com/content/klu/466/2004/00000033/00000004/art00007
47. MacQueen, J.B.: Some methods for classification and analysis of multivariate observations. In: Cam, L.M.L., Neyman, J. (eds.) Proc. of the fifth Berkeley Symposium on Mathematical Statistics and Probability, vol. 1, pp. 281–297. University of California Press, Berkeley (1967)
48. Mallapragada, P.K., Jin, R., Jain, A.K., Liu, Y.: Semiboost: Boosting for semi-supervised learning. IEEE Transactions on Pattern Analysis and Machine Intelligence 31(11), 2000–2014 (2008), doi:10.1109/TPAMI.2008.235
49. Marlow, C., Naaman, M., Boyd, D., Davis, M.: Ht06, tagging paper, taxonomy, flickr, academic article, to read. In: Hypertext, pp. 31–40 (2006)
50. Meadow, C.T.: Text Information Retrieval Systems. Academic Press, Inc., Orlando (1992)
51. Meyer, D., Leisch, F., Hornik, K.: The support vector machine under test. Neurocomputing 55(1-2), 169–186 (2003)doi:10.1016/S0925-2312(03)00431-4, http://www.sciencedirect.com/science/article/B6V10-49CRCBP-1/2/346ddc665b1b67be089a7d5d46edca07
52. Mezaris, V., Kompatsiaris, I., Strintzis, M.G.: Still image segmentation tools for object-based multimedia applications. IJPRAI 18(4), 701–725 (2004)
53. Mezaris, V., Kompatsiaris, I., Strintzis, M.G.: Still image segmentation tools for object-based multimedia applications. IJPRAI 18(4), 701–725 (2004)
54. Mika, P.: Ontologies are us: A unified model of social networks and semantics. Web Semant. 5(1), 5–15 (2007), http://dx.doi.org/10.1016/j.websem.2006.11.002
55. O'Really, T.: What is Web 2.0: Design Patterns and Business Models for the Next Generation of Software. O'Reilly Media Inc., Sebastopol (2005)
56. Palen, L., Hiltz, S.R., Liu, S.B.: Online forums supporting grassroots participation in emergency preparedness and response. Commun. ACM 50(3), 54–58 (2007), http://doi.acm.org/10.1145/1226736.1226766
57. Quack, T., Leibe, B., Gool, L.J.V.: World-scale mining of objects and events from community photo collections. In: CIVR, pp. 47–56 (2008)
58. Russell, B.C., Freeman, W.T., Efros, A.A., Sivic, J., Zisserman, A.: Using multiple segmentations to discover objects and their extent in image collections. In: CVPR, vol. (2), pp. 1605–1614 (2006)
59. van de Sande, K., Gevers, T., Snoek, C.: Evaluating color descriptors for object and scene recognition. IEEE Transactions on Pattern Analysis and Machine Intelligence 99(1) (doi:5555), http://doi.ieeecomputersociety.org/10.1109/TPAMI.2009.154
60. Schmitz, P.: Inducing ontology from flickr tags. In: Proc. of the Collaborative Web Tagging Workshop (WWW 2006) (2006), http://www.rawsugar.com/www2006/22.pdf
61. Scholkopf, B., Smola, A., Williamson, R., Bartlett, P.: New support vector algorithms. Neural Networks 22, 1083–1121 (2000)
62. Shi, J., Malik, J.: Normalized cuts and image segmentation. In: IEEE Computer Society Conference on Computer Vision and Pattern Recognition, vol. 0, p. 731 (1997), http://doi.ieeecomputersociety.org/10.1109/CVPR.1997.609407
63. Sivic, J., Russell, B.C., Efros, A.A., Zisserman, A., Freeman, W.T.: Discovering objects and their localization in images. In: ICCV, pp. 370–377 (2005)

64. Sivic, J., Zisserman, A.: Video google: A text retrieval approach to object matching in videos. In: ICCV 2003: Proceedings of the Ninth IEEE International Conference on Computer Vision, p. 1470. IEEE Computer Society, Washington (2003)
65. Sun, Y., Shimada, S., Taniguchi, Y., Kojima, A.: A novel region-based approach to visual concept modeling using web images. In: ACM Multimedia, 635–638 (2008)
66. Sung, K.K., Poggio, T.: Example-based learning for view-based human face detection. IEEE Trans. Pattern Anal. Mach. Intell. 20(1), 39–51 (1998)
67. Torralba, A.B., Murphy, K.P., Freeman, W.T.: Contextual models for object detection using boosted random fields. In: NIPS (2004)
68. Tsikrika, T., Diou, C., de Vries, A.P., Delopoulos, A.: Image annotation using clickthrough data. In: 8th ACM International Conference on Image and Video Retrieval, Santorini, Greece (2009)
69. Vasconcelos, M., Vasconcelos, N., Carneiro, G.: Weakly supervised top-down image segmentation. In: CVPR, vol. (1), pp. 1001–1006 (2006)
70. Viola, P.A., Jones, M.J.: Rapid object detection using a boosted cascade of simple features. In: CVPR, vol. (1), pp. 511–518 (2001)
71. Wang, Z., Feng, D.D., Chi, Z., Xia, T.: Annotating image regions using spatial context. In: International Symposium on Multimedia, vol. 0, pp. 55–61 (2006), http://doi.ieeecomputersociety.org/10.1109/ISM.2006.32
72. Wu, L., Hua, X.-S., Yu, N., Ma, W.-Y., Li, S.: Flickr distance. In: ACM Multimedia, 31–40 (2008)
73. Yanai, K.: Generic image classification using visual knowledge on the web. In: ACM Multimedia, 167–176 (2003)
74. Zhang, J., Marszalek, M., Lazebnik, S., Schmid, C.: Local features and kernels for classification of texture and object categories: A comprehensive study. Int. J. Comput. Vision 73(2), 213–238 (2007), http://dx.doi.org/10.1007/s11263-006-9794-4