International Journal of Applied Engineering Research
ISSN 0973-4562 Volume 10, Number 11 (2015) pp. 29845-29856

http://www.ripublication.com

A Survey on Context Based Image Annotation Techniques

Swati Nair L[1], Ms R Manjusha[2], Dr.LathaParameswaran[3]
M.Tech scholar[1], Assistant Professor[2], Professor[3]
Department of Computer Science Engineering
Amrita VishwaVidyapeetham, Coimbatore

Abstract

The importance of image acquisition and then analysing them for various purposes is increasing everyday .Image annotation and retrieval is a vital process for analysis of large data. Context based annotation systems labels the images based on the context of the scene and provides accurate results for automatic annotation compared to the earlier Content based systems and thus has become a very important research domain in image processing. Many approaches and representations are proposed and developed for context based image annotation .This paper provides anoverview of some of the important approaches and representations of objects and their relationship used widely for context based image annotation.

Keywords: Context Based Image annotation , Automatic Image Annotation, Content Based Image Retrieval, Semantic Gap, Hierarchical generative model

Introduction

Due to the rapid growth of archiving of images, the need for indexing and searching images effectively has increased significantly today. In spite of the fact that many Content Based Image Retrieval methods are prevalent, searching based on image feature is rather difficult for the users. Most of the users prefer searching with textual queries. This can be achieved by annotating the images manually and then searching the annotated images using textual queries. But it is a known fact that manual annotation of a large number of images is very much time consuming, expensive and involves considerable efforts. Hence, automatic image annotation methods are preferred over manual annotation for efficient retrieval of images. Thus, automatic image annotation with keywords is widely used which involves the learning of semantics of images. Hence the context based image annotation plays a very important role. The automatic image annotation cannot be accurate if the context of the scene is not taken into account for any object in the scene.

The earlier image retrieval methods were content based. The CBIRsproposed in [1] extracts the low level features like colour, texture, shape which is used for image annotation. But the main drawback is that it does not take into account the semantics which results in the wrong annotation of objects which are entirely different from each other but have exactly the same color or texture features (e.g. cannot distinguish between a cheetah and a tiger due to the same color and texture features. Context based image annotation on the contrary deals with recognizing and categorization of the objects taking into account the context of the scene. This method is different from content based image annotation as context deals with the objects and its neighbourhood (surroundings) whereas content deals only with the objects. It deals with semantic relationship between the object and its surroundings. The semantic relations are obtained from the size, probability and position which defines the interactions between the various objects in a scene. These semantic relations are denoted as contextual features.

The application of Context based image annotation is in browsing with image queries, unmanned navigation of robots, helping children with autism, assisting blind people for moving without human supported. The system proposed in [2] describes a system for navigation of robots for blind people taking into account the context of the scene and thus annotating different objects in a particular scene .Amethod which takes the context of the scene into consideration and can easily differentiate between the objects of the same shape, colour etc. has been proposed in [12]. For example one can differentiate between a lemon and a tennis ball which is same in color and shape but contextually different objects.

The different types of contextual features are scale context, semantic context and spatial context. Semantic context is defined in terms of co-occurrence of an object with other objects and its occurrences in scenes. Spatial context is the likelihood of the presence of a particular object in any place or position and its absence in other places while comparing it to any other object in that particular scene. Scale context is defined based on the comparison between the sizes of one object with another. Context based systems take into account the surroundings of the object also into account whereas content based systems takes into account the object alone. Hence there is always a semantic gap between the low level features extracted from images and the high level information needed for the user. The main objective of a context based annotation system is to bridge this gap and to provide a perfect annotation system.

Context Based Image Annotation Techniques

There are many techniques used for Context based image annotation for understanding semantics of the scene and to understand the relationship between the object and the scene.This work presents an overview of the various techniques related to the context based image annotation systems currently available.

A. Graph based method

The graph based method takes into consideration the objects and their relationship. The objects are shown by nodes and edges. The nodes denote objects whereas the weight of the edges denotes the probability of co-occurrence between the object pairs (fig 1).

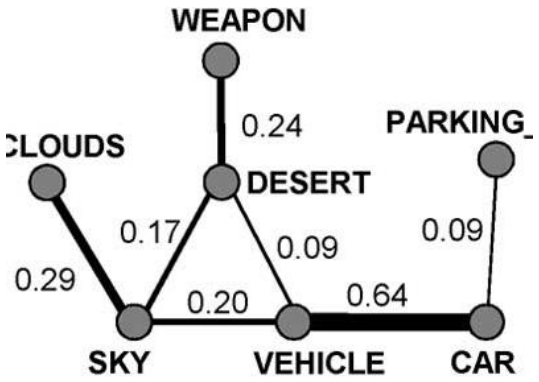

Figure 1: Graph Based Model

The above figure shows a weighted graph .The objects are sky, vehicles, desert, car, parking, weapon and clouds.

The interaction between known categories and unknown regions are modelled in [3] and a method to discover new categories among object categories which are not labelled is designed. In the proposed method two variations of graph based object descriptors are introduced. They are used to capture the two dimensional and three dimensional co-occurrence patterns with respect to a particular region spatially. By using this model the interactions among known objects and unknown objects can be computed to find new visual object classes. Instead of detecting all categories right from the start, new objects are identified by extracting useful cues from already known classes .The texture, shape and color features are combined in this method by Multiple Kernel Learning (MKL) framework. The posterior probabilities of any region in an image can be determined by using classifiers like Support Vector Machine (SVM) classifier which is a very simple classification method for classifying objects into two categories.The method was evaluated on various datasets and showed very good results for the discovery of unknown objects in an image. Detection on unknown or known objects is a very difficult problem. In this approach a robust method is designed by investigating the confidence scores between the known and unknown objects and clustering the objects.

A method assigning a label to each pixel of a given image from a set of possible object classeshas been proposed in [4]. Basically conditional random fields are used to estimate the interactions and correlation between the pixels. The major cue that aids in recognizing objects is by getting the statistics of the object's co-occurrence globally i.e. by finding out the objects or classes that are likely to occur together in an image. The experiment was conducted on MSRC dataset and VOC dataset .Low level

features like color ,texture and Texton response is extracted. A controlled test for evaluating the performance of the CRF models, both without and with co-occurrence potentials is performed .The results showed that consideration of these potentials better labelling results are obtained .Many methods are proposed based on this strategy but have many limitations.They involve complex computations, are costly and their application on a large dataset is limited. But the method proposed has improved labelling results compared to using only pairwise models.

The work proposed in [5] a hierarchical generative model for any given image that can classify the entire scene, then recognizes and segments every object, and also annotates or labels the image with a list of keywords or tags. In this work a hierarchical model is developed to combine the patch-level, object-level, and scene-level informationImages are modelled as a visual and a textual model .Images and tags are used as data fromFlicker.com.SIFT features are extracted. This method is compared with Bag of Words model and corr-LDA model .It has been proved that this method outperforms the two methods.But in this method the geometry and appearance information of objects are not captured.

B. Tree Based methods.

Tree based method involves hierarchical representation of objects based on their dependencies. A tree based method which is also a type of graph based method where objects are represented as nodes and its relationship is represented as edges has been proposed in [6]. The tree based representation is shown in Fig 2.Here the nodes represent the objects and the thickness of the edges between the objects indicates the probability of co-occurrence between the objects. Car, bus, dog, bird, chair, table, sofa, etc. are the nodes in the tree and they denote the objects.

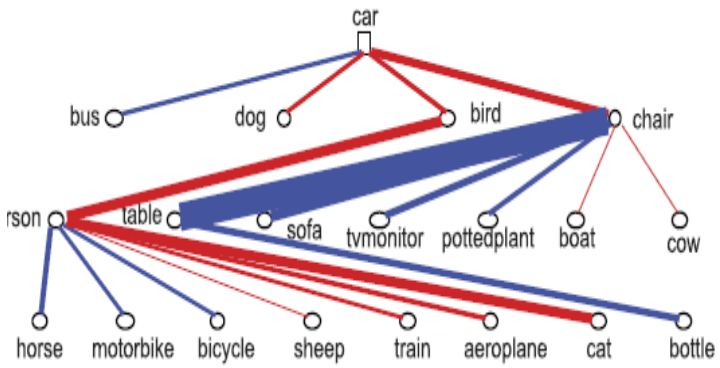

Figure 2: Tree Based Model

An efficient model to capture information between more than a hundred object categories contextually using a hierarchical tree based structure has been proposed in [7]. Hereanew data set is introduced and the dataset comprises images containing different instances of different object classes. In this model global image features and dependencies between object categories are incorporated and local detector outputs

are incorporated into a single framework based on probability. The object recognition performance is improved in this contextual model and it enables querying of images by multiple object classes, thus providing efficient scene interpretation. This approach can be used for understanding of scenes which cannot be learnt using local detectors, for example detecting objects which are out of context or finding the most likely scene and unlikely object in a dataset. Markov Random fields and texture features are used and can capture the dependencies of over hundred object categories. However this method does not capture spatial relationships.

C. Model based on co-occurrence, location and appearance (probability)

In this method the relationship between the objects is defined in terms of contextual features like co-occurrence, location and appearance (fig. 3). The contextual features include:

1. **Semantic context –** It is based on probability and defined by the co-occurrence of an object with other objects and its occurrences in scenes.
2. **Spatial context –** It is based on position and defined in terms of likelihood of the presence of a particular object in any place or position and its absence in other places while comparing it to any other object in that particular scene.
3. **Scale context –** It is based on the size or appearance and defined based on the comparison between the size of one object with another.

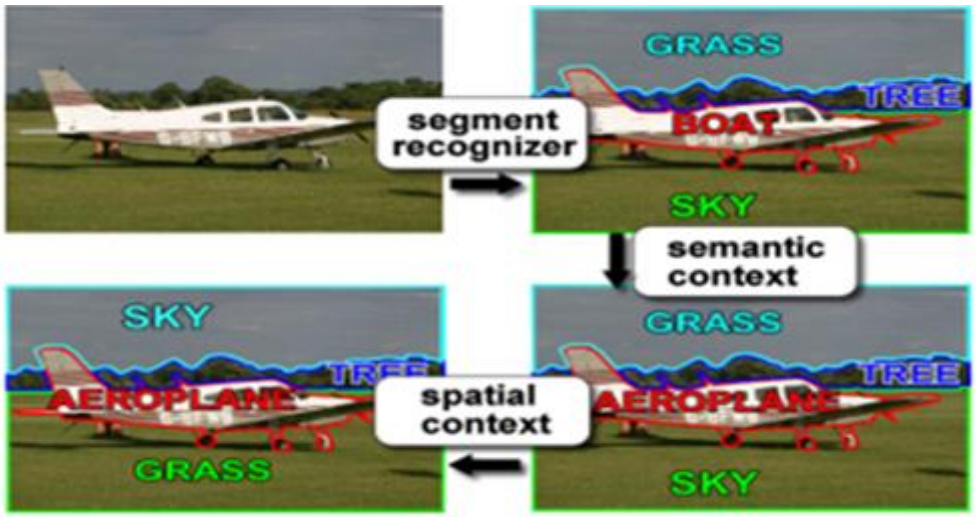

Figure 3: System incorporating Spatial and Semantic context (from [8])

In [8] an approach for classification of objects that incorporates two probabilities viz., contextual co-occurrence and its relative position or location based on local features and appearance is explained .The CoLA (Co-occurrence, Location and Appearance) maximizes the agreement of object label in accordance with both spatial and semantic relevance using a conditional random field (CRF) .The relative location between objects or categories is modelled using pairwise features which are simple. By vector quantizing this feature space a small set of prototypical spatial relationships

directly from the data is learned. The results from evaluation conducted on two datasets i.e. PASCAL and MSRC showed that the combination of both co-occurrence and spatial context improves the system's accuracy in most of the cases when compared to using only co-occurrence based techniques. But the limitation is that it cannot be extended for a large number of objects and dependencies.

A method that deals explicitly with multiple categories of objects which co-exist in a particular image has been developed for object recognition [9]. The main aim of this method is object recognition by taking into account the contextual information of the scene which is defined by the relationship of co-occurrence among different object categories or class.The mixture ratios for different object classes or categories present in an image is estimated using maximum a priori (MAP) regression. In this method, the prior probability is estimated from the co-occurrence relation among the objects and then the likelihood of a particular event is estimated by defining combination model of distribution of frequencies for the local features and the model is linear. The features extracted are DOG and SIFT. Various experiments conducted on PASCAL datasets showed that this method is very effective. Poor AUC performance has been reported when background categories are not incorporated. The performance could be improved by taking into account the co- occurrence relationship among various object categories and background categories also.

D. Markov Random field (MRF)

The spatial constraints for an object can be modeled using MRFs taking into account the smoothness in a particular area in an image and the texture pattern variations in a small image area.A generative model which is hierarchical is proposed in [8] using Markov Random Fields and SIFT descriptors. The method segments and recognizes each object in a scene and then it classifies the entire scene. Each image is then annotated with a list of keywords or tags. All the three operations are done in a single framework in this model. For example, consider an image which is a scene of a game of polo which consists of several categories of objects such as grass, horses, human, etc. This can be annotated further with some abstract tags such as dusk, evening or less important tags such as saddle, stick tags. This hierarchically generative model explains images by incorporating both visual context and a textual context model in one framework. Visually important objects are denoted by small regions and corresponding patches, while visually uncorrelated annotation tags depends on the overall context of the scene .A completely automatic framework for learning the sematic is proposed in this model.It is able to learn from web data scene models that are noisy, for example images and tags from Flickr. This model significantly outperforms state of-the-art algorithms in context based annotation. But the geometry of the objects and information about the appearance of the objects is not considered. This has to be taken into account for improvement.

E. Bag of Words Model

The Bag of words model explained in [10] is a very efficient classifier. The classification is done by extracting the interest points. Features are extracted from the image and local patches are extracted from the image. The patches are represented as

numerical vectors called feature descriptors which should be invariant to scale and rotation. Hence SIFT (Scale-invariant feature transform) are best suited. After this the vectors are clustered by any of the clustering techniques like K- Means and a codebook is generated by assigning certain code word to each patch.

F.Bayesian Belief Network (BBN)

A BBN network as explained in [14] is based on the probability and is denoted by a graph, consisting of a set of edges and vertices. The variables are denoted by vertices or nodes and the conditional probability is denoted by the edges or arcs in the model. When there is no arc between two nodes it shows that the two corresponding variables are conditionally independent and there will be no situation wherein the state of one variable depends on the state of the other variable.

The two types of probability considered in BBN are Joint probability and Conditional probability. The joint probability of two events that are independent is defined by the product of the probabilities of the two events occurring independently. The Joint probability is estimated as $P(x,y) = P(x)P(y)$. Conditional probability$P(x|y)$for two dependent events x and y is the probability of occurrence of event y when x has already occurred. The joint probabilities for two dependent events is estimated as$P(x, y) = P(x)\ P(y|x)$. The other way of defining joint probability is $P(x,y) = P(x|y)\ P(y)$.

In [13], a system that can perform the analysis of any visual content based on any prior knowledge about the scene has been proposed. The domain knowledge is modelled using Ontologies and the application context is modelled using conditional probabilities and the prior knowledge about the scene.The statistical and explicit knowledge is integrated using a Bayesian network.The hypothesis is formulated using evidence-driven probabilistic inference. The incorporation of focus-of-attention (FoA) mechanism is also proposed in this method and this is based on the information obtained mutually between various categories. The most prominent hypotheses are selected to be evaluated by the Bayesian Belief Network. So there is no need to exhaustively test all the probable combinations present in the hypotheses set .The framework is evaluated after performing experiments using the contents from the three major domains and performed the following three operations: 1) image categorization or classification; 2) region labelling which is done locally and 3) annotation of video shot key frames. The results obtained showed the improved performance compared to a set of basic classifiers that do not incorporate any context, domain or scene knowledge. But the prior probabilities have to be known in advance. An enormously large number of training data is required for the approximation of prior and conditional probabilities.

G. Kernel Based Methods

These methods use kernel classifiers,where a kernel is a similarity measure. Assuming that

$K(y1, y2) > 0$ is the "similarity" of $y1, y2 \in Y$ .The kernel is computed as

$$k(y1, y2) = f(y1).f(y2) = \sum_{j=1}^{m} f_{j(y1)} f_{j(y2)}$$

A kernel classifier follows the Mercer's theorem which can be explained as follows. For every semi-definite kernel K which is continuous and symmetric positive there is a feature vector function f such that

$$k(y1, y2) = f(y1).f(y2)$$

where, function f may have infinitely many dimensions. Feature-based approaches and kernel-based approaches are often mathematically interchangeable. Feature and kernel representations are duals of each other. There is always a need for capturing patterns that are not linear in the data.For nonlinear classification the classes cannot be separated by a boundary.The linear models like linear regression, Support Vector Machine are not sufficient enough for classification. The main advantage of Kernels is that it transforms the linear models into nonlinear settings. This is achieved by mapping the lower dimensional data to higher dimensionsto obtain linear patterns, or by applying the linear model to all the space or by changing the mapping representation of features for a particular object.

In the context based image annotation approach explained in [12], a thirty dimensional feature vector is obtained by extracting the texture features(Gabor features in 3 scales and 4 orientations) and six color features (mean and standard deviation of the three color channels).Then the image is divided into blocks and then features are extracted . The image is then converted into a vocabulary of visual words and a kernel is designed for classification. This method was validated on datasets from the University of Washington and IAPR dataset. There are around 600 annotation keywords. These keywords are propagated to the test image by Contextual Keyword propagation method by determining the confidence score and thus annotating the image with the keywords with top five or seven confidence scores.Then an annotation refinement is done by contextual spectral embedding method to give a very refined annotation. The annotation results obtained were accurate when compared with the ground truth keywords. In this method the semantics of the image as well as the spatial, location and the co-occurrence information of the objects in the images are taken into consideration. The method can be utilised for large number of data.

Conclusion

A large amount of research has been done in the domain of context based image annotation and there are many methods of which a few are discussed in this paper. Each method uses different features, classifiers and approach and has its own strengths and weaknesses.In any conventionalCBIR system the main challenging aspect is the semantic gap. To overcome this gap context based annotation systems are more preferred to CBIR systems. In this paper, an attempt is made to deliver a comprehensive literature survey on context based image annotation techniques. As a literature survey paper, all aspects of individual work is not focussed .However the

main focus is given on the representation of objects and their relationship with each other , the features ,classifiers and the datasets the implementation is carried out .

Table 1: Summary of Document Retrieval Methods

Technique	**Strengths**	**Weaknesses**
Content Based Image Retrieval	Computationally easy	Does not take into account the semantics of the image which may result in wrong annotations
Tree based model	Can easily capture the dependencies of over 100 object categories	Construction of tree is difficult .
Graphical based method	Outperforms the three approaches. Bag of words, region based model and correspondence-LDA for classification and annotation	The geometry and appearance information of the objects is not taken into account.
Hidden Markov Model	Efficient scene classification and object detection	Prior probabilities should be known.
Spatial spectrum kernel	Accurate results	Complex kernel design
Spatial and semantic context approach	Easy to implement	The location details of the objects are not taken into account

Table 2: Comparison of Context Based Annotation Methods

Authors	**Methods**	**Features extracted**	**Classifier**	**Dataset**	**Performance Measure**
Yong Jae Lee and Kristen Grauman [3]	Graph Based Method	Texton Histo-grams(TH), Color Histograms (CH), and pyramid of HOG (pHOG)	SVM classifier	MSRC dataset, Corel and Pascal	Purity and mean Average Precision (mAP)
LuborLadicky, Chris Russell1, PushmeetKohli and Philip H.S. Torr1[4]	Graph Based Method	Color, location and Texton response	Baye's classifier	VOC dataset and MSRC dataset	Recall
Li-Jia LiRichard Socher and Li Fei-Fei [5]	Graph Based Method	SIFT features	Bayesian classifier	Flicker dataset	Precision and recall
Myung Jin Choi ,AntonioTorralba and Alan S. Willsky [6]	Tree Based Method	Global features	Bayesian classifier	SUN dataset and PASCAL dataset	Precision and recall
R. Zhang and Z. Zhang[7]	Tree Based Method	Color, texture, shape features	Bayesian classifier	Corel	Acuracy

Carolina Galleguillos Andrew Rabinovich Serge Belongie [8]	Model based on co-occurrences, location and appearances	SIFT features	Bayesian classifier	PASCAL dataset and MSRC dataset	Accuracy
Takahiro Okabe, Yuhi Kondo, Kris M. Kitani,And Yoichi Sato,*[9]*	Model based on co-occurrences, location and appearances	DoG and SIFT features	Bayesian classifier	PASCAL dataset	AUC (Area Under Curve)
Jun Li1 , Hongmei Zhang and Yuanjiang Liao [10]	Bag of Words Model	SIFT features and Shape features	Semi-supervised learning function based on the distance metrics	PASCAL dataset	Precision
Zhiwu Lu, Horace H. S. Ip, and Yuxin Peng [11]	Kernel based method	Gabor features and color features	Spatial spectral kernel classifier	UW dataset, IAPR dataset, PASCAL dataset, COREL daatset	Precision and recall

References

[1] S. MangijaoSingh, K. Hemachandran, "Content-Based Image Retrieval using Color Moment and Gabor Texture Feature", IJCSI International Journal of Computer Science Issues, Volume 9, Issue 5,pp:299-309,September 2012.

[2] AntonioTorralba, Kevin P. Murphy, William T. Freeman and Mark A. Rubin," Context-based vision system for place and object recognition",Massachusetts Institute of Technology, 2003 .

[3] Yong Jae Lee, Kristen Grauman, "Object-Graphs for Context-Aware Visual Category Discovery", IEEE Transactions on Pattern Analysis and Machine Intelligence, Volume 34, Issue 2, pp: 346-358, February 2012.

[4] LuborLadicky, Chris Russell1, PushmeetKohli and Philip H.S. Torr1, "Graph Cut based Inference with Co-occurrence Statistics". 11th European Conference on Computer Vision, Proceedings, Part V , Volume 6315, 2010, pp: 239-253, September 2010.

[5] Li-JiaLiRichardSocher and Li Fei-Fei, "Towards Total Scene Understanding: Classification, Annotation and Segmentation in an Automatic Framework", IEEE Conference on Computer Vision and Pattern Recognition, pp: 2036-2043,2009.

[6] Myung Jin Choi, Antonio Torralba and Alan S. Willsky, "A Tree-Based Context Model for Object Recognition", IEEE Transactions on Pattern

Analysis and Machine Intelligence, Issue No.02 , vol.34, pp: 240-252, February 2012.

[7] R. Zhang and Z. Zhang, "Effective image retrieval based on hidden Concept discovery in image database," IEEE Transactions on Image Process, Volume 16, Issue 2, pp: 562-572, February, 2007.

[8] Carolina Galleguillos Andrew Rabinovich Serge Belongie, "Object Categorization using Co-Occurrence, Location and Appearance", IEEE Transactions on Pattern Analysis and Machine Intelligence, pp: 1-8, February 2009.

[9] Takahiro Okabe, Yuhi Kondo, Kris M. Kitani, And Yoichi Sato, "Recognizing multiple objects based on co-occurrence of categories", The Pacific –Rim Symposium on Image and Video Technology (PSIVT) 2009, pp: 497-508.

[10] Jun Li, Hongmei Zhang and Yuanjiang Liao,"Image Annotation Based On Bag Of Visual Words And Optimized Semi-Supervised Learning Method", ICTACT Journal On Image And Video Processing: Special Issue On Video Processing For Multimedia Systems, Volume 5, Issue 01, August 2014.

[11] Zhiwu Lu, Horace H. S. Ip, and YuxinPeng, "Contextual Kernel and Spectral Methods for Learning the Semantics of Images", Volume:20, Issue: 6, pp:1739 - 1750 ,IEEE Transactions on Image Processing , 2011.

[12] Rabinovich, A. ; Vedaldi, A. ; Galleguillos, C. ; Wiewiora, E.," Objects in Context", Computer Vision, 2007. ICCV 2007.

[13] SpirosNikolopoulos, GeorgiosTh. Papadopoulos, IoannisKompatsiaris, IoannisPatras," Evidence-Driven Image Interpretation by Combining Implicit and Explicit Knowledge in a Bayesian Network", IEEE Transactions On Systems, Man, And Cybernetics,Volume:41 , Issue: 5, pp: 1366 - 1381, October 2011.

[14] P. Carbonetto, N. de Freitas, and K. Barnard, "A statistical model for general contextual object recognition", European Conference on Computer Vision, Volume 3021, pp: 350-362, 2004.

[15] A. Torralba, "Contextual Priming for Object Detection," International Journal for Computer Vision, Volume 53, Issue 2, pp: 169-191, July 2003.

[16] C. H. Hu, "Graph-based Semi-supervised Machine Learning", Zhejiang University, 2008.

[17] L. Wu, S. C. H. Hoi and N. Yu, "Semantics-preserving bag-of-words models and applications", Volume 19, Issue 7 ,pp: 1908-1920,IEEE Transactions on Image Processing, 2010.

[18] S. Belongie, J. Malik and Jan Puzicha, "Shape Matching and Object Recognition Using Shape Contexts", volume 24, pp: 509-522,IEEE Transactions on Pattern Analysis and Machine Intelligence 2002.

[19] M. R. Naphade, S. Basu, J. R. Smith, Ching-Yung Lin and B. Tseng, "Statistical modeling approach to content-based video retrieval", IEEE Proceedings of 16th International Conference .

[20] C.K. Chow and C.N. Liu, "Approximating Discrete ProbabilityDistributions with Dependence Trees," IEEE Trans. InformationTheory, May 1968.

[21] S.K. Divvala, D. Hoiem, J.H. Hays, A.A. Efros, and M. Hebert, "An Empirical Study of Context in Object Detection," Proc. IEEE Conference on Computer Vision and Pattern Recognition,pp: 1271 – 1278, June 2009.

[22] www.wikipedia.com

[23] Jianjun Liu, Zebin Wu, Zhihui Wei, Liang Xiao, and Le Sun ,"Spatial-Spectral Kernel Sparse Representation forHyperspectral Image Classification", Volume:6, Issue:6,pp: IEEE Journal Of Selected Topics In Applied Earth Observations And Remote Sensing, December 2013.