Unsupervised Learning Objects Categories using Image Retrieval System

Karina Ruby Perez Daniel, Enrique Escamilla Hernandez,
Mariko Nakano Miyatake, and Hector Manuel Perez Meana

National Polytechnic Institute IPN, ESIME Culhuacan, Av. Santa Ana No. 1000 Col. San, Francisco Culhuacan, Del. Coyoacan, Zip Code 04430, Mexico City
krperezd@hotmail.com,eescamillah@ipn.mx,mariko@infinitum.com.mx,
hmpm@prodigy.net.mx
http://www.posgrados.esimecu.ipn.mx

Abstract. Since several years ago artificial intelligent systems have become in a big challenge and learning of object categories is one of the most important parts in this field. Unsupervised learning of object categories provides the considerably high intelligence to apply several ambitious tasks, such as robot vision and powerful image retrieval engine, etc. In the learning object categories, a fast and accurate unsupervised learning model is required. In this paper we propose an unsupervised learning method to categorize the objects using images retrieved by Internet, in which a keyword is introduced as input data. For this purpose, all retrieved images are described using the Pyramid of Histogram of Oriented Gradients (PHOG) algorithm and the resultant PHOG vector is clustered to get a dataset for learning object categories. Two clustering methods are used in the proposed method, which are K-means and Chinese Restaurant Process (CRP), to make the learning method more efficient and simple.

Key words: Unsupervised learning objects, PHOG, k-means, Chinese restaurant process.

1 Introduction

Learning Object Categories is a very important tool in computer vision systems, which has attracted the researchers' attention during the last several years. Most of currently learning object methods are based on the manually gathered and labeled images [1–3]. However recently, with the fast developing of internet and fast growing users number, the requirement of more efficient methods have stimulated the developing of new learning methods to handle the images retrieved by internet [4–7]. Because of that it is a fast developing field in which many researchers around the world are concentrating their efforts, given as a result the developing of learning objects methods to be used with internet connection based on labels or word annotations [8], complex images training to classify them [9] or probabilistic methods using text and images. However, when used in internet such methods still retrieve several images unrelated with the keyword.

To reduce this problem, a learning object method in which the Chinese Restaurant Process CRP is used as a clustering method for learning object purposes is proposed since, as shown in [10–12], CRP is very simple and efficient for clustering data making it possible the unsupervised learning of objects and the image classification according to their vector features. Thus, the learning object method proposed in this paper is based on a simple techniques that let us building a visual model from a query (word) given by the user, where the main the aim is to construct a visual representation of any object without previous knowledge about it.

Is well known that is almost impossible to store an image database large enough to represent all the existing objects related to a given keyword provided for a given user, thus the web appears to be a desirable alternative and then the internet connection is fundamental for the proposed system. To search the image associated to the given keyword avoids the construction of an extensive database, however many images obtained from internet may have few relation with the desired object and then these images must be filtered out. In order to filter the image database obtained from internet to achieve the acquisition of objects concept, all images must be clustered according to the similarity existing among their main features. This process, if it is possible, must be carried out in an unsupervised way. To this end several unsupervised and semi-supervised clustering algorithms have been proposed in the literature. Among them, one of the most widely used is the k-means algorithm. The K-means [13] is easy to implement, simple and efficient, however this method needs the number of clusters as an input. To solve this problem, this paper proposes to use "Chinese Restaurant Process" (CRP) [12] as clustering method. CRP implements a model-based Bayesian clustering algorithm, in which the cluster assignment procedure can be regarded as an iterative Chinese restaurant process. The CRP, unlike the K-means, is a probabilistic method and do not need the number of clusters as an input.

Taking in account the above mentioned issues, this paper proposes a learning object algorithm in which, all gathered images from Internet are transformed into vectors features which are clustered using CRP according to the similarities existing among their main features. To this end, firstly the feature extraction is done using PHOG (Pyramid of Histogram of Oriented Gradients) [14, 15] method. Then the PHOG, CRP and color segmentations are combined to achieve the Learning Object Category. Here as first step the K-means algorithm is used and next the CRP is used in order to get a successful method at learning from Google images. Finally the "Ground Trut" test was used as evaluation criteria. The rest of this paper is organized as follows, in Section 2 a brief description of PHOG, K-means, color segmentation and CRP is given, in Section 3 provides the proposed algorithm. Section 4 provides the experimental results and finally in section 5 the conclusion of this work is given.

2 Basic Concepts used in the Proposed Algorithm

In the proposed algorithm, several important tools, such as Pyramid of Histogram of Oriented Gradient (PHOG), K-means clustering algorithm, object segmentation and Chinese Restaurant Process (CRP) are used. In this section we describe basic concept of the before mentioned algorithms in general manner.

2.1 Pyramid of Histogram of Oriented Gradient (PHOG)

PHOG is a global feature descriptor based on distribution of the edge direction of the image, then using PHOG the global shape of each object in the image can be extracted as a vector representation. Therefore recently PHOG is considered as adequate tool for the image classification [4, 7, 9, 14]. PHOG extracts an image description based on hierarchical representation which consists of several levels of descriptions. In the first level, Histogram of Oriented Gradients (HOG) is applied into the original whole image, while in the subsequent levels, the image or sub-image is segmented into four non-overlapped sub-images and a HOG is applied to each of them. Once the HOG vectors of sub-images of each level are obtained, the final PHOG vector is obtained concatenating each single HOG vector. The detail operation of PHOG is described below.

Firstly, the edge contours of the entry image must be extracted using the Canny edge detector. The resultant edge image is split into four non-overlapped sub-images called cell in the first level of pyramid, in the second level of pyramid each cell of first level of pyramid moreover is split into four non-overlapped cells. Consecutively this operation is done until L level of PHOG. The HOG operation is applied to each cell of each level of pyramid, getting histogram of the direction of existent edges in each cell. This operation is performed using Sobel operator of 3 x 3 without Gaussian smoothing filter. The edges direction is divided into N intervals, which forms N bins of a histogram of a single cell. The values of all bins of the histogram of a single cell form a vector of N elements, called HOG vector. Once HOG vectors of all cells are obtained, these are concatenated at each pyramid level, which means the HOG vector of 0-level pyramid is concatenated with four HOG vectors of 1-level pyramid, and so on. The concatenated vectors form the PHOG vector, which introduces the spatial information of the image, giving the ability of detection of global shape and also local features of the object, which corresponds to human learning mechanism of objects [14, 15]. The number of elements (vector size) of PHOG vector is determined by number of bins of each cell, which is given by 1.

$$PHOG\ vector\ size = N \sum_{l \in L} 4^l \tag{1}$$

Where N is the number of bins, l is number of bins used in each cell and L is total pyramid levels. If we use $L = 0$ (only original whole image is used) and $N = 20$, PHOG vector is a 20-dimentioanl vector. Thus if $L = 1$, and $N = 20$, PHOG vector has 100 elements, when $L = 2$ and $N = 40$, the size of PHOG

vector is 840. Worth noting that if the HOG quantize 20 edge direction ($N = 20$), the range of orientation angle $[0, 180]$ is divided by 20 and if $N = 40$, the angle range is $[0, 360]$, but the interval of each angle is same with $N = 20$.

2.2 K-means

The K-means clustering algorithm [13] is an unsupervised method to cluster input feature vectors into some meaningful subclasses, i. e., the members of the same cluster share similar features while the members from different clusters are sufficiently different each other. Considering that each input feature vector x_i is d-dimension vector, the data set X is given by 2.

$$X = \{x_i \mid x_i \in R^d,\ i = 1, 2, \ldots M\} \tag{2}$$

where M is number of input vectors.

The K-means clustering algorithm [16] is described as follows:

1. *Initialization:* k-centroids $(c_1, c_2, \ldots, c_k)$ of k clusters $C_1, C_2, \ldots, C_k$ are randomly selected from data set X. These centroids are initial cluster centers of each cluster.
2. *Assignment:* Each elements $x_i (i = 1 \ldots M)$ of the data set X is assigned to a cluster with closest centroid from x_i which is determined taking account of minimum distance between and all centroids. That is if $d(x_i, c_j) < d(x_i, c_m)$ for all $m = 1, \ldots, k$; $j \neq m$ then x_i is assigned to the cluster C_j.
3. *Updating:* Recalculate the centroids $c_1^*, c_2^*, \ldots, c_k^*$ of clusters, using members of clusters.
4. *Iteration:* Repeat steps 2 and 3 until the centroids no longer move. That is if $c_i^* = c_i$ for all $i = 1, \ldots, k$ then the current $c_1^*, c_2^*, \ldots, c_k^*$ are considered as the final *cluster centroids*, otherwise assign $c_i = c_i^*$ and then repeat steps 2 and 3.

Finally all elements of data set X are classified into k clusters. A principal inconvenience of the K-means algorithm is that the number of cluster must be determined in advance. In the many applications, this number is unknown.

2.3 Object Segmentation

In almost all image processing techniques, complex background in image causes several difficulties an adequate process. Then the complex background must be discarded using image segmentation method, before the principal processing. The image segmentation is typically used to locate objects and boundaries in the images. A segmented region of an image should be uniform and homogeneous with respect to some characteristic such as color, intensity or texture. Therefore the image segmentation provides homogeneous regions.

Although according to Lucchese and Mitra [17], the object segmentation algorithms can be divided into feature-space, image-domain and physics based

techniques, all these techniques use a same assumption that color is a constant property of the surface of each object. The formal definition of the object segmentation is given as following way [18].

Let $\mathcal{I}$ denote an image and let $\mathcal{H}$ define a color homogeneity; then the image $\mathcal{I}$ is segmented into $\mathcal{N}$ regions $\mathcal{R}_n$, $n = 1, 2, \ldots, \mathcal{N}$ such that

1. $\bigcup_{n=1}^{N} \mathcal{R}_n = \mathcal{I}$ with $\mathcal{R}_n \cap \mathcal{R}_m \neq \emptyset$, $n \neq m$, i.e., states that the union of all region cover the whole image.
2. $\mathcal{H}(\mathcal{R}_n) = \text{true}\ \forall\ n$ states that each region has to be color homogeneous, and
3. $\mathcal{H}(\mathcal{R}_n \cup \mathcal{R}_m) = \text{false}\ \forall\ \mathcal{R}_n$ and $\mathcal{R}_m$ adjacent, i.e., two adjacent region cannot be merged into a single region that satisfies the color homogeneity $\mathcal{H}$

Nowadays, there are several color-spaces used in different applications, although the RGB color-space is most commonly used color-space, it doesn't represent the color perception of human visual system. In this sense, the color-space HSV (Hue, Saturation and Value) is considered as better color-space than RGB [17–19], because HSV is more intuitive than RGB. For example, the variation of the saturation (S) presents the variation of perceptual color intensity and the variation of value (V) presents the perceptual illumination intensity. Taking account of the property of the HSV color-space , a circular histogram HSV color segmentation was proposed [19]. Then HSV color-space can be obtained from the RGB color-space, performing as follows.

$$
\begin{aligned}
H &= \tan^{-1}\left(\sqrt{3(G-B)}, (2R-G-B)\right) \\
S &= 1 - \min(R,G,B)/I \\
V &= \max(R,G,B)
\end{aligned}
$$

2.4 Chinese Restaurant Process

Chinese Restaurant Process [11, 12] refers to an analogy with a real Chinese restaurant where the number of tables is infinite. The first customer sits down at a table. The ith customer sits down at a table with a probability that is proportional to the number of people already sitting at that table or if a new table is opened up with a probability proportional to the hyperparameter α. Because of exchangeability, the order in which customers sit down is irrelevant and we can draw each customers table assignment z_i by pretending they are the last person to sit down. Let K be the number of tables and let n_k be the number of people sitting at each table. For the ith customer, then is defined a multinomial distribution over table assignments conditioned on z_{-i}, i.e. all other table assignments except the ith:

$$
p(z_i = k \mid z_{-i}, \alpha) \propto \begin{cases} n_k & \text{if } k \leq K \\ \alpha & \text{if } k = K+1 \end{cases} \tag{3}
$$

Given the cluster assignment each data point is conditionally independent of the other ones. The exchangeability assumption in this process holds for some datasets but not in others. While several special models for spatial and temporal

dependencies have been proposed, the distance-dependent CRP offers an elegant general method to modeling additional features and non-exchangeability. For example if 10 customers are clustered using the CRP method, the first customer choosed the firts table with $p = \frac{\alpha}{\alpha} = 1$. The second customer chooses the first table with probability $\frac{1}{1+\alpha}$ and the second table with probability $\frac{\alpha}{1+\alpha}$. After the second customer chooses the second table, the third customer chooses the first table with probability $\frac{1}{2+\alpha}$, the second table with probability $\frac{1}{2+\alpha}$, and the third table with probability of $\frac{\alpha}{2+\alpha}$. This process continues until all customers have seats, defining a distribution over allocations of people to table or object to classes.

This method employed as a clustering algorithm shows an advantage compared with K-means clustering algorithm. In the K-means algorithm, the number of clusters must be determined in advance, while in the CRP algorithm, an adequate number of clusters can be determined though the process.

3 Proposed Algorithm

In the proposed algorithm, to cluster successfully the retrieved images from Internet, a PHOG, K-means clustering, object segmentation and CRP algorithms are employed in the cascade structure with 4-stages. This proposed 4-stages approach is shown by figure 1. Firstly using keyword introduced by user, the Google image retrieval engine "Google Image Search^{TM}" extracts the corresponding images from Internet to generate a database, which becomes the input of the proposed 4-stages clustering algorithm. The dataset consist of 64 images for each keyword. The use of Internet is convenient, because it allows uploading the dataset for any object at any time. In this section process of each stage is described.

3.1 Image Clustering using PHOG and K-means

The data set generated using any image search engine contains many junk images which is not related to the introduced keyword. Firstly in the proposed algorithm junk images are discarded using PHOG and K-means clustering algorithm. The PHOG is applied to all images retrieved by a specific keyword to get M PHOG vectors, where M is number of retrieved images. Figure 2 shows some retrieved image together with the PHOG vector (level 0) represented by histogram form. From the figure, the images which share similar object shape yield quite similar PHOG vector. For instance the PHOG vector shown in figure 2b is very similar to figure 2b. On the other hand, the representation of figures 2f and 2h are significantly different, indicating that two images 2e and 2g are different. The pyramid levels of the PHOG used for this task is one, that is level-0 PHOG vectors is used. Thus, in order to get the most common group of the images retrieved by the same keyword, K-means clustering algorithm is used.

K-means algorithm requires the number of clusters as one of input data. After previous test, the clustering performance cannot be improved using more

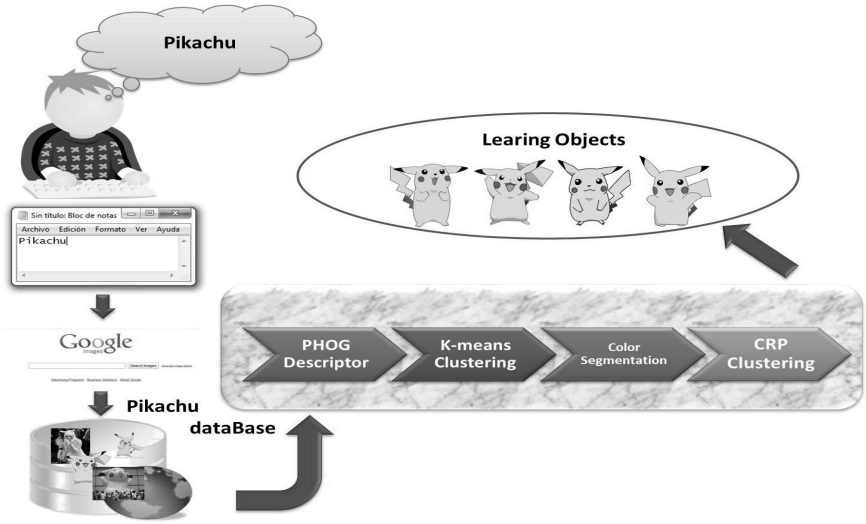


Fig. 1: Outline of the proposed learning object method.

than three clusters, so adequate number of cluster is considered as three ($k = 3$). Actually using three clusters, 64 retrieved images using same query (keyword) can be clustered correctly according with image appearance.

When K-means algorithm is converged, we analyze the number of elements of three clusters. The cluster with largest number of elements (images) is considered as winner and is kept for further process, while other two clusters are discarded, because images in these clusters are junk images whose relation with the query is very low.

3.2 Object Segmentation

In this section, all images in the winner cluster obtained in the previous stages are analyzed. When an image which contains a specific object is classified using this object, the background of the image can be interfered with the classification process causing an error. As mentioned in section 2.3, to segment the object from the background, we use color property of the surface of the object in HSV color-space.

The color reference is randomly selected from the center region of the image, which is defined as Region of Interest (ROI). The segmentation is done using a HSV color filter inspired on the circular filter introduced in [19]. If extracted area (object region) is less than the pre-established threshold, then the image is considered as a junk image, in other words, the image does not contain the object indicated by keyword, otherwise the image is considered as useful image and it is analyzed furthermore. The junk images are discarded in this stage.

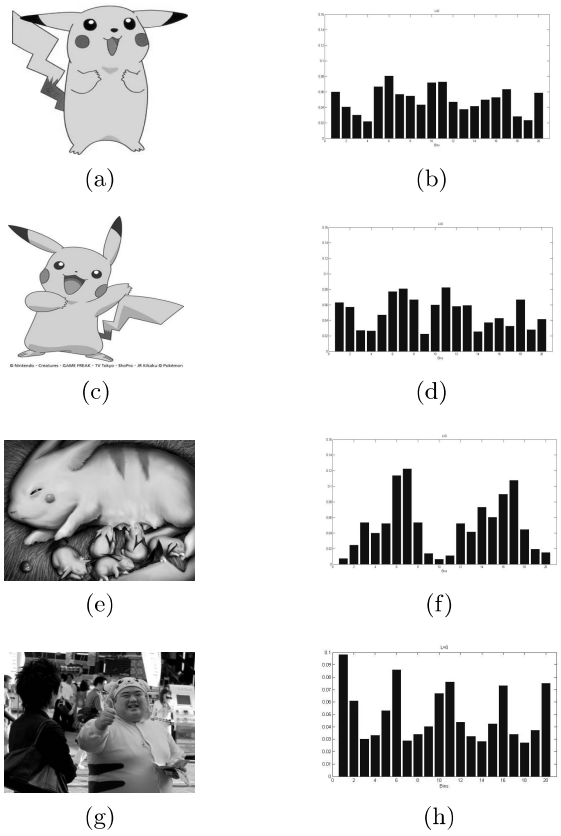


Fig. 2: Example of retrieved images and their PHOG representation (L=0).

The threshold value is established by a heuristic way, retrieving more than 1000 images using 50 keywords indicating cartoon's characters, animals and several objects. Fig. 3 shows an example of useful images and discarded images through the segmentation process, when a word "Pikachu" is given as keyword.

3.3 Chinese Restaurant Process (CRP)

Once the most of junk images were discarded, the most representative images (cluster of images) with the desired object indicated by keyword, can be selected using the CRP. CRP is implemented by a model-based Bayesian clustering algorithm with two input parameters n and α (see 2.4). The first parameter n is the maximum number of elements of each cluster and another parameter α determines the probability that a new table is opened up.

According with the total number of the retrieved images and the number of the discarded images during the previous stages, we determine these two

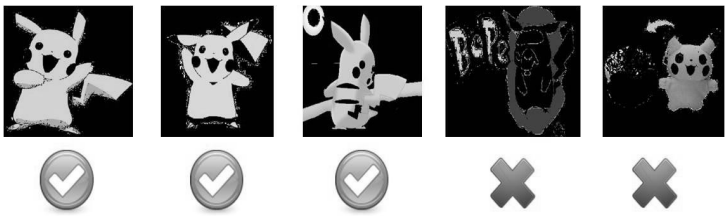


Fig. 3: Segmentation results.

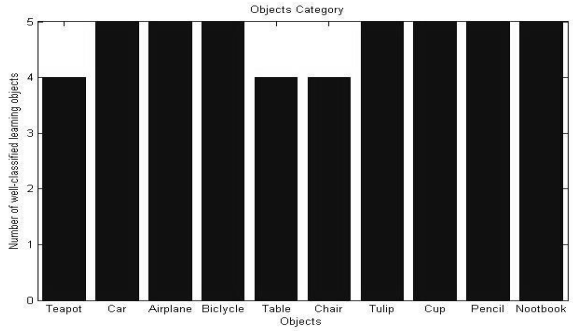
parameters as $n = 5$ and $\alpha = 0.2$. These values guarantee that the highest probability is assigned to the most popular cluster. The final cluster obtained by this process is called as "Learning Object Category"

4 Experimental Results

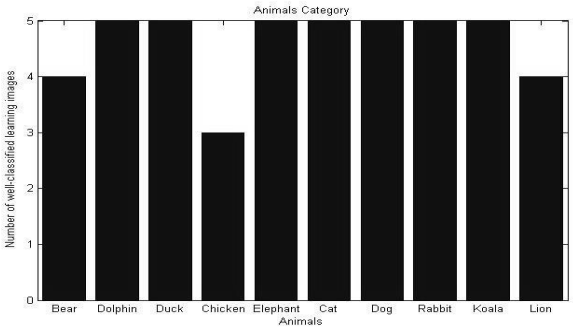
To evaluate the proposed algorithm, in this section some experimental results are shown. The evaluation criteria used here is "Ground truth", which indicate if the elements of final cluster obtained through the proposed algorithm are corresponded to the query (keyword) or not. The figure 4 shows the number of well-classified objects after all four stages on the proposed algorithm. Worth noting that for each keyword, 64 images are retrieved and finally only five images are allowed as the most popular cluster of the CRP algorithm.

The two graphs shown by Fig. 4 depict the level of well-classified objects according to the Ground Truth test. Although the values of the Ground Truth test of some objects such as teapot and chair are not sufficiently high compared with other objects, this situation can be improved if more specific keywords are used. This refers to a semantic problem. When the user query implies an ambiguous concept, for instance,"mouse", in the 64 retrieved images, 52 images contains computer mouse, 10 of them are an animal mouse and only 2 images are Mickey Mouse. In this case, the most of the retrieved images regards to a "computer mouse", therefore the learning object by the proposed algorithm is obviously a computer mouse. If a user desires to retrieve"animal mouse" in place of "computer mouse", he has to specify the query specifying his keyword as "animal mouse" or "mouse animal". Figure 5 shows the PHOG vectors obtained from given three images with the three different "mouse" object. From the figure, we can observed that three PHOG vectors are considerably different among them.

The learning objects obtained, are similar images to each other, i.e., this method can be used as similar-images retrieval system.



(a)



(b)

Fig. 4: Number of well-classified objects. a) Varied Objects. b) Animals.

5 Conclusions

In this paper, we proposed learning object category algorithm, which is composed by the following four stages: Pyramid of Histogram of Oriented Gradient (PHOG), K-means clustering algorithm, image segmentation on HSV color filter and Chinese Restaurant Process (CRP). From the experimental results, we conclude that the rate of the wrong classified objects in the final cluster (learning objects) is very low according to the figure 4. Due to the shape description capacity of PHOG, K-means algorithm and moreover efficient CRP classifier allow improving the learning methodology. Since CRP is a non parametric clustering method, the proposed algorithm can be combined with other techniques for some applications in which the learning of object category is important. Furthermore, since the proposed algorithm can use the huge image database of Internet, it can be an alternative method to learn any object at any time without a prior stored database. Even though the totally unsupervised learning object category task is still far, we consider that this work contributes to achieve this goal.

Acknowledgements. This work was supported by CONACYT and IPN.

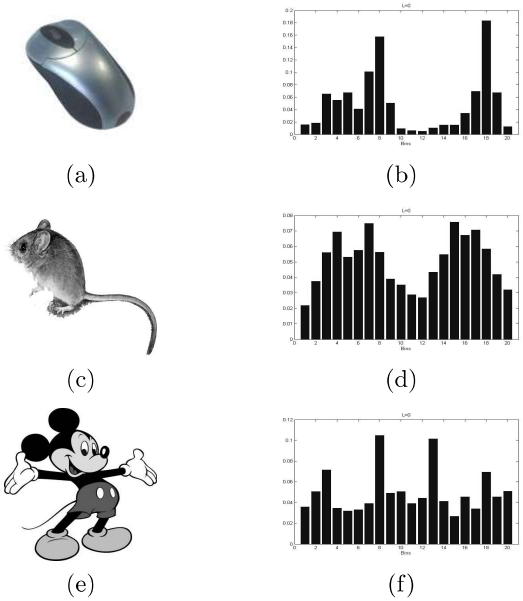


Fig. 5: Example of retrieved images by the "mouse" querry and their PHOG representation.

References

1. S. Agarwal, A. Awan, and D. Roth: Learning to detect objects in images via a sparse, part-based representation. IEEE PAMI, Vol. 20, Number 11, pp. 1475–1490 (2004)
2. K. Barnard, P. Duygulu, N. de Freitas, D. Forsyth, D. Blei, and M. Jordan: Matching words and pictures. JMLR, Vol. 3, pp. 1107–1135 (2003)
3. A. Berg, T. Berg, and J. Malik: An improved cluster labeling method for support vector clustering. Pattern Analysis and Machine Intelligence, IEEE Transactions, Vol. 27, pp. 461–464 (2005)
4. R. Fergus, L. Fei-Fei, P. Perona, and A. Zisserman: Learning Object Categories from Google's Image Search. In Proc. of ICCV, Vol. 2, pp. 1816–1823 (2005)
5. F.Schroff, A.Criminisi, A.Zisserman: Harvesting Image Database from the Web. in Proc. of International Conference on Computer Vision, pp. 1–8 (2007)
6. B. C. Russell, A. Torralba, K. P. Murphy, and W. T. Freeman: Labelme: a Database and Web Based Tool for Image Annotation. IJCV, Vol. 77, Number 1, pp.1453–1466 (2010)
7. R. Fergus, L. Fei-Fei, P. Perona, and A. Zisserman: Learning Object Categories from Internet Image Search. JPROC. of IEEE, Vol. 98, pp. 1453–1466 (2010)
8. J. Liu, R. Hu, M. Wang, Y. Wang and E. Chang: Web-Scale Image Annotation. Proceedings of the 9th Pacific Rim Conference on Multimedia, pp. 663–674 (2008)
9. F.Schroff, A.Criminisi, A.Zisserman: Harvesting Image Database from the Web. In Proc. of International Conference on Computer Vision, pp. 1–8 (2007)

10. D. M. Blei and P. I. Frazier: Distance dependent Chinese restaurant processes. In ICML 2010 (2010)
11. D. M. Blei, T. L. Griffiths, M. I. Jordan and J.B. Tenenbaum: Hierarchical Topic Models and the Nested Chinese Restaurant Process. In Neural Information Processing Systems(NIPS) (2003)
12. R. Socher, A. Maas and Christopher D. Manning: Spectral Chinese Restaurant Processes: Nonparametric Clustering Based on Similarities. In Fourteenth International Conference on Artificial Intelligence and Statistics (AISTATS) (2011)
13. K. Jain: Data clustering: 50 years beyond K-means. Pattern Recognition Letters. Elseiver Journal. Vol.31, pp. 651–666 (2010)
14. A. Bosch, A. Zisserman, and X. Munoz: Representing shape with a spatial pyramid kernel. In Proceedings of the International Conference on Computer Vision, pp. 401–408 (2007)
15. A. Bosch, A. Zisserman, and X. Munoz: Image Classification using Random Forests and Ferns. In Proceedings of the International Conference on Image and Video Retrieval, pp. 1–8 (2007)
16. J. Meng, H. Shang and L. Bian: The Application on Intrusion Detection Based on K-means Cluster Algorithm. In Proceedings of the 2009 International Forum on Information Technology and Applications, Vol.1, pp. 150–152 (2007)
17. L. Lucchese and S.K. Mitra: Image Segmentation A State-Of-Art Survey for Prediction. In Advanced Computer Control, 2009. ICACC '09., pp. 420–424 (2009)
18. N.R. Pal and S.K. Pal: A Review on Image Segmentation Techniques. Pattern Recognition, Vol. 26, Number 9, pp. 1277–1294 (1993)
19. Din-Chang Tseng, Yao-Fu Li, and Cheng-Tan Tung: Circular histogram thresholding for color image segmentation. Pattern Recognition, Vol. 2, pp. 673–676, (1995) Current version: 2002