2012 International Conference on Advanced Computer Science Applications and Technologies

Semantic Similarity Measure with Conceptual Graph-Based Image Annotations

Nutchanun Chinpanthana
Faculty of Information Technology, Dhurakij Pundit University
110/1-4 Prachachuen Rd. Laksi, Bangkok 10210, Thailand
nutchanun.cha@dpu.ac.th

Abstract— **This paper presents a novel approach of the semantic similarity measure that support the image retrieval systems. The approach is composed of five stages: (1) data collection, (2) image annotation, (3) conceptual graph representation, (4) similarity matching, and (5) shows a semantic search result. First stage is collecting the contents into database archive. LabelMe tool is used to annotate images. Next stage is representing an image into the conceptual graph. Third stage is finding the similarity matching between the conceptual graph and representative graph. Last stage is showing the set semantic of image results. The results are compared to the classification methods. The experimental results indicate that our proposed approach offers significant performance improvements in the interpretation of semantic images, compared, with the maximum of 88.8% accuracy.**

Keywords; semantic images; image retrieval; similarity matching; graph representation

I. INTRODUCTION

Digital media has been a phenomenal growth in the field of computer vision. Traditional methods in multimedia retrieval have managed search like a query-by-example paradigm as a Content-Based Image Retrieval (CBIR). CBIR is developing techniques that support effective finding or browsing in a large corpus image data. Image is usually represented as a set of low level features: color, texture, or shape which extracted using image processing algorithms. Too many researchers are finding and representing the semantics associated with homogeneous content [1]. Later research group has introduced the representation with segmentation is applied to decompose an image into the homogenous regions, which correspond to objects: VisualSeek[2], NETRA [3] and Blobworld [4]. System is finding the relative images by using an example image at the beginning of a query. After that, it returns a set of relevance images from measure similarity of low-level features between two images. The result, however, is still unsatisfactory due to the inclusion of irrelevant images. These systems failed to retrieve many relevant images moreover retrieved too many irrelevant images.

Although CBIR has been extensively studied, it has been abundance of prior work. Due to the problem of visual similarity, it is not a semantic similarity. The systems retrieved images based on the corresponding the set of features not with a concept of semantic images. There is a gap between low-level image features and semantic meanings. To overcome these difficulties, many researches used text to represent the image contents. Annotation texts playing an increasingly vital role, images are annotated manually with keywords and then retrieved them by their annotations [5]. Unfortunately, manually annotating a lot of images by hand is too tedious and time-consuming task. To solve this problem, many automatic image annotation systems have been proposed in recent year. Most of the existing the annotation techniques are the classification images. Classification relies on training a classifier that is defined as the task of taking a keyword of the dataset and assigning it into their several categories. Many researches have been used, such as the two-dimensional multi-resolution hidden Markov model [6], support vector machine [7] , and Mixture Hierarchical Model [8]. Even there have been many techniques on image annotations; the features do not adequately into the semantic concepts. In order to reduce this gap, a promising paradigm of multimedia retrieval has been introduced into the form of content description, transcribed text, or captions in the past few years. The extraction of meaningful image features and an understanding are required for supporting the retrieval systems. The paper begins with reviewing related work on traditional techniques in multimedia retrieval. This is followed by an explanation of the conceptual graph representation. A set of experiments to classify the semantic images is then presented. The paper concludes with a discussion of the result of the proposed method and directions of the future work.

The rest of paper is organized as follows. In Section II, we first present the overview of related work in image annotation systems. The proposed conceptual graph representation is described in Section III. Section IV represents the semantic similarity measure to find the relevance images. The traditional classification is presented in Section V. After that, Section VI gives comprehensive experimental results, and the conclusions and discussion in the future work are given in Section VII.

978-0-7695-4959-0/13 $25.00 © 2013 IEEE
DOI 10.1109/ACSAT.2012.33

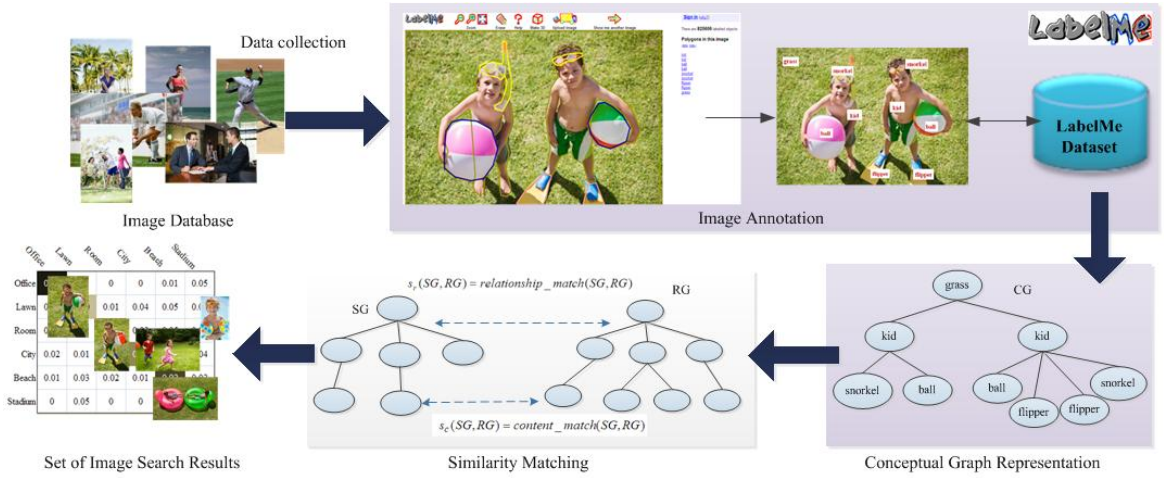


Figure 1. Framework of graph-based image annotations

II. OVERVIEW OF PURPOSED SYSTEM

In this paper, we purpose a novel system of semantic images. The system can retrieve a set of relevance semantic images with semantic similarity measure. The proposed is represented by using the conceptual graph. We then describe in the details of automatic image annotation tools and the major combining of image components. Overall structure of proposed system is shown in the Figure 1. It contains five stages: (1) data collection, (2) image annotation, (3) conceptual graph representation, (4) similarity matching, and (5) shows a semantic search result.

The first thing to consider within the proposed approach of semantic image representation is what image information to use in a conceptual graph. We first concentrate on the image contents how we find the essential meaningful features. Based on our observation, a semantic image is emerging from major contents and the association between image contents. Each content has different types however some types possibly have similar semantics. For example, "car" and "bicycle" are semantically analogous because they are both instances of "vehicle". This relation might be proven useful in the part of semantically related contents. Hence, we mapped various contents into a predefined ontology of keyword archive from WordNet [10]. But, automatic construction of concept ontologies is in general a hard problem and ill-defined. LabelMe [9] is popular annotation tools that are evaluated with ontology and linked to WordNet [10]. The LabelMe database is the most comprehensive public image database that is manually annotated by online volunteers. Therefore, we use image contents automatically annotated from LabelMe Tool [9] that contains large scale image collections.

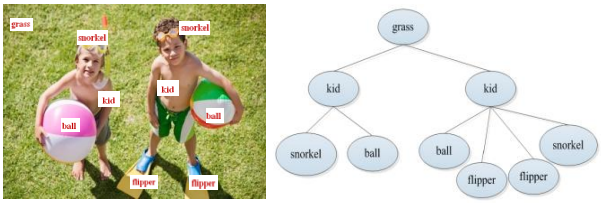
For our purpose, we chose the class object, which is defined in Wordnet as "physical object". This definition seemed to be suitable for our purposes because contents are supposed to be visible. We initially expected that most of contents would appear within the class "object". We then chose a large class entity at this top level. Wordnet defines an entity as having "a distinct separate existence (living or nonliving)". In general, one possible way to describe ontologies [11] can be formalized as: $o = \{C, \{R_{c_i,c_j}\}\}$. O is an ontology, C is a set of concepts described by the ontology, c_i and c_j are two concepts, $c_i, c_j \in C$ and $R_{c_i,c_j} : C \times C$ is the semantic relation amongst these concepts.

III. CONCEPTUAL GRAPH REPRENTATION

In semantic representative, we briefly describe how to construct an image graph that defined on the edges reflect semantic relationships between images. We construct a conceptual graph that represents the contents and spatial relationships among objects by a simple form of concept nodes. In our model, the concepts in the previous section are represented by a simple form of concept nodes. We integrated all components by using graph representation called a conceptual graph.

Formally, CG is defined by spatial entities represented as a set of vertices V and binary spatial relationships represented as a set of edges E : $CG \equiv \langle V, E \rangle$. Let C be the set of all concepts, and image contents, then a content $c \in C$ of the image is represented in the graph by vertex $c \in V$. The link relation between two nodes $a, b \in V$ of the image is represented by graph's edge $e_{ab} \equiv \langle v_a, v_b \rangle \in E$. For example, to represent the relationship between nodes c_1 and c_2 we defined an edge (c_1, c_2) . To clarify the explanation, we use the example shown in Figure 2 that shows the description of the proposed image (CG) .

The representative graph (RG) can extract the structural information embedded in the CG . For implemented, we are discovery a structure of contents and an associated content from each CG. The incorporation of the structure of contents with corpus statistics can be built



(a) (b)

Figure 2. (a) An example of the lawn image. (b) The semantic conceptual graph model

a representative graph where the actual understanding of the semantics is unobtainable. By doing so, the statistics model can take advantage of a conceptual space. Calculating the semantic association can be transformed to the estimation of the conceptual similarity (or distance) between nodes (contents or concepts) and the relationship patterns in the representative graph.

We evaluate the semantic representation between node based and edge based approaches, which correspond to the information content approach, respectively. Following the notation in information theory, the information content (IC) of a concept/class c can be quantified as follows:

$$IC(c_i) = \log^{-1} P(c_i) \tag{1}$$

where $P(c_i)$ is the probability of encountering an instance of concept C_i.

$$P(c_i) = freq(c_i) \Big/ \sum_i freq(c_i), \tag{2}$$

where $freq(c_i)$ represents the frequency of the concept in the graph. $f\ r(c_i) = \sum_{w \in c_i} f\ r(w) / |\ l(w)|$. Following the standard argument of information theory, we are considering the link strength of an edge that links a parent node to a child node. The link strength (LS) is taking the negative logarithm of the above probability. The strength of a child link is proportional to the conditional probability of encountering an instance of the child concept c_i given an instance of its parent concept p: $P(c_i | parent(c_i))$. We then obtain the following formula:

$$LS(c_i) = -\log(P(c_i | parent(c_i)) = IC(c_i) - IC(parent(c_i)),$$

where

$$P(c_i / parent(c_i)) = P(c_i \cap parent(c_i)) / P(parent(c_i)),$$

$$P(c_i / parent(c_i)) = P(c_i) / P(parent(c_i)), \tag{3}$$

Considering other factors, the relation pattern is a more natural and a direct way of evaluating semantic images. Human demonstrates a pattern of activity by relating with other contents. Therefore, we estimate the linkage between nodes and body parts which correspond to the type of relation being compared. We consider three factors in the relationship pair. The set of edge pair between two nodes c_a and c_b is defined, $R_i = \{(c_a, c_b), ..., (c_m, c_n)\}$. The relationship pattern probability can be simplified as follows:

$$LP(R_i) = freq(R_i) / \sum_i freq(R_i) \tag{4}$$

where $freq(Pat_i)$ represent the frequency of the pattern relation Pat_i in the data set.

Algorithm 1: Content_Matching.

1. input: Conceptual Graph CG, Representative Graph RG;
2. output: the content matching coefficient.
3. BEGIN
4. Initialize $c_i \in SG, c_j \in RG$ where c_i, c_j are leaves, $cm = 0, total\_cm = 0.$
5. **While** $SG(c_i)$ is not NULL
6. if $find\_content(SG(c_i) == RG(c_j))$
7. $cm = cm + LS(RG(c_j)).$
8. $increment(total\_cm, i).$
9. else if $find\_content(SG(parent(c_i)) == RG(parent(c_j)))$
10. $cm = cm + LS(RG(parent(c_j))).$
11. $increment(total\_cm, i).$
12. else increment i.
13. end if
14. end if
15. **end while**
16. RETURN $\left(\frac{cm}{total\_cm} \right)$.

Algorithm 2: Relationship_Matching.

1. input: Conceptual Graph CG, Representative Graph RG;
2. output: the relationship matching coefficient.
3. BEGIN
4. Initialize $R_i \in SG, R_j \in RG, rm = 0, total\_rm = 0.$
5. **While** $SG(R_i)$ is not NULL
6. if $find\_relationship(SG(R_i) == RG(R_j))$
7. $rm = rm + LP(RG(R_j)).$
8. $increment(total\_rm, i).$
9. end if
10. **end while**
11. RETURN $\left(\frac{rm}{total\_rm} \right)$.

IV. Semantic Similarity Measuring

In this phase, we use a semantic schema matching that is a task of finding similar structures between the representative graph (RG) and the semantic conceptual

graph (SG). The SG is matching the structure of their contents in RG. We are computing similarity coefficients between elements of the two graphs. The coefficients in the [0, 1] range, are calculating in 2 steps. The first step, called a content matching, matches the labeled contents of the SG with the RG. A main step of process is finding the contents based on the ontology archive. The similarity are mapped into a content coefficient, s_c:

$$s_c(SG,RG) = content\_match(SG,RG) \quad (5)$$

The second step is a relationship matching, matches between a set of relationships of SG and RG. The algorithm is iteration mapping the relationship patterns in RG. The result is a relationship similarity coefficient, s_r:

$$s_r(SG,RG) = relationship\_match(SG,RG) \quad (6)$$

From s_c and s_r, the weighted similarity w_{cr} is computed. $w_{cr} = w_c.s_c + (1-w_c).s_r$, where w_{cr} is weighted semantic similarity which is a mean of s_c and s_r. The constant w_c is in the [0, 1] range. A mapping is created by choosing pairs of schema elements with maximal weighted similarity. The output of schema matching is a weight of mapping graph.

V. TRADITIONAL CLASSIFIERS

In this section, we descript the three traditional classifiers: Bayesian classification, Support vector machine and Multi-layer perception networks.

A. Bayesian Classification

Bayesian theory [12][13] is the basis of statistical classification procedures. It requires all methods that provide the fundamental probability model for be assumptions explicitly built into models which are constructed by using the training data to estimate the probability of each class. Consider a general N-group classification problem in which each object has an associated attribute vector x of d dimensions. Let c denote the member that takes a value of c_i if an object is belong to group i. Define $p(c_i)$as the prior probability of class i and $f(X|c_i)$ as the probability density function. According to the Bayes rule:

$$p(c_i|X) = f(X|c_i)p(c_i)/f(X), \quad (7)$$

where $p(c_i|X)$ is posterior probability of class c_i and $f(X)$ is the probability density function:

$$f(X) = \sum_{i=1}^{M} f(X|c_j)p(c_i). \quad (8)$$

The Bayesian classifier learns the conditional probability of each attribute x_i (i = 1, 2, …, N) of X given the class label c_i , denoted as:

$$p(c_i|X) = \prod_{i=1}^{N} p(x_i|c_i). \quad (9)$$

Therefore, the image classification can be stated as given a set of observed features x_i from an image X, classify X into one of the classes c_i.

B. Multi-layer perception networks

Artificial neural networks architectures [12][13] have been increasingly employed to deal with many tasks of image processing especially image classification and retrieval. The neural classifier has the advantage of being fast, easily trainable and capable of creating arbitrary partitions of feature space. We used the multi-layer perception networks (MLPN) is typical neural networks. Back Propagation (BP) learning algorithm is often widely used in MLPN model, it turns the input and output problem of samples to non-linear optimization problems. It is trained by error back propagation. BP algorithm is a sort of supervised learning algorithm, it has hidden nodes. It minimizes a continuous error function or objective function. BP is a gradient descent method of training. If there is existed deviation when compared the output gained by network forward reasoning to expected output sample, the weight coefficient should be adjusted. For forward propagation, the BP algorithm is as follows:

$$Y_j = f\left[\sum_{i=1}^{n} W_{ji} \cdot X_i\right] \quad (10)$$

where X is input samples

$$W_{ji}(n+1) = W_{ji}(n) + \rho \cdot \varepsilon_j \cdot x_i \quad (11)$$

where ρ is learning rate, ε is error signal

$$\rho_j = (T_i - Y_i)f'\left[\sum_{i=1}^{n} \omega_{ji} \cdot X_i\right] \quad (12)$$

where Y is output samples, ω_{ji} is weight coefficient and T is expected output samples

$$\rho_j' = f'\left[\sum_{i=1}^{n} W_{ji} \cdot X_i\right] \cdot \sum_{i=1}^{n} \rho_k \, \omega_{ji} \quad (13)$$

where $(x) = \frac{1}{1-\exp(-x)}$, action function of networks.

C. Support Vector Machines

Support Vector Machine (SVM) [13][14] is another simple classifier that is designed for binary classification. That is, to separate a set of training vectors which belong to two different classes. Let $(x_i, y_i)_{1 \le i \le N}$ be a set of training examples, each example $x_i \in \mathbb{R}^d$, d being the dimensional of input feature space, belongs to a class labeled by $y_i \in \{-1, 1\}$. During the SVM model generation, the input vectors, are mapped into a new higher dimensional feature space. Then, an optimal separating hyperplane in the new feature space is constructed by a kernel function which products between input vectors x and , $K(x,y)$. Two most used kernel functions are Polynomial and Gaussian Radial Basis Function (RBF) kernel functions which are:

$$K_{poly}(x_i, y_j) = (x_i \cdot y_j + 1)^p, \quad (14)$$

TABLE I. CONFUSION MATRIX OF THE CLASSIFICATION RESULTS WITH BAYESIAN.

	Office	Lawn	Room	City	Beach	Stadium	Performance (%) Pr	Recall	F1
Office	0.79	0.03	0.05	0.07	0.01	0.05	84.0	79.0	81.4
Lawn	0.03	0.65	0.08	0.05	0.14	0.05	67.7	65.0	66.3
Room	0.04	0.09	0.82	0.02	0.02	0.01	83.7	82.0	82.8
City	0.07	0.04	0.01	0.78	0.04	0.07	79.6	77.2	78.4
Beach	0.01	0.13	0.02	0.03	0.74	0.07	77.9	74.0	75.9
Stadium	0	0.02	0	0.03	0	0.96	79.3	95.0	86.5
Average									78.70

TABLE II. CONFUSION MATRIX OF THE CLASSIFICATION RESULTS WITH SVM

	Office	Lawn	Room	City	Beach	Stadium	Performance (%) Pr	Recall	F1
Office	0.80	0.04	0.06	0.08	0	0.02	82.5	80.0	81.2
Lawn	0.01	0.85	0.03	0.02	0.07	0.02	75.9	85.0	80.2
Room	0.04	0.06	0.80	0.07	0.03	0	85.1	80.0	82.5
City	0.09	0.04	0.01	0.79	0.00	0.07	79.0	79.0	79.0
Beach	0.02	0.11	0.03	0.03	0.78	0.03	86.7	78.0	82.1
Stadium	0.01	0.02	0.01	0.01	0.02	0.93	86.9	93.0	89.9
Average									82.5

TABLE III. CONFUSION MATRIX OF THE CLASSIFICATION RESULTS WITH THE MULTI-LAYER PERCEPTION NETWORKS

	Office	Lawn	Room	City	Beach	Stadium	Performance (%) Pr	Recall	F1
Office	0.71	0.03	0.11	0.07	0.01	0.07	70.3	71.0	70.6
Lawn	0.08	0.65	0.04	0.05	0.11	0.07	69.9	65.0	67.4
Room	0.05	0.02	0.89	0.02	0.02	0	78.8	89.0	83.6
City	0.12	0.06	0.04	0.64	0.05	0.09	77.1	64.0	69.9
Beach	0.03	0.14	0.04	0.02	0.72	0.05	75.0	72.0	73.5
Stadium	0.02	0.03	0.01	0.03	0.05	0.86	75.4	86.0	80.4
Average									74.50

TABLE IV. CONFUSION MATRIX OF THE CLASSIFICATION RESULTS WITH SIMILARITY MATCHING

	Office	Lawn	Room	City	Beach	Stadium	Performance (%) Pr	Recall	F1
Office	0.95	0	0	0	0.01	0.05	90.5	94.1	92.2
Lawn	0.04	0.80	0.01	0.04	0.05	0.06	88.9	80.0	84.2
Room	0.03	0.01	0.85	0.02	0.05	0.04	93.4	85.0	89.0
City	0.02	0.01	0.03	0.85	0.02	0.04	92.4	87.6	89.9
Beach	0.01	0.03	0.02	0.01	0.92	0.02	87.6	91.1	89.3
Stadium	0	0.05	0	0	0	0.95	81.9	95.0	88.0
Average									88.80

p is the degree of polynomial and $K_{gaussian}(x_i, y_j) = e^{\|x_i - x_j\|^2/2\sigma}$, σ is Gaussian sigma respectively. All vectors lying on one side of the hyperplane are labelled as -1, and all vectors lying on another side are labeled as +1. The training instances that lie closest to the hyperplane in the transformed space are called support vectors. The number of these support vectors is usually small compared to the size of the training set and they determine the margin of the hyperplane, and thus the decision surface.

VI. EXPERIMENTAL RESULTS

In this section, we evaluated the semantic classification results by comparing with four methods [12]; Naive-Bayes, the multi-layer perception networks (MLPN), Support Vector Machine (SVM), and graph similarity matching, the purposed method. In this work, we used open source software Waikato Environment for Knowledge Analysis (WEKA) to apply a traditional classification [15].

A. Dataset and evaluation methods

In our experiments, we manually selected the probe images from two data sources. The dataset contains approximately 1,500 images where 750 images of natural scene selected from the Corel Image Database [16] and the Corbis Image Database [17]. We selected the important image classes that most people are six semantic classes: *office*, *home, beach, lawn, city,* and *stadium*. For evaluation of this technique, four most popular measurements have been applied precision, recall, f-measure, and accuracy [12]. Precision is defined to the total numbers of retrieved images with all corpuses, while recall is the specific related image with retrieval images. The highest value of the both measurements is 1. Their definitions are shown below.

$$\text{precision}_i = \frac{\#\text{of correctly classified images of class } i}{\#\text{of images classified to class } i}$$

$$\text{recall}_i = \frac{\#\text{of correctly classified images of class } i}{\#\text{of images in the class } i}$$

$$\text{f-measure}_i = \frac{2 \cdot \text{precision}_i \cdot \text{recall}_i}{\text{precision}_i + \text{recall}_i}$$

$$\text{accuracy} = \frac{\#\text{of correctly classified images}}{\#\text{of images}}$$

B. Experimental results

The experiment, we compare four types classification method for searching a set of semantic images. Table 1

to 4 show the confusion matrices and evaluation methods. Each column of the matrix represents one class and shows how the instances of this class are classified. Each row represents the instances that are predicted to belong to a given class, and shows the true classes of these instances. Comparing the results, we can observe that the accuracy of stadium class in Bayesian provide higher than other methods as shown in Table 1. The results of lawn class in SVM can achieve the better accuracy of 85% when the lawn class in Bayesian and MLPN gains 65%. Whereas the city and beach class with the SVM only 79% and 78%, similarity matching obtains up to 85% as show in Table 2. The similarity matching seems the best classifier since it can produce the highest average accuracy of 88.8%, compared to 82.5% for SVM, 78.7% for Bayesian, 74.5% for MLPN. The results of similarity matching show in Table 4. We are concluding that, the accuracy of the similarity matching based on graph representation that are suitable for semantic classification. The example of classification results as shown in Figure 3.

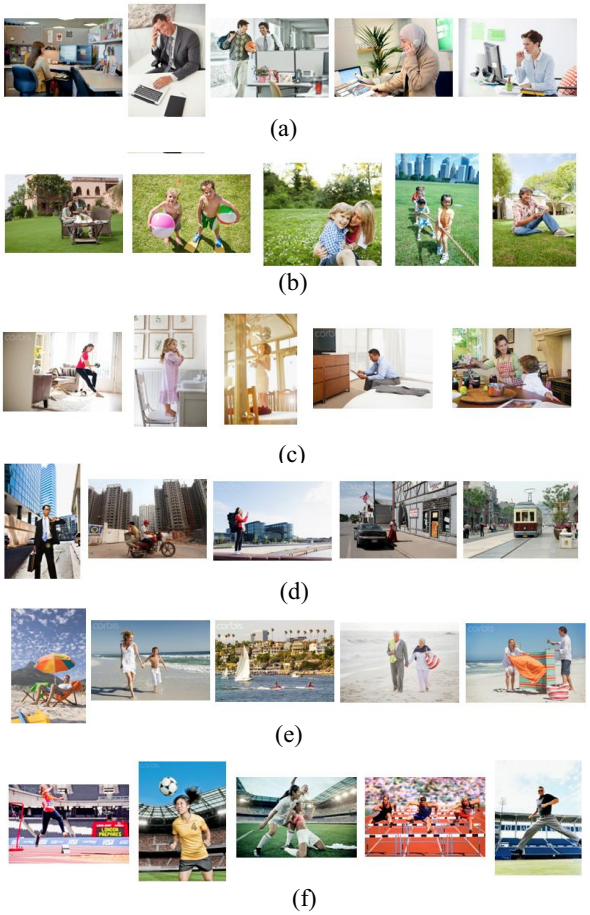


Figure 3. The example of images in each class (a) office (b) lawn (c) room (d) city (e) beach (f) stadium.

VII. CONCLUSIONS

Researchers have attempted available methodologies and techniques to interpret the semantic images. In This paper, we proposed a novel technique to improve semantic performance results in image retrieval systems. The semantic conceptual graph can be representing the concept of image contents. Then, the images can be classified into the high-level semantics. The results indicated that the proposed method offers good classification of semantics. To improve the algorithm to be able to classify more the group of semantic concepts and human activities by adding more human features is interesting for the future work. Also, a new method for representing features of the image will be defined to make the classification work on more human semantic.

REFERENCES

[1] Jeroen Steggink and Cees G. M. Snoek, "Adding Semantics to Image-Region Annotations with the Name-It-Game," Multimedia Systems, vol. 17, iss. 5, pp. 367-378, 2011

[2] Smith, J., and Chang, S., "VisualSEEK: a fully automated content-based image query system," In Proc. of the fourth ACM Int. Conf. on Multimedia, pp. 87-98, 1996.

[3] Ma, W. and Manjunath B.,"NETRA: A toolbox for navigating large image database," *In Proc. of the IEEE Int. Conf. on Image Processing*, Vol. 1, pp. 568-571, 1997.

[4] Carson, C., et al., "Blobworld: A system for region-based image indexing and retrieval," *In Proc. Int. Conf. Visual Information System*, 1999.

[5] Changhu Wang, Feng Jing, Lei Zhang, Hong-Jiang Zhang: Scalable search-based image annotation. Multimedia Syst. 14(4): 205-220, 2008.

[6] Li, J., Wang, J.Z., "Automatic linguistic indexing of pictures by a statistical modeling approach," IEEE Trans. Pattern Anal. Mach. Intell. **25**(9), 1075–1088, 2003.

[7] Gao, Y., Fan, J., Luo, H., Xue, X., Jain, R., "Automatic image annotation by incorporating feature hierarchy and boosting to scale up SVM classifiers," In: Proceedings of ACM ultimedia, Santa Barbara, 2006.

[8] Carneiro, G., Vasconcelos, N., " Formulating semantic image annotation as a supervised learning problem," In: Proceedings of CVPR, 2005.

[9] Russell, B.C., Torralba, A., Murphy, K.P., Freeman, W.T, "LabelMe: a database and web-based tool for image annotation," Int. J. Comput. Vis. 77(1–3), 2008.

[10] Miller, George A., "WordNet: An on-line lexical database," *International Journal of Lexicography*, 3:235–312, 1990.

[11] Gruber, T. R., "A Translation Approach to Portable Ontology Specifications," *Knowledge Acquisition*, Vol.5, No.2, pp. 199-220, 1993.

[12] T. Mitchell, *Machine Learning*. McGraw Hill , 1997.

[13] R.C. Holte. Very simple classification rules perform well on most commonly used datasets. Machine Learning, Vol. 11, No. 1, pp. 63–90. 1993.

[14] Richard O. Duda, Peter E. Hart and David G. Stork, Pattern classification, New York, Wiley, 2nd, 2001.

[15] Website at http:// www.cs.waikato.ac.nz/ml/weka.

[16] The Corel Corporation home page: http://www.corel.com/

[17] The Corbis Corporation home page: http://pro.corbis.com/