International Journal of Applied Information Systems (IJAIS) – ISSN : 2249-0868
Foundation of Computer Science FCS, New York, USA
Volume 5– No.1, January 2013 – www.ijais.org

Unsupervised Object Annotation through Context Analysis

A.M. Riad
Dean of Faculty of Computers, Mansoura University Mansoura, Egypt

Hamdy.K. Elminir
Head of electronic and communication Dept, Misr Higher Institute of Engineering and Technology Mansoura, Egypt

Sameh Abd-Elghany
Faculty of Computers and Information Sciences, Mansoura University
Mansoura, Egypt

ABSTRACT
The goal of object level annotation is to locate and identify instances of an object category within an image. Nowadays, Most of the current object level annotation systems annotate the object according to the visual appearance in the image. Recognizing an object in an image based visual appearance yield ambiguity in object detection due to appearance confusion for example "sky" object may be annotated as "water" according to similarity in visual appearance. As a result, these systems don't recognize the objects in an image accurately due to the lack of scene context. In the task of visual object recognition, scene context can play important role in resolving the ambiguities in object detection. In order to solve the ambiguity problem, this paper presents a new technique for a context based object level annotation that considers both the semantic context and spatial context analysis to reduce ambiguous in object annotation.

General Terms
Image Annotation and Retrieval

Keywords
Image Annotation; Semantic Context; Objects Recognition.

1. INTRODUCTION
Online photo-sharing web sites, such as Flicker, encourage internet users to share their personal photos on the web, as well as to manually annotate the photos with tags (i.e., keywords). however, the manual tags on the images websites suffer from significant problems, such as incomplete ("missing") for tags, misspelling, noisy ("incorrect") tags ,and subjectivity and time consuming[1],[2]. To reduce these problems, automatic image annotation has received a lot of attention recently. In most current researches, image annotation refers to the process of automatically labelling the image contents with a predefined set of concepts representing the semantic content of images, and that method can be called the traditional image annotation. A number of approaches have been proposed in the literature on automatic image annotation. These approaches belong to two categories. In The first category, image annotation is formulated as a supervised classification problem and Different machine learning techniques are used to predict the annotations of new images [3-5]. The second category learn the correlation between image features and textual words from the examples of annotated images and then apply the learned correlation to predict words for unseen images[6-9]. These methods infer the association between the images and their related tags only at the image level, and utilize the image-to-image visual similarity to refine or predict Image tags. However, the annotations generated at image level models present poor performance as a result two images with partial common keywords may be considerably different in terms of image features, and thus image level similarity may be inadequate to describe the similarity of the underlying concepts among the images. To achieve reliable and visible image annotation and retrieval systems, it is critical to obtain the exact correspondence between the tags and the individual regions within an image [10]. However, in practice, it is tedious if not impossible task to manually assign each tag to its corresponding region, for large-scale image set and most users are prefer to annotate the tags at the image level. This inspires interesting and practically valuable research problems which automatically reassign the tags annotated at the image level to those derived semantic regions.

Currently, Traditional automatic object detection solutions are developed for specific applications and usually consider only one tag type, e.g., faces, locations, or events and can be formulated as a process of classification, in which images are automatically classified into a set of pre-defined categories (keywords). Namely, given a set of training images with keywords that describe image semantic contents, low-level features of the training images are extracted. Then, classifiers are constructed with low-level features to give the class decision. Lastly, the trained classifiers are used to classify new instances and annotate unlabelled images automatically. Before a learning machine can perform classification, it needs to be trained first, and training samples need to be accurately labelled[11]. The labelling process suffers from these problems (a) time consuming, (b) error-prone, (c) limited word vocabulary, (d) extensive manual works. Furthermore, the rapid growth of new multimedia data makes the trained models outdated quickly and difficult to scale up.

This paper presented a semantic object level annotation approach by extracting semantic properties of images tags in conjunction with image segmentation algorithm to reassign the image labels to those objects. The proposed approach involves three kinds of critical relations in image annotation; one is the object- to -object relation,-word-to-image relation and word-to-word relation. These are combined together to perform semantic object detection. The result is the ability to automatically formulate annotations and localization to large numbers of objects in side images which endow image retrieval systems with a new level of semantic richness. The proposed framework has the following contributions: 1) it can remove most of the unrelated tags associated with each image

to avoid the propagation of incorrect information, since each image will be connected to a small number of most probably concept-related samples; 2) it is robust to noise in the visual features; since the proposed annotation technique rely on object level not image level. 3) Compared to existing approaches that require training for each object category, the proposed is naturally effective for large-scale applications since the proposed approach does not rely on supervised learning requirement. 4) The proposed approach is able to solve the ambiguous in object level annotation by taking into account the semantic context and spatial context.

This paper is organized as follows; section 2 presents the related work. Section 3 illustrates the proposed framework in details. Section 4 shows the experiments and results. Section 5 concludes this work.

2. Related Work

Most existing works of image annotation simply construct a graph to model the relationship among individual images, and a single edge is connected between two vertices for capturing the image-to-image visual similarity [12-15]. Nevertheless, using a single edge to model the relationship of two images is inadequate in practice, especially for the real-world images typically associated with multiple tags [16].To effectively search the visual content of the Internet, it is necessary to recognize the object categories present in an image [17]. In earlier works to object recognition are relying on strongly supervised learning of the visual appearance of object categories an. For instance, [18] uses manual annotations to train face detection classifiers, [19] proposes a method for the recognition of buildings, while [20] introduces an implicit shape model for the detection of cars. The common assumption shared by these approaches is that they require huge human-labelled collection of training images, each containing an instance of the object. Such supervised learning methods are not practical due difficult to scale, out-of vocabulary problem, and being time-consuming in acquiring training data and learning. [21].

There exist many related works have been proposed to solve supervised learning problems in object recognition and perform annotation and recognition at region level using unsupervised approaches [22-24]. In [22] Proposed to perform object localization, but it focused on single object category or assumed there is no overlapping between multiple objects in the training images. In [23] the visually similar objects are clustered together then the frequent item set mining rule to annotate each cluster with most frequent tag. [24] Presents an algorithm for extracting region level annotations from flickr images using a small set of manually labelled regions to guide the selection process. However, these works tried to tackle the problem of annotation at image level; the relationship between tags has not been considered. They all neglected the semantic relatedness between different tags. Taking into account semantic relations between tags improves the recognition rate [25].

The use of semantic representations for image retrieval has already been explored in previous works for example [26-29] which propose to refine the tags based on the visual and semantic consistency residing in the images, and assign similar tags to visually similar images. Although this work benefits from the semantic relatedness between tags, it performs this task at image level not at region level. To address the scalable object recognition problem, we propose to propagate the labels annotated at the image-level to those local semantic regions. Generally, one label of an image only characterizes a single local semantic region, and two images with common labels often share similar semantic regions. Inversely, if two local semantic regions from different images are visually similar, there usually exist labels in the two images are semantically similar. The proposed approach recognizes the object based on two viewpoints, objects visual similarity and tags semantic similarity. This technique based on the idea that if two regions are visually similar in different images, the semantic annotations for these objects would be related. Thus the proposed approach presents a new combined semantic feature-based models and regional visual feature based models to perform unsupervised object recognition.

The proposed approach does not ignore the spatial information between different objects in the image and knowing the label for one object in image provides information about the labels of other objects. For example, the knowledge of object being sea is informative about objects satisfying the "upper" relationship with respect to it, since they are highly likely to be a boat.

3. The context based object recognition framework

3.1 General overview

The main contribution of the paper is label-to-region assignment task. The proposed approach is built on nearest-neighbor hypothesis (i.e., visually similar objects likely share semantically similar tags). Advantage of the proposed approach is that it does not requires a training stage which depends on number of training images in order to build a good model for the target object recognition. Given a set of images and their associated tags, the contribution of this research is to segment image into a set of meaningful objects (regions), and then to find out the proper associations between tag and each object in the image.

3.2 System Architecture

The system architecture of the proposed framework is breakdown into a set of processes in order to offer the functionality. Fig.1. depicts the architecture of the framework. This section presents the proposed framework in detail.

3.2.1 Image segmentation

The purpose of image segmentation is to partition an image into meaningful regions such that each region is homogeneous and the union of no two adjacent regions is homogeneous. In this work, we use a combination of watershed segmentation algorithm and edge detection method [30], which is an efficient, automatic, and unsupervised segmentation method for gray-level images, to partition an image into non-overlapping regions.

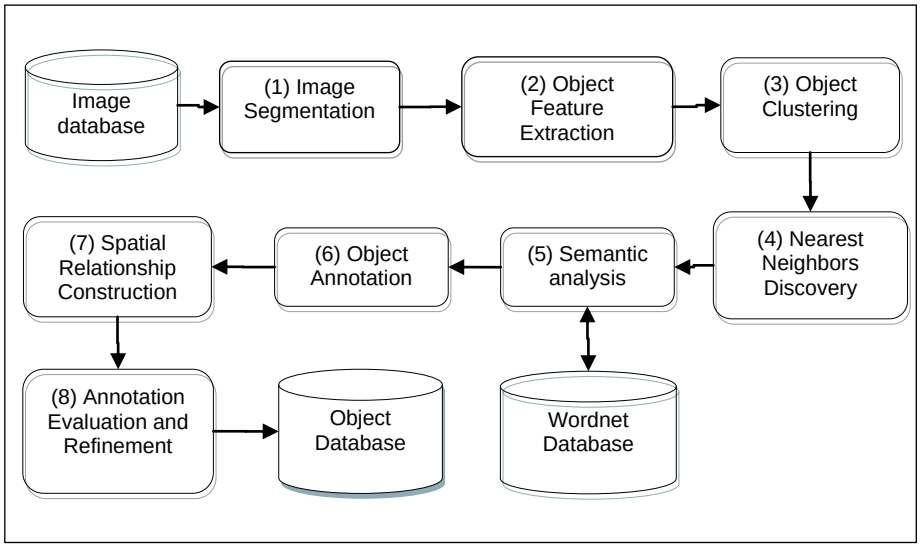


Fig 1: Semantic Object Recognition System Architecture

The framework, the converts the colour images to the grey images and then partition them by watershed segmentation and edge detection technique.

3.2.2 Object Feature Extraction

Scale Invariant Features Transform (SIFT) is an approach for detecting and extracting local feature descriptors [31]. SIFT descriptors are invariant to image scaling, transformation, rotation and partially invariant to illumination changes and affine, gives the local features of an image. Therefore, feature from the images can be extracted more accurately by using SIFT than color, texture, shape [32].

3.2.3 Object Clustering Process

After segmentation, the objects are clustered into a certain number of groups. One of the most popular and widely clustering methods that minimize the clustering error for points in Euclidean space is K-means clustering algorithm. This choice was mainly motivated by the comparably fast processing of the k-means algorithm compared to other unsupervised clustering methods [33].

3.2.4 Object Nearest Neighbours Discovery

In our framework, we use the k-nearest neighbours strategy to discover the closeness relations between the regions in a graph manner which contains edges and vertices, the vertices represented by regions itself, while the edges between vertices represented by the similarity measure between object visual features. Specifically, two regions Ri and Rj are linked with an edge if Ri is among the k1 nearest neighbours of Rj and vice versa. Here the k1 nearest neighbors are measured by the usual Euclidean distance.

Once the linkages between the regions are created, the adjacency relations between the image vertices can be naturally constructed where any two vertices with at least one edge connection are considered as adjacent. As Fig. 2.

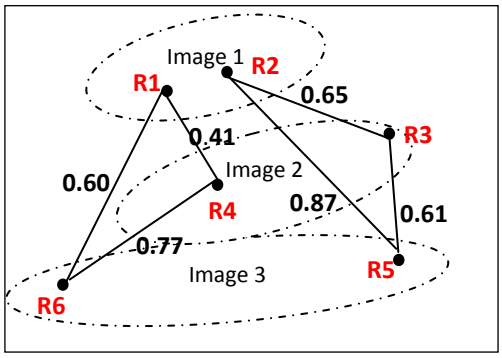


Fig.2: A simple graph o f six regions relationship

3.2.5 Semantic analysis

The major approaches for image searches that are currently available use text descriptions or tags annotating to the images. However, it is still difficult to search for images by referring to only their text information because sometimes the information is not directly linked to the image or the text may have different interpretations. In order to bridge the gap between image and text, we need to understand the image semantics.The most popular ontology used to extract semantic knowledge in image annotation is WordNet [34-39]. wordnet is a semantic lexical database providing that contains over 155,000 words arranged in hierarchical groups of related words called synsets . In Wordnet about 117, 000 synsets that linked to other synsets by means of a small number of conceptual relations [40]. These relations such as hypernyms (Y is a hypernym of X if every X is a (kind of) Y), hyponyms (Yis a hyponym of X if every Y is a (kind of) X), antonym (opposite meaning of each other) [43]. A simple way to measure the semantic similarity between two synsets is to treat taxonomy as an undirected graph and measure the distance between them in WordNet.The path length is measured in nodes/verteces rather than in links/edges. The length of the path between two members of the same synsets is 1 (synonym relations). Fig.3 shows an example of the

hyponym taxonomy in WordNet used for path length similarity measurement.

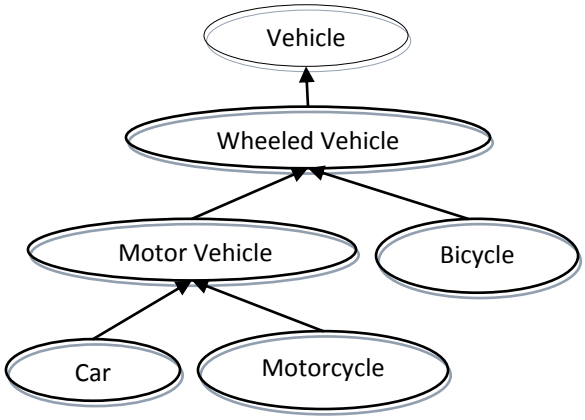


Fig 3: shows an example of the hyponym taxonomy in WordNet

The shared parent of two synsets is known as a sub-sumer. The least common subsume (LCS) of two synsets is the sumer that does not have any children that are also the sub-sumer of two synsets. In other words, the LCS of two synsets is the most specific sub-sumer of the two synsets. Back to the above example, the LCS of {car} is {motor vehicle}, since the {automotive, motor vehicle} is more specific than the common sub-sumer {wheeled vehicle}. The path length gives us a simple way to compute the relatedness distance between two word senses. when looking up a word in WordNet, the word is first lemmatized. Therefore, the distance between "car" and "cars" is 0 since they are identical. There are many proposals for measuring semantic similarity between two synsets: In this work we used the formula proposed in [41], because this measure takes into account both path length and depth of the least common sub-summer:

$$Sim(s, t) = 2 * depth(LCS)/[depth(s) + depth(t)] \quad (1)$$

• where s and t: denote the source and target words being compared.

• Depth(s): is the shortest distance from root node to a node S on the taxonomy where the synset of S lies .

• LCS: denotes the least common sub-summer of s and t.

3.2.6 Object Annotation

As a pair of visually similar images should have smaller semantic distance and vice versa [42]. It is based on the intuition that if different persons use same tags to label visually similar images, then these tags are likely to reflect the visual contents of the annotated images.[44] For each image there are a set of tags and a set of visual objects but we don't know which tag corresponds to which object, and number of objects and tags can be different; even when they are the same, we may have more object for a single tag, or more than one tag for a single object, we try to align the tags and objects. in this process let the image collection D contains D={di},i=1,….N) , di={(Xi ,Yi)} Where Xi=(xi1 , xi2,…. xin) represents the set of objects from 1 to n associated with image di and Yi=(yi1 , yi2,…. yim) represents the set of words from 1 to m associated with image di.

The steps for object recognition process are:

Step 1: find vf (Xi, Xj) where vf denotes to visual similarity function between two objects measured on low level visual feature.

Step 2: Construct the simple graph Gs. For each vertex (represented by object), based on the similarity function vf, connect it to its k-nearest neighbors(e.g. xith , xith).

Step 3: find the set of words {Yi,Yj}associated with the two images di , dj where objects xith, xith belonging to.

Step 4: find Max Sf (Yi,Yj)where Sf denotes to semantic similarity function between two tags measured on high level semantic concept

Step 5: Since object xith is most similar visual object to xith, word yith most semantically similar to word yjth while; (xith, yith) belongs to image di and (xith, yjth) belongs to image dj. Then xith object is initially recognized with word yith while object xith initially recognized with word yith.

3.2.7 patial relationship construction

As spatial relations improve results of image retrieval besides, the object localization in this work, the spatial relationships between objects has been considered.

We also present a systematic approach to employing contextual information for object recognition. The use of contextual information is very important during object recognition because it can help resolve the ambiguities due to appearance confusion in many cases. For example, a blue homogeneous object can be recognized as "water" as well as "sky" due to the similarity in appearance. However, the relation of the object A to other nouns such as the "sun" can resolve the ambiguity. If the relation below (A, sun) is more likely than in (sun, A), then the object A can be recognized as "water" (and vice-versa).

In the retrieval process, the user query in text form is automatically translated to semantic meaning and representation. Moreover the multiple objects with combined spatial relation such as Object A is Left to Object B and Right-below Object C are considered. It is fully automatic image and spatial relation semantic extraction without involving any user or relevance feedback during the retrieval process. This process automatically extracts and identifies spatial information among objects in the images. This process contains these stages:

(A) Minimum Bound Object (MBO) construction

Any image may contains a set of objects, Each of the objects is represented using Minimum Bound Region (MBR) that indicated using a box as show in figure 4, the Image objects, I={Object A, Object B, Object C}.

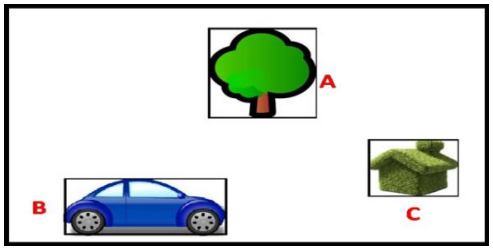


Fig4: three objects in one image

(B) Reference point estimation

Each object has a reference coordinate which indicated using the centroid MBO.

Object A=RA (xA,yA)
Object B=RB (xB,yB)
Object C=RC (xC,yC)

(C) Calculate the Slope of each pair of objects in image

The slope of each possible pair of objects is calculated based on their object's reference as in figure 5.

The slope between each two pair of objects as below:

Object A and Object B = S (AB) = $(yB - yA)/(xB - xA)$ (2)

Object A and Object C = S (AC) = $(yC - yA)/(xC - xA)$ (3)

Object B and Object C = S (BC) = $(yC - yB)/(xC - xB)$ (4)

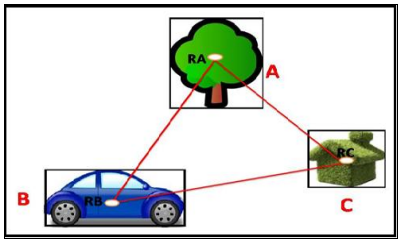
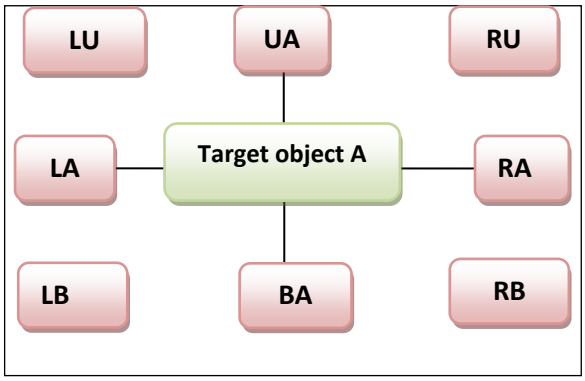


Fig 5: objects with slope between references point



(D) Spatial relationship Rule construction

There are 8 spatial relationship concepts are determined as shown in figure 6, each relationship are determined through a rule, these relations are:

LU(A,B): B is located in Left upper A
LB(A,B): B is located in Left below A
RU(A,B): B is located in Right upper A
RB(A,B): B is located in Right below A
UA (A,B): B is located in Upper- Alignment A
BA(A,B): B is located in Below- Alignment A
RA (A,B): B is located in Right- Alignment A
LA(A,B): B is located in Left- Alignment A

Spatial relationship rules are

LU (A,B) = S(A,B) $<$ 0 and xA $>$ xB, yA $>$ yB
LB (A,B) = S(A,B) $<$ 0 and xA $>$ xB, yA $<$ yB

Fig 6: spatial relationships between objects

RB (A,B) = S(A,B) $>$ 0
UA (A,B) = S(A,B) = 0 and yA $>$ yB
BA (A,B) = S(A,B) = 0 and yA $<$ yB
RA (A,B) = S(A,B) = ∞ and xA $<$ xB
LA (A,B) = S(A,B) = ∞ and xA $>$ xB

3.2.8 Object annotation evaluation and refinement

This process tries to decrease the error prone of the object annotation process that appeared in the annotation process, in this framework three levels of refinement; visual appearance based, semantic context based, spatial context based refinement. Our assumption is that if certain objects in the database are visually similar, the words of these images should also be semantically similar and vice versa. To discover the Association between the high-level concept and low-level visual features of objects, we need to quantify the visual features by clustering, because the concept space is discrete while the visual feature space is continuous in general [23]. Therefore, we aim to associate the semantic concept and the visual feature cluster.

Dentitions

An association rule is a pattern that states when C occurs, Y occurs with certain probability.

Ci: the cluster i that contains visually similar objects{x1,.. ,xk},

ycj : is a concept which initially assigned to object in cluster ci

Count of Ci: number of objects contained in cluster ci.

Support, yj, is the frequency of occurrence of concept yj in the database; a concept yj is frequent if yj's support in database is more than a minsup threshold

Support, Ci → yj, probability that a cluster ci contains yj

Confidence, c, conditional probability that a transaction having Ci also contains yj

The steps of object annotation evaluation process using association rule mining technique as follows:

(1) Scan the Database to get the support S of each concept y_i and visual cluster Ci, and select those concepts and clusters with support less than user specified minimum support.

(2) Construct the transaction database D and the basic candidate 2-itemsets based on the existing inverted file. Our goal is not the association between concept and concept. We are interested in the association between concepts and visual feature clusters. Therefore, only the objectset containing one concept and one visual feature cluster are considered. The existing inverted file relates the concepts to their associated objects.

(3) For each concept y_j in the cluster c_i, calculate the support between concept y_j and cluster c_j.

$$\text{Support}(y_j, c_i) = \text{Count}(y_i, c_i) / \text{count}(C_i) \quad (5)$$

Where count (y_j) is the frequency of occurrence of concept y_j in the cluster c_i; and count (c_i) is the total number of visual objects in the cluster.

(4) All concepts that have support less than the user specified minimum support are selected. The objects which are recognized with these concepts are considered as invalid recognized.

(5) For all frequent concepts, calculate the confidence between concept y_j and cluster c_i.

$$\text{Confidence}(y_j, c_i) = \text{Support}(y_j, c_i) / \text{Support}(y_j) \quad (6)$$

(6) The objects that have confidence <= minimum Confidence are should be invalid initially recognized.

To refine annotation;This process reconstructs the visual similarity between the objects lies in the weak rule to the next neighbour similar object and then repeating the steps for object recognition process from step 2 to step 5 presented in object annotation process; then evaluation the annotation process; and repeating these process until all objects lies in the strong rule.

Semantic context based evaluation and refinement

Most of the current object annotation systems are not categorizing the objects in an image accurately due to the lack of semantic context. In the task of visual object recognition, semantic context can play the very important role of reducing ambiguity in objects' visual appearance. Semantic context corresponds to the likelihood of an object to be found in some scenes but not others. Hence, we can define semantic context of an object in terms of its co-occurrence with other objects and in terms of its occurrence in scenes. For example the image may contains three objects, "Person", "Tennis Racket", and "Lemon". Using a recognition system without a semantic context, these labels would be final; however, in context, one of these labels is not satisfactory. Namely, the object labelled "Lemon", with an appearance very similar to a "Tennis Ball" is incorrect due to the ambiguity in visual appearance. But if we provide semantic context in which "Person", "Tennis Racket", are co-occurred, most likely with the yellow ball which is labelled as "Tennis Ball" instead of lemon. It means that the Tennis Ball" object is semantically related with "Person" and "Tennis Racket". Semantic Context is commonly obtained from strongly labelled training data, which is impractical in the era of large scale web images. In the proposed approach, the semantic context is presented in the form of a list of labels of objects indicating the occurrence of these objects together in a set of images. This process based mainly on the frequent item set algorithm [45], which is Fasting algorithms for mining association rules. Our goal is to find the association between tags, so we begin with candidate 2-tag-sets, which contain two different tags, after that, we get the support for these tag-sets and confidence as in eq. (1) and eq. (2). For all set that are not satisfy the frequent and confidence conditions are should be invalid annotated. By extending to the 3- tag-set and more, we can determine the error object label for example if there is a two instance of images m1 and m2; m1 containing two objects ("person", "Lemon") while m2 contains three objects ("person", "Lemon", "racket"). In the first image, it is not correct to consider the "lemon" is incorrectly labelled, but in the second image it is easy to discover the error in "lemon" label according to the semantic context analysis. To refine the object annotation in the image that contains objects incorrectly labelled, the image is globally matched to get those similar images, after that getting the distinct high frequency occurrence label against those images, finally assigning this label to the candidate object.

Spatial context based evaluation and annotation

We also present a systematic approach to employing spatial context analysis for object recognition. The use of spatial context information is very important during object annotation because it can help resolve the ambiguities due to appearance confusion in many cases. For example, a blue homogeneous object can be annotated as "sea" as well as "sky" due to the similarity in appearance. However, since we know that the relationship sky above sea occurs more frequently; After performing the semantic context analysis using prioir algorithm between tags and their relative location, in this step we will get the association between each pair objects and their spatial relation.

4. Implementation and evaluation

In our experiments we use The PASCAL 2007 object recognition dataset that contains nearly 10,000 images of 20 different object categories e.g. aeroplane ,bicycle , bird ,bottle , bus , car, cat , chair , cow, dining table, dog, horse, motorbike, person, potted plant, sheep , sofa , train , tv monitor. The domain of the images in this collection is very generic in that it covers a wide range of daily life situations with many different images of similar visual content but with varying illumination, viewing angle and background. The recognition was performed on the ground truth segmentations.

In these experiments, we present some analysis to evaluate the contribution of the different forms of context. The Average Precision (AP) is used to evaluate and compare the various methods, which is defined as the average of the precisions computed at all recall levels. The Mean Average Precision (MAP) is the average of the APs across all queries.

4.1 Appearance based object annotation

In Figure 8 several examples are shown where appearance confusion was occurred in different images examples. Let us consider the last example, where the test image contains bird. The appearance alone labels the objects as aeroplane, and vice versa. In the appearance-only scenario, the MAP estimates of the data terms were used to label the segments.

Figure 8 examples for appearance confusion between bird and aeroplane

The person, chair and train get high accuracy because their visual appearance is rarely confused with the other objects, while cat, dog and horse, cow objects get less accuracy they visually confused together.

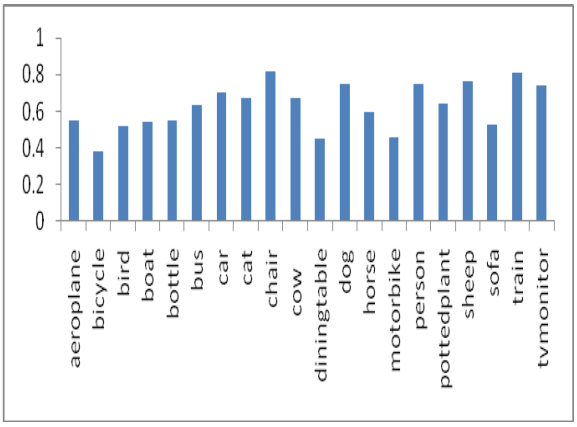


Fig 9: mean average precision for the 20 categories in the VOC dataset using appearance based visual recognition.

4.2 semantic based object annotation

Using the priori data mining algorithm the support/confidence object association was estimated and the result is shown in table [1]. In This table we show only the occurrence of objects that contains association rule with support over 0.002and confidence over 0.1.Categories such as dog , cat and horse gain significantly from context. Their appearance cues are very similar but they are very strongly associated with other categories (chair, person respectively) whose appearance cues are quite reliable.

Table 1 The 2nd support/confidence object association

Keyword 1	Keyword 2	Support	confidence
horse	person	0.01756147	1
car	person	0.01505268	0.8571429
motorbike	person	0.01455093	0.8285714
bottle	person	0.01304566	0.7428572
bicycle	person	0.01204215	0.6857143
person	bicycle	0.01204215	0.6857143
chair	sofa	0.01154039	0.6571429
person	diningtable	0.01154039	0.6571429
chair	diningtable	0.01103863	0.6285715
chair	tvmonitor	0.009533367	0.5428572
person	pottedplant	0.009031611	0.5142857
bus	car	0.008028098	0.4571429
person	tvmonitor	0.008028098	0.4571429
bus	person	0.00652283	0.3714286
person	bus	0.00652283	0.3714286
person	sofa	0.006021074	0.3428572
chair	pottedplant	0.005519317	0.3142857
person	cow	0.005519317	0.3142857
person	train	0.005017561	0.2857143
train	person	0.005017561	0.2857143
bottle	diningtable	0.004515805	0.2571429
pottedplant	sofa	0.004515805	0.2571429
boat	person	0.004014049	0.2285714
person	boat	0.004014049	0.2285714
aeroplane	person	0.003512293	0.2
chair	bottle	0.003512293	0.2
bottle	tvmonitor	0.003010537	0.1714286
sofa	diningtable	0.003010537	0.1714286
bird	person	0.002508781	0.1428571
cat	person	0.002508781	0.1428571
person	bird	0.002508781	0.1428571
bicycle	car	0.002007025	0.1142857
car	bicycle	0.002007025	0.1142857
cat	sofa	0.002007025	0.1142857
diningtable	tvmonitor	0.002007025	0.1142857

4.3 spatial based recognition

As in figure 10 , the first image contains person and horse objects while in the second image contains dog object, the appearance based object recognition should present confusion in dog and horse objects since they are visually nearly similar. Since the person may be come with dog in the same image and also come with horse in the same image , the system will not be able to detect error in horse detection in co-occurrence based object recognition the co-occurrence test left the labels untouched. However, the location based recognition discarded the possibility of the person being above the dog according to the support/confidence relation table[3] the person always com a above the horse and never come above the dog. So the error label in the first image will be modified to be a horse which matches the ground truth labeling

Fig 10: example of location based object recognition in voc 2007 image dataset.

Average Mean precision for the 20 categories in the VOC 2007 dataset using appearance alone, and using subsequently more complex context models with appearance. This figure indicates that the object recognition based combined appearance and context outperforms the traditional methods that depends only on visual appearance for object recognition specially in images with multiple different objects. While Some categories such as bird and aeroplane do not receive significant benefit from Co-occurrence context due to peculiarities of the dataset, such as they rarely co-occur with other objects.

Table 1 support/confidence relation

Keyword 1	Keyword 2	location	Support	conf
car	person	Right-Below	0.01154	1
bicycle	person	Left-Below	0.009031	0.7826
bottle	person	Right-Below	0.007526	0.6521
person	car	Right-Below	0.007526	0.6521
bicycle	person	Right-Below	0.006522	0.5652
person	horse	Right-Below	0.00652	0.5652
motorbike	person	Left-Below	0.00602	0.5217
person	chair	Right-Upper	0.00551	0.4782
person	tvmonitor	Right-Below	0.005519	0.4782
diningtable	person	Left-Below	0.005017	0.4347
tvmonitor	person	Right-Below	0.005017	0.4347
chair	tvmonitor	Right-Below	0.004515	0.3913
chair	diningtable	Left-Below	0.004014	0.3478
person	bicycle	Right-Below	0.004014	0.3478
chair	sofa	Right-Upper	0.003512	0.3043
chair	person	Right-Upper	0.003010	0.2608
person	bicycle	Left-Below	0.003010	0.2608
person	dog	Right-Below	0.003010	0.2608
pottedplant	chair	Right-Below	0.003010	0.2608
bottle	chair	Right-Below	0.002508	0.2173
chair	pottedplant	Left-Below	0.002508	0.2173
horse	person	Left-Below	0.002508	0.2173
person	sofa	Right-Upper	0.002508	0.2173

5. Conclusion and Future Work

After a review of existing techniques related to automatic image annotation, we point out that these methods are not powerful enough to retrieve efficiently annotate images including semantic concepts. This paper proposes a framework that combines object recognition, semantic annotation, data mining and image visual features processing. This system expected to improve the automatic image annotation process since the system remove the noise tags by associating each image with only the semantic shared tags that

collected as a voting process from different images. Moreover, this system uses different context method to solve the ambiguous in image visual appearance problem, which indicate that the combined appearance and context performs much better than appearance information alone .The future work will be concerns of involving the colour feature in the object detection process example need "red car". Also, we will extend the proposed approach to Web video domain.

6. REFERNCES

[1] Sevil, S. G.; Kucuktunc, O.; Duygulu, P. & Can, F. (2010), 'Automatic tag expansion using visual similarity for photo sharing websites.', *Multimedia Tools Appl.* **49** (1) , 81-99 .

[2] Weinberger, K. Q.; Slaney, M. & van Zwol, R. (2008), Resolving tag ambiguity., *in* Abdulmotaleb El-Saddik; Son Vuong; Carsten Griwodz; Alberto Del Bimbo; K. Selçuk Candan & Alejandro Jaimes, ed., 'ACM Multimedia' , ACM, , pp. 111-120 .

[3] arneiro, G.; Chan, A. B.; Moreno, P. J. & Vasconcelos, N. (2007), 'Supervised Learning of Semantic Classes for Image Annotation and Retrieval.', *IEEE Trans. Pattern Anal. Mach. Intell.* **29** (3) , 394-410.

[4] Zhang, L. & Ma, J. (2011), 'Image annotation by incorporating word correlations into multi-class SVM.', *Soft Comput.* **15** (5) , 917-927 .

[5] S. Zhang, B. Li, and X. Xue, "Semi-automatic dynamic auxiliary-tag-aided image annotation", presented at Pattern Recognition, 2010, pp.470-477.

[6] Ding, G.; 0001, J. W.; Xu, N. & 0014, L. Z. (2009), Automatic Image Annotations by Mining Web Image Data., *in*, 'ICDM Workshops' , IEEE Computer Society, , pp. 152-157 .

[7] Wang, X.-J.; 0001, L. Z.; Jing, F. & Ma, W.-Y. (2006), AnnoSearch: Image Auto-Annotation by Search., *in* 'CVPR (2)' , IEEE Computer Society, , pp. 1483-1490 .

[8] Llorente, A.; Motta, E. & Rüger, S. M. (2009), Image Annotation Refinement Using Web-Based Keyword Correlation., in 'SAMT' , Springer, , pp. 188-191 .

[9] Weston, J.; Bengio, S. & Usunier, N. (2010), 'Large scale image annotation: learning to rank with joint word-image embeddings', Machine Learning 81 , 21-35 .

[10] Liu, D.; Hua, X.-S. & Zhang, H.-J. (2011), 'Content-based tag processing for Internet social images.', Multimedia Tools Appl. 51 (2) , 723-738 .

[11] Liu, D.; Hua, X.-S.; Yang, L. & Zhang, H.-J. (2009), Multiple-Instance Active Learning for Image Categorization., in Benoit Huet; Alan F. Smeaton; Ketan Mayer-Patel & Yannis S. Avrithis, ed., 'MMM' , Springer, , pp. 239-249 .

[12] Liu, J.; Wang, B.; Lu, H. & Ma, S. (2008), 'A graph-based image annotation framework.', *Pattern Recognition Letters* **29** (4) , 407-415

[13] Wang, X.-J.; Ma, W.-Y.; 0001, L. Z. & Li, X. (2005), Multi-graph enabled active learning for multimodal web image retrieval., *in* HongJiang Zhang; John R. Smith & Qi Tian, ed., 'Multimedia Information Retrieval' , ACM, , pp. 65-72

[14] Jing, Y. & Baluja, S. (2008), 'VisualRank: Applying PageRank to Large-Scale Image Search.', *IEEE Trans. Pattern Anal. Mach. Intell.* **30** (11) , 1877-1890 .

[15] Wu F, Han YH, Zhuang YT, "Multiple hypergraph clustering of Web images by mining Word2Image correlations", presented at COMPUTER SCIENCE AND TECHNOLOGY , 2010, pp 750-760

[16] Liu, D.; Yan, S.; Rui, Y. & Zhang, H.-J. (2010), Unified tag analysis with multi-edge graph., *in* 'ACM Multimedia' , ACM, , pp. 25-34 .

[17] Fergus, R.; 0002, F.-F. L.; Perona, P. & Zisserman, A. (2010), 'Learning Object Categories From Internet Image Searches.', *Proceedings of the IEEE* **98** (8) , 1453-1466 .

[18] Viola, P. & Jones, M. (2001), ' Rapid Object Detection using a Boosted Cascade of Simple Features' ' Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition' , Hawaii .

[19] Li, Y. & Shapiro, L. G. (2002), Consistent Line Clusters for Building Recognition in CBIR., *in* 'ICPR (3)' , pp. 952-956 .

[20] Leibe, B.; Leonardis, A. & Schiele, B. (2006), An Implicit Shape Model for Combined Object Categorization and Segmentation., *in* Jean Ponce; Martial Hebert; Cordelia Schmid & Andrew Zisserman, ed., 'Toward Category-Level Object Recognition' , Springer, , pp. 508-524 .

[21] Hsieh, L.-C. & Hsu, W. H. (2010), Search-Based Automatic Image Annotation via Flickr Photos Using Tag Expansion., *in* 'ICASSP' , IEEE, , pp. 2398-2401 .

[22] Chen, Y.; Zhu, L.; Yuille, A. L. & Zhang, H. (2008), Unsupervised learning of probabilistic object models (POMs) for object classification, segmentation and recognition., *in* 'CVPR' , IEEE Computer Society, .

[23] He, R.; Xiong, N.; Yang, L. T. & Park, J. H. (2011), 'Using Multi-Modal Semantic Association Rules to fuse keywords and visual features automatically for Web image retrieval.', *Information Fusion* **12** (3) , 223-230 .

[24] hatzilari, E.; Nikolopoulos, S.; Papadopoulos, S.; Zigkolis, C. & Kompatsiaris, Y. (2011), Semi-supervised object recognition using flickr images., *in* José M. Martinez, ed., 'CBMI' , IEEE, , pp. 229-234 .

[25] Barrat, S. & Tabbone, S. (2010), 'Modeling, classifying and annotating weakly annotated images using Bayesian network.', *J. Visual Communication and Image Representation* **21** (4) , 355-363 .

[26] Liu, D.; Hua, X.-S.; Wang, M. & Zhang, H.-J. (2010), Image retagging., *in*, 'ACM Multimedia' , ACM, , pp. 491-500 .

[27] Chang, C.-Y.; Wang, H.-J. & Li, C.-F. (2009), 'Semantic analysis of real-world images using support vector machine.', *Expert Syst. Appl.* **36** (7) , 10560-10569 .

[28] Rahman, M. M.; Bhattacharya, P. & Desai, B. C. (2009), 'A unified image retrieval framework on local visual and semantic concept-based feature spaces.', *J. Visual Communication and Image Representation* **20** (7) , 450-462 .

[29] Wong, R. C. F. & Leung, C. H. C. (2008), 'Automatic Semantic Annotation of Real-World Web Images.', *IEEE Trans. Pattern Anal. Mach. Intell.* **30** (11) , 1933-1944 .

[30] Salman, N. (2006), 'Image Segmentation Based on Watershed and Edge Detection Techniques.', *Int. Arab J. Inf. Technol.* **3** (2) , 104-110 .

[31] Lowe, D. G. (2004), 'Distinctive Image Features from Scale-Invariant Keypoints', *Int. J. Comput. Vision* **60** (2) , 91--110 .

[32] Z. Wang, Y. Mei, F. Yan, "A New Web Image Searching Engine by Using SIFT Algorithm", in Proc. WISM, *2009*,pp 366-370

[33] Quack, T.; Mönich, U.; Thiele, L. & Manjunath, B. S. (2004), Cortina: a system for large-scale, content-based web image retrieval., *in*, 'ACM Multimedia' , ACM, , pp. 508-511

[34] Deng, J.; Dong, W.; Socher, R.; Li, L.-J.; Li, K. & 0002, F.-F. L. (2009), ImageNet: A large-scale hierarchical image database., *in* 'CVPR' , IEEE, , pp. 248-255 .

[35] Liu, D.; Hua, X.-S.; Wang, M. & Zhang, H.-J. (2010), Retagging social images based on visual and semantic consistency., *in* Michael Rappa; Paul Jones; Juliana Freire & Soumen Chakrabarti, ed., 'WWW' , ACM, , pp. 1149-1150 .

[36] Kilinç, D. & Alpkocak, A. (2011), 'An expansion and reranking approach for annotation-based image retrieval from Web.', *Expert Syst. Appl.* **38** (10) , 13121-13127 .

[37] Yang, C.; Dong, M. & Fotouhi, F. (2005), I2A: an interactive image annotation system., *in* 'ICME' , IEEE, , pp. 948-951 .

[38] Jin, Y.; 0021, L. W. & Khan, L. (2005), Improving Image Annotations Using WordNet., *in* K. Selçuk Candan & Augusto Celentano, ed., 'Multimedia Information Systems' , Springer, , pp. 115-130 .

[39] Z. Wang, K. Jia, P. Liu," A Novel Image Retrieval Algorithm Based on ROI by Using SIFT Feature Matching" in Proc. MultiMedia and Information Technology ,2008,pp 338-341

[40] The Wordnet website. [Online]. Available: http://wordnet.princeton.edu

[41] Verb Semantics and Lexical Selection

[42] G. Qi, X. Hua, and H. Zhang, "Learning semantic distance from community-tagged media collection", in Proc. ACM Multimedia, 2009, pp.243-252.

[43] Wang, Y. & Gong, S. (2007), Refining image annotation using contextual relations between words., *in* Nicu Sebe & Marcel Worring, ed., 'CIVR' , ACM, , pp. 425-432 .

[44] Li, X.; Snoek, C. G. M. & Worring, M. (2009), 'Learning Social Tag Relevance by Neighbor Voting.', IEEE Transactions on Multimedia 11 (7) , 1310-1322 .

[45] Agrawal ,R. and Srikant, R(1994).. Fast algorithms for mining association rules. VLDB'94.