Available Online at www.ijcsmc.com

International Journal of Computer Science and Mobile Computing

A Monthly Journal of Computer Science and Information Technology

ISSN 2320–088X

IJCSMC, Vol. 2, Issue. 2, February 2013, pg.74 – 79

REVIEW ARTICLE

A REVIEW OF OBJECT RECOGNITION USING VISUAL CODEBOOK

MS. RUPALI D. RANE[1], *PROF. BHARATI K. KHADSE*[2], *PROF. S.R.SURALKAR*[3]

[1]**Department of E&TC, S.S.B.T's College of Engineering and Technology, (Bambhori) Jalgaon-425001 (M.S.)**

ruprane296@gmail.com

[2]**Department of E&TC, S.S.B.T's College of Engineering and Technology, (Bambhori) Jalgaon-425001 (M.S.)**

bharatikhodpe@yahoo.co.in

[3]**Department of E&TC, S.S.B.T's College of Engineering and Technology, (Bambhori) Jalgaon-425001 (M.S.)**

Shekhar_srs@rediffmail.com

Abstract*— Automated recognition of object categories in images is a critical step for many real-world computer vision applications. Interest region detectors and region descriptors have been widely employed to tackle the variability of objects in pose, scale, lighting, texture, color, and so on. This review paper studies codebook models which is used for various computer vision tasks and the different features of object recognition system. The codebook model-based approach provides state-of-the-art performances on current datasets. This approach is impressive because we are simply modeling the statistical distributions of low-level image features into a fixed-length vector in histogram space to which standard classifiers can be directly applied. The discriminative power of such a visual codebook determines the quality of the codebook model and its size controls the complexity of the model.

Indexed Terms: - codebook models, Segmentation & Classification, Semantic Image Annotation.

I. INTRODUCTION

Object recognition deals with identifying the object from an image. Object recognition is the field in computer vision which aims at recognizing objects from image data and to estimate the position and orientation of the recognized object. Object recognition aims at representing the objects in the surrounding and creating their relations with the 3D structures of the object. Shapes can differ in appearance especially in the way of perception. Objects can be viewed from different angles and positions which make an object look different from all the angles.

Two trees are also different but we label them as one object. Many theories have been developed on how humans perform object recognition. Humans have the capability to easily identify objects in our surroundings, regardless of their circumstances, whether they are upside down, of different color or occluded. Object recognition is one of the most fascinating abilities that humans easily possess since childhood. By just looking at the object, humans are able to tell its identity or category despite of the appearance variation due to change in pose, illumination, texture, deformation, and under occlusion [4] [11]. Object recognition utilizes certain other features also for the better recognition like orientation, distance, size of object etc. Humans can uniquely identify individual objects not just face but also other objects like cars, houses, animals etc. Just the presence of a hand can indicate that there's a person present or a set of wheels can give reason to believe a car might be in the scene.

But shape is the main feature of identity. For identification, discrimination between physically similar objects is made but for the characterization of objects their generalization is necessary [8].To recognize, identify or do categorization of an object involves comparing its visual representation with some stored information[12]. Object recognition sometimes becomes difficult as objects change themselves with respect to the scene for e.g. in case of varying lightning conditions or the viewpoint of the object, presence of background clutter etc.

Different steps are involved in the object recognition with certain assumptions in each case. In identification stage, object shape is considered fixed. In the generalization task recognition of an object is done despite the changes in its appearances due to the transformations. Categorization involves the assignment of the object to a similar class of objects.

Some issues that are needed to be considered in the object recognition system are:

- Object representation: How an object is represented is a matter of concern as it determines the perception by a person to that particular object. The primary attribute of any object is the shape of the object which is a matter of concern whereas the secondary attributes like color, size, texture are also needed but they only aid in getting to know more information about any object. The representation of any object should be such that it gives all the information about it at a single instance.
- Feature extraction: This step deals with the features that are needed to be extracted.
- Feature model matching: This step deals with how the features that were extracted are matched with the model already in the database. This matching helps in the recognition process.
- Hypothesis formation: This deals with how the probabilities are assigned to the objects. The presence of an object is verified by using their models. This step aims at reducing the size of search space.
- Object verification: The presence of objects is verified from their models.

Figure 1: Image showing clean and dirty room

Figure 1 shows two images. In the first image a clean room is shown from which the things needed are easily visible. Whereas in the second image there is large amount of clutter because of that recognition of object is a difficult task.

II. CODEBOOK MODELS

A codebook model yields a distribution over code word that models the whole image, making this model well-suited for describing context. Unlike text, visual words are not intrinsic entities and different quantization methods can lead to very different performances. The size of the codebooks that have been used in the literature range from 10^2 to 10^4, resulting in very high-dimensional histograms. A larger size of codebook increases the computational needs in terms of memory usage, storage, the computational time to construct the codebook and to train a classifier. On the other hand, a smaller size of codebook lacks good representation of true distribution of features. Thus, the choice of the size of a codebook should be balanced between the recognition rate and computational needs. The compactness constraint is typically ignored by several systems that mainly focus on categorization performance.

Now, we provide a review of a selective research work from patch-based visual object recognition literature. In general, there are two types of codebook that are widely used in the literature: global and category specific (or concept-specific) codebook. A global codebook is category-independent and has less discriminative power, whereas a category specific codebook may be too sensitive to noise. The conventional approach to constructing either a global or category-specific codebook is achieved by cluster analysis, usually by the *K*-means method. The learnt cluster centers are not semantically meaningful since the clustering is based on appearance similarity only.

A. Global Codebook

A globally-clustered codebook is usually constructed by clustering visual descriptors that are randomly chosen from each class of a training set. Thereafter, each image is represented as a feature vector by computing the frequency histograms with the learnt clusters. This mapping produces a bag-of-features representation. Several authors have used the globally-clustered codebook at some stage in their framework.

- Csurka et al. [1] used the Harris affine region detector to identify the interest points in the images which are then described by SIFT descriptors. A visual codebook was constructed by clustering the extracted features using *K*-means method. Images are then described by histograms over the learnt codebook. The authors run the *K*-means several times over a selected size of *K* and different sets of initial cluster centers. The reported results were the clusters that gave them the lowest empirical risk in classification. The size of the codebook used in reporting the results is 1000. The authors compared Naive Bayes and Support Vector Machine (SVM) classifiers in the learning task and found that the one-versus-all SVM with linear kernel gives a significantly (i.e. 13%) better performance. The proposed framework was mainly evaluated on their 'in-house' database that is currently known as 'Xerox7' image set containing 1,776 images in seven object categories. The overall error rate of the classification is 15% using SVMs.
- Jurie and Triggs [2] proposed a mean-shift based clustering approach to construct codebooks in an under sampling framework. The authors sub sample patches randomly from the feature set and allocate new cluster centroids for a fixed-radius hyper sphere by running a mean-shift estimator on the subset. The mean-shift procedure is achieved by successively computing the mean-shift vector of the sample key points and translating a Gaussian kernel on them. In the next stage, visual descriptors that fall within the cluster are filtered out. This process is continued by monitoring the in formativeness of the clusters or until a desired number of clusters is achieved.

The features used in their experiments are the gray level patches sampled densely from multi-scale pyramids with ten layers. Three different feature selection methods were used in the experiments: maximization of mutual information, odds of ratio, and training an initial linear SVM on the entire training set to select the features that have the highest weight. Two different ways of producing fixed-length feature vectors from the learnt codebook were used in the experiments: Binary indicator vectors which were produced by thresholding the frequency counts of the codeword in the image and the histograms.

- Nister and Stewenius [3] proposed a hierarchical *K*-means clustering that constructs a vocabulary tree in an offline training stage for image retrieval from a large database. Features were extracted using maximally stable extremal regions (MSERs) which are then described by SIFT descriptors. SIFT features were then quantized with the vocabulary tree. The vocabulary tree is constructed by a hierarchical scoring scheme based on the term frequency-inverse document frequency (*tf-idf*) score. First, an initial *K*-means process is run on the training data, defining *K* centroids. The training data is then partitioned into *K* groups, where each group consists of the features closest to a particular centroid. The second step is then recursively processed by quantizing each node into *K* new parts, where *K* defines the number of children of each node. The tree is constructed level-by level up to a maximum number of levels. Following the recursive process, in the online phase, each visual descriptor is propagated down the vocabulary tree by coding the closest node at each level. The proposed technique was tested on a ground truth database containing 6,376 images in groups of four of the same object but under different conditions. From their experimental results, they found that larger vocabulary (between 1 and 16 million leaf nodes) improves retrieval performance. They claim that this methodology provides the ability to make fast searches on extremely large databases (i.e. one million images).
- Mikolajczyk et al. [4] find local features by extracting edges with a multi-scale Canny edge detector with Laplacian-based automatic scale selection. For every feature, a geometry term gets determined, coding the distance and relative angle of the object centre to the interest point, according to the dominant gradient orientation and the scale of the interest point. These regions are then described with SIFT features that are reduced to 40 dimension via principal component analysis (PCA). The visual codebook is constructed by means of a hierarchical *K*-means clustering. Initially the features are clustered using *K*-means algorithm and then agglomerative clustering is performed to obtain compact feature clusters within each partition. Given a test image, the features were extracted and a tree structure is built using the hierarchical *K*-means clustering method in order to compare with the learnt model tree. Classification is done in a Bayesian manner computing the likelihood ratio. This test is done at

local maxima of the likelihood function of the object being present. Some additional tests are applied to determine whether objects of different classes share similar clusters or whether overlapping objects exist. In this manner, the location, scale and orientation of multiple objects can be determined. Experiments were performed on a five class problem taken from the PASCAL VOC 2005 image dataset containing four classes and a RPG (rocket-propelled grenade) shooter that was collected from various sources.

- Wu and Rehg [5] showed that when the histogram intersection kernel (HIK) are used in clustering patch-based visual descriptors that are histograms, the codebooks constructed produce improved bag of features classifiers. The proposed method replaces *K*-means clustering that uses the *L*2 distance measure with HIK for better performance when the choice of feature representation is histograms. When comparing *K*-means with *K*-median, the latter uses the *L*1 distance measure. In the first step, features are extracted to construct a visual codebook of size 200. At the next step, an image or image sub-window is represented by a histogram of code words in a specified image region. An image is represented by the concatenation of histograms from all 31 sub-windows that split an image into three levels, resulting in a histogram of dimension 6200. Spatial and edge information are incorporated as an additional input, and histograms are concatenated from the original input and Sobel gradient image. The authors also propose a one class SVM formulation using HIK that can be used to improve the effectiveness of the HIK-based codebook, by compact clusters in histogram feature space.

B. Category Specific Codebook

A category-specific or concept-specific codebook is usually constructed by clustering the extracted features from images in a single class only. Sometimes, the features can also be extracted with a concept that covers different and independent regions of the same category or scene. This makes the resulting clusters depend on only that subset of the feature space which is relevant for the concept. The construction process of a codebook is identical to the globally-clustered codebook, and is carried out separately for each of the categories or concepts

- Sivic and Zisserman [6] proposed an approach to retrieve visual objects and scenes from a movie using a text retrieval approach. Local regions were extracted from each frame in the video in the following two different ways: One method is referred to as a shape-adapted (SA) region which surrounds an interest point by an elliptical shape. The second method is referred to as a maximally stable (MS) region which is constructed by intensity watershed image segmentation. The SA regions are detected on corner like regions and the MS regions correspond to blobs of high contrast with respect to the surroundings. Both SA and MS regions are then described by SIFT descriptors. The authors were aware of the difficulty in clustering a very large scale of descriptors extracted from their movies, so instead they selected 10,000 frames which represent about 10% of all the frames in the movie, resulting in 200,000 averaged track descriptors to construct a codebook. A visual codebook is constructed using *K*-means clustering algorithm and Mahalanobis distance measure
- Leibe and Schiele [7] used the Harris interest point detector to extract image patches. The pixel gray values of those patches are then clustered using the agglomerative clustering method to generate a visual codebook. The size of the learnt codebook was further reduced by merging the most similar clusters in a pair-wise manner when the similarity between clusters exceeds a predefined threshold *t*. Instead of assigning image patches to their nearest codeword in the learnt codebook, every patch casts probabilistic votes to the codebook using the NGC measure whose similarity is above *t*. For classification, a generalized Hough transform-like voting scheme is applied. The proposed method was evaluated on a database of 137 images of scenes containing one car each in varying poses. The size of the codebook was around 2,500.
- Farquhar et al. [8] proposed alternatives to the scheme introduced by Csurka et al. [1]. The Gaussian mixture model (GMM) was proposed as a replacement of the *K*-means based codebook construction, and summed responsibility replacing bin membership for histogram generation. The GMMs were all trained for category-specific codebooks and were then combined into a single codebook. Features were extracted using multi-scale Harris affine region detector that are then described by SIFT descriptors. The features were pre-processed to reduce its dimensionality. The authors used two different methods to reduce dimensions: the PCA and partial least squares (PLS), and found that PLS improves classification performance over the PCA method for the same number of reduced dimensions.
- Zhang et al. [9] compare sets of local features in two different methods. Their first method involved clustering a set of patch-based descriptors in each image to form a representation of *(ci,wi)* pairs, that they refer to as image signature where *ci* is the cluster centre and *wi* is the proportional size of the *i* th

cluster. Cluster centers were obtained using *K*-means algorithm with $K = 40$. Earth Mover's Distance (EMD) was the choice for measuring similarities between image representations.

III. FUTURE DIRECTIONS FOR OBJECT RECOGNITION SYSTEMS

Currently, most object recognition systems use either purely visual features or textual metadata associated with images. They have advantages and disadvantages respectively. To overcome their drawbacks and improve the performance without sacrificing the efficiency, the new web object recognition systems should pay a great attention for these features:

a. Automatic Image Segmentation and Classification

As the image passes through the segmentation and classification process the system automatically identifies regions, scenes, objects, facial aspects and spatial positions of those regions, objects and faces within the image. As part of this process the attributes within the image are given statistical relevancy based on how they typify the concept. Automatic classification of the content of an image lends itself to many applications, combining this with existing metadata allows users to search more accurately, for many more things in an image, in addition to making images with poor or non-existent keywords visible for the first time at a dramatically reduced cost compared with manually adding keywords.

b. Semantic Image Annotation

The objective of semantic annotation is to describe the semantic content in images and retrieval queries. Semantic annotation requires some understanding of the semantic meaning in images and retrieval query, and standardization of representation of images. Based on the semantic annotation of images and retrieval queries, we can compare semantic similarity between images and a retrieval query. At present, semantic annotation is implemented by some markup language such as XML based on a shared ontology definition.

c. Semantic Object recognition

This feature of the system is the retrieval architecture, which understands the syntax and meaning of a user's query and uses a linguistic ontology to translate this into a query against the visual ontology index and any metadata or keywords associated with the image. The retrieval system takes textual queries and reasons about them through understanding their syntax and meaning. For example, in a traditional system if a user queries "beach without people" the text system looks for the words "beach" and "people" and does not understand the meaning of "without".

d. Ontology Reasoning

Ontological reasoning is the cornerstone of the semantic web, a vision of a future where machines are able to reason about various aspects of available information to produce more comprehensive and semantically relevant results to search queries. Rather than simply matching keywords, the web of the future will make use of ontology to understand the relationship between disparate pieces of information in order to more accurately analyze and retrieve information.

e. Multi-Object Search

Object-based image retrieval has recently become an important research issue in retrieving images on the basis of the semantics of images. However, most existing object-based image retrieval systems are based on single object matching, with its main limitation being that one individual image region (object) can hardly represent the users' retrieval target especially when more than one object of interest is involved in the user query. An important aspect of the system is that users are allowed to formulate a query based on multi objects of an image.

f. Spatial Search

As part of the classification process, the spatial context of identified regions, objects, scenes and faces is encoded within the index. This means the system can return semantically accurate results for queries involving spatial prepositions such as "with", "next to", "on", "beside" "against" etc. In addition to querying properties which are in the "top" "bottom" "center" "left" or "right" of an image. In other word, enable the user to search for multiple Object criteria as the same as text based information Retrieval.

g. Queries with Different Forms

The system should enable the user to search for images using text or using image example or using combination of text and image query.

h. Text Query with Multilanguage

Current search engines face the problem of- Limited Resource Languages. The lower the web presence of a language, the fewer hits a speaker of that language gets from a query. A query for grenivka (Slovenian for grapefruit) produces only 24 results, of which only 9 are images of grapefruits. Yet, translating the query into English produces tens of thousands of images with high precision.

i. Semantic Recommendation

Most object recognition method always assumes that users have the exact searching goal in their mind. However, in the real world application, the case is that users do not clearly know what they want. Most of the times, they only hold a general interest to explore some related images. As a result, building a recommendation system based on the user query is necessary.

The system should be able to represent common search terms used in object recognition. This is used by a keyword-generation tool to expand a user's search keyword. This is achieved by finding which concepts in the ontology relate to a keyword and retrieving information about each of these concepts.

IV. CONCLUSION

A wide variety of researches have been made on object recognition. This paper attempts to deal with a detailed review of the most common traditional and modern object recognition systems from early text based systems to content based retrieval and ontology based schemes. This paper review those works mainly based on the methods/approaches they used to come up to an efficient retrieval system together with the limitations/challenges and tried to give a constructive idea for future work in this field.

REFERENCES

[1] Csurka, G., Dance, C. R., Fan, L., Willamowski, J., & Bray, C. (2004). "Visual categorization with bags of key points". In *Workshop on statistical learning in computer vision, ECCV'04* (pp. 1–22).

[2] Jurie, F., & Triggs, B. (2005). "Creating efficient codebooks for visual recognition". In *Proceedings of the tenth IEEE international conference on computer vision (ICCV'05)* (Vol. 1, pp. 604–610).

[3] Nister, D., & Stewenius, H. (2006). "Scalable recognition with a vocabulary tree". *IEEE Computer Society Conference on Computer, 2*, 2161–2168.

[4] Mikolajczyk, K., Leibe, B., & Schiele, B. (2006). "Multiple object class detection with a generative model". In *Proceedings of the IEEE conference on computer vision and pattern recognition (CVPR'06)* (Vol. 1, pp. 26–36).

[5] Wu, J., & Rehg, J. (2009). "Beyond the Euclidean distance: Creating effective visual codebooks using the histogram intersection kernel". In *Proceedings of the IEEE international conference on computer vision (ICCV'09)*.

[6] Sivic, J., & Zisserman, A. (2003). "Video google: A text retrieval approach to object matching in videos". In *Proceedings of the ninth IEEE international conference on computer vision (ICCV'03)* (pp. 1470–1478).

[7] Leibe, B., & Schiele, B. (2003). "Interleaved object categorization and segmentation". In *Proceedings of the British machine vision conference (BMVC'03)* (pp. 759–768).

[8] Farquhar, J. D. R., Szedmak, S., Meng, H., & Shawe-Taylor, J. (2005). "Improving "bag-of-key points image categorisation: Generative models and PDF-kernels". In *LAVA report*. U.K.: University of Southampton.

[9] Zhang, J., Marszalek, M., Lazebnik, S., & Schmid, C. (2007). "Local features and kernels for classification of texture and object categories: A comprehensive study". *International Journal of Computer Vision, 73*, 213–238.

[10] Wong, R.C.F. and Leung ,C.H.C. 2008. "Automatic Semantic Annotation of Real-World Web Images". *IEEE transactions on pattern analysis and machine intelligence,* vol. 30, no. 11, pp 1933- 1944.

[11] Toshev A, Taskar B, Daniilidis K, "Object detection via boundary structure segmentation" *IEEE Computer Society Conference on Computer Vision and Pattern Recognition 2010.*

[12] Gu C, Lim J, Arbelaez P, Malik J, "Recognition using regions". *IEEE Computer Society Conference on Computer Vision and Pattern Recognition*,2009.

[13] Perronnin, F. (2008). "Universal and adapted vocabularies for generic visual categorization". *IEEE Transactions on Pattern Analysis and Machine Intelligence, 30*, 1243– 1256.

[14] Ramanan, A., & Niranjan, M. (2010). "A one-pass resourceallocating codebook for patch-based visual object recognition". *In IEEE international workshop on machine learning for signal processing (MLSP'10)* (pp . 35–40).

[15] Chunjie Zhang; Jing Liu; Qi Tian; Yanjun Han; Hanqing Lu; Songde Ma , "A Boosting, Sparsity-Constrained Bilinear Model for Object Recognition".*IEEE Digital Object Identifier*: 10.1109/MMUL.2011.20 Publication Year: 2012 , Page(s): 58 - 68

