VISUAL RECOGNITION BY OBJECT-CENTRIC AND SCENE-CENTRIC METHODS

V. Suvetha[1], D. Chitra[2]

ABSTRACT

Computer vision is a field that includes methods for acquiring, processing, analyzing and understanding of images and also has been described as the enterprise of automating and integrating a wide range of processes and representations for vision perception. Visual recognition is one of the major problems in computer vision. It includes the problems of scene classification, image annotation, image retrieval, object recognition and object detection. Context is a rich source of information about an object identity, location and scale. A novel frame work to context modeling is based on the probability of co-occurrence of objects and scene is proposed. Images are represented by their posterior probabilities with respect to set of contextual models build upon the Bag-of-Features image representation. Representing images by posterior probabilities are remarkably noise-free and an effective model of the contextual relationships between semantic concepts. This modeling is based on two classes of representation. The first class consists of methods that model contextual relationships between sub-image entities. Methods in the second class adopt a scene-centric representation. Context models are learned from entire images by generating a holistic description of the scene or its gist.

Keywords : *Computer vision, context, image retrieval, object recognition, scene classification.*

I. INTRODUCTION

Visual recognition is one of the major problems in computer vision. It includes the problems of scene classification [3], [14], [15], [16], image annotation [1], [6], image retrieval [21], object recognition [8] and object detection [22]. Object recognition is an attempt to mimic the human capability to distinguish different objects in an image. The automatic object recognition concept is used in industry. Many applications involve recognition of patterns in data. The field of pattern or object recognition provides domain independent technique for data analysis. The recognition has been used to refer many different visual abilities, identification, categorization and discrimination.

Scene classification differs from the conventional object detection and image retrieval to the extent that a scene is composed of several entities often organized in an unpredictable layout. A given scene, it is virtually impossible to define a set of properties that would be inclusive of all its possible visual manifestations.

There are two main elements in an image classification system. The first one refers to the computation of the

[1]*Assistant Professor, Department of Information Technology., PA College of Engineering & Technology,*
E-mail : suvethame@gmail.com

[2]*Professor and Head, Department of Computer Science & Engineering, PA College of Engineering & Technology,*
E-mail : chitrapacet@gmail.com

feature vector representing an image and the second is the classifier, the algorithm that classifies an input image into one of the predefined category using the feature vector. Classifying scenes such as mountains, forests and offices is not an easy task owing to the variability, ambiguity and wide range of illumination.

In general there are two basic strategies found in scene classification. The first uses low level features such as global color or texture histograms, the power spectrum and is normally used to classify only a small number of scene categories indoor versus outdoor. The second strategy uses an intermediate representation before classifying scenes and has been applied to cases that are larger number of scene categories. The last decade has produced significant advances in visual recognition. These methods follow a common recognition strategy that consists of

1) Identifying a number of visual classes of interest,

2) Designing a set of appearance features or some other visual representation,

3) Postulating an architecture for the classification of those features,

4) Relying on sophisticated statistical tools to learn optimal classifiers from training data.

The resulting classifiers are referred as strictly appearance based classifiers [7]. Recent innovations produced better features. The ubiquitous SIFT descriptor, methods for fast object matching, sophisticated discriminate classifiers such as Support Vector Machines (SVMs) with various kernels tuned for vision [1] and sophisticated statistical models [3]. Compared to the recognition strategies of biological vision, strictly appearance-based methods have the limitation of not exploiting contextual cues. Psychophysical studies have also shown that context can depend on multiple clues [2]. Object recognition is known to be affected by several properties such as support objects do not float in the air, interposition objects occupy different volumes, probability objects appear in different scenes with different probabilities, position objects appear in typical locations and size objects have typical relative sizes [1].

In recent years, several efforts had taken to account for context in recognition. Such efforts can be broadly classified into two classes. The first consists of methods that model contextual relationships between sub-image entities. Methods in the second class learn a context model from the entire image and generating a holistic representation of the scene known as its gist. More precisely, image features are not grouped into regions or objects, but treated in a holistic scene-centric framework. Various recent works have shown that semantic descriptions of real world images can be obtained with these holistic representations, without the need for explicit image segmentation. The holistic representation of context has itself been explored in two ways. One approach is to rely on the statistics of low-level visual measurements that span the entire image. A second approach is to adopt the popular Bag-of-Features representation to compute low-level features locally and aggregate this across the image to form a holistic context model. These methods usually ignore spatial information.

An approach to context modeling based on the probability of co-occurrence of objects and scenes. This modeling is quite simple, and builds upon the availability of robust

appearance classifiers. A vocabulary of visual concepts is defined and statistical models learned for all concepts with appearance modeling techniques. These techniques are typically based on the Bag-of-Features (BoF) representation. The outputs of the appearance classifiers are interpreted as the dimensions of a semantic space. Images are represented by the vector of its posterior probabilities under each of the appearance models. This vector is denoted as a Semantic Multinomial (SMN) distribution.

II. Related Work

There are many works done regarding object-centric and scene-centric approaches. In the earlier works of object centered approach uses to represent object intrinsic features exclusively for performing object detection and recognition tasks. Antonio Torralba uses a scheme that includes context information in object representations and to demonstrate its role in facilitating individual object detection. This approach is based on using the differences of the statistics of low-level features in real-world images [13]. In that object locations and scales can be inferred from a simple holistic representation of context based on the spatial layout of spectral components that captures low-resolution spatial and spectral information of the image [18].

Exploiting both local image data as well as contextual information Torralba et al. introduces Boosted Random Fields (BRFs), uses boosting to learn the graph structure and local evidence of a Conditional Random Field (CRF). Boosting is a simple way of sequentially constructing strong classifiers from weak components and has been used for single class object detection with great success. [19].

Fink et al. proposed an efficient method for detection of multiple objects in complex scenes. In this dynamic programming mutual boosting, multiple detectors of objects and parts are trained simultaneously using AdaBoost and object detectors might combine the remaining intermediate detectors to enrich the weak learner set [4]. Boosting examines a larger features set during training but it has the disadvantage that an iteration of mutual boosting detection of M objects is time consuming process.

Sivic et al. proposed a method to automatically discover the visual categories present in the data and localize the visual categories in the image from a set of unlabelled images. It investigates three major areas including topic discovery categories are discovered by pLSA , classification of unseen images – topics corresponding to object categories are learnt on one set of images and used to determine the object categories present in another set and object detection is to determine the location and approximate segmentation of objects in each image [17].

A method for recognizing scene categories based on approximate global geometric correspondence is given by Lazebnik et al [9]. The technique works by partitioning the image into increasingly fine sub-regions and computing histograms of local features found inside each sub-region. The resulting spatial pyramid is a simple and computationally efficient extension of an order less Bag-of-Features (BoF) image representation. The result shows that global representations can be surprisingly effective not only for identifying the overall scene for categorizing images as containing specific objects.

Li et al. proposed a Bayesian hierarchical model for learning natural scene categories to learn and recognize natural scenes. The model represents the image of a scene by a collection of local regions, denoted as code words obtained by unsupervised learning. Each region is represented as part of a theme. It provides a principled approach to learning relevant intermediate representations of scenes automatically and without super vision and also it uses a principled probabilistic framework for learning models of textures via code words [10].

Zhang et al., proposed an approach that permits recognition based on color, texture and particularly shape in a homogeneous framework. This approach can be applied to large, multiclass data sets it outperforms nearest neighbor and support vector machines. In this framework, scaling to a large number of categories does not require adding new features [24].

A novel approach for visual scene modeling and classification investigating the combined use of text modeling methods and local invariant features was proposed by Quelhas et al. 2007. The frame work for scene classification integrates scale-invariant feature extraction and latent space modeling methods. This approach uses probabilistic Latent Semantic Analysis (pLSA) for scene ranking and clustering. pLSA is able to automatically capture meaningful scene aspects from data, scene similarity is evident and was useful to explore the scene structure of an image collection and turning it into a tool with potential in visualization, organization, browsing and annotation of images in large collections was proposed by (Hofmann 1999). The approach exploits the output of one-against-all classifiers to derive multiple class labels [14].

An approach for automatically discovering intermediate concepts from scenes by Maximization of Mutual Information (MMI) was proposed by Liu et al [12]. The approach to capture the spatial information of the semantic concepts in the scene, the Spatial Pyramid Matching (SPM) and weighted Spatial Concept Corellogram (SCC) [22] is used. The method uses SVM as a classifier to train and test the models.

A nonparametric, data–driven model for image features that captures spatial dependencies via a multi scale graphical model was proposed by Kivinen et al [8]. The individual features or wavelet coefficients are marginally described by Dirichlet Process (DP) mixtures, yielding the heavy-tailed marginal distributions characteristic of natural images. Dependencies between features are then captured with a hidden Markov tree and Markov chain Monte Carlo methods used to learn models latent state space.

A framework for object categorization named CoLA for Co-occurrence, Location and Appearance was proposed Galleguillos et al. 2008 uses a Conditional Random Field (CRF) to maximize object label agreement according to both semantic and spatial relevance. Model the relative location between objects using simple pair wise features. By vector quantizing the feature space learn a small set of prototypical spatial relationships directly from the data [5], [11].

A new descriptor for images that allows the construction of efficient and compact classifiers with good accuracy on object category recognition was given by Torresani et al.. The descriptor is the output of a large number of weakly trained object category classifiers on the image. The trained categories are selected from ontology of visual

concepts and it accepts the existing object category classifiers often encode ancillary image characteristics. The advantage of the descriptor is that it allows object-category queries to be made against image databases using efficient classifiers efficient at test time such as linear support vector machines [20].

An implementation of contextual modeling was proposed, in this approach concepts are modeled as mixtures of Gaussian distribution on appearance space and mixtures of Dirichlet distributions on semantic space. This model shows that 1) the contextual representation outperforms the appearance based representation and 2) this holds irrespectively of the choice and accuracy of the underlying appearance models. This model combines the object-centric and scene-centric approaches.

III. Appearnce Based Models

A. Appearance Based Classifiers

At the visual level, images are characterized as observations from a random variable X, defined on some feature space χ of visual measurements. χ could be the space of Discrete Cosine Transform (DCT) or SIFT (Scale Invariant Feature Transform) descriptors. Each image is represented as a bag of N feature vectors $I = \{x_1, \ldots, x_N\}$, $xi \in \chi$ assumed to be sampled independently. Images are labeled according to a vocabulary of semantic concepts $L = \{w_1, \ldots, w_L\}$. Concepts are drawn from a random variable W, takes values in $\{w_1, \ldots, w_L\}$. Each concept induces a probability density on χ by using the Equation (1)

$$P_{x/w}(I/w) = \Pi P_{x/w}(x_j|w) \qquad (1)$$

The densities $P_{x/w}(x_{/w})$ are learned from a training set of images $D = \{I_1, .., I_{|D|}\}$, annotated with captions from the concept vocabulary L. For each concept w, the concept density $P_{x/w}(x|w)$ is learned from the set D_w of all training images and caption includes the w^{th} label in L. $P_{x/w}(x|w)$ is an appearance based model and the observations drawn from concept w in the visual feature space χ. Given an unseen test image I, minimum probability of error concept detection is achieved with a Baye"s decision rule based upon the posterior probabilities for the presence of concepts $w \in L$ given a set of image feature vectors I by using the Equation (2)

$$P_{w/x}(w/I) = \frac{P_{x/w}(I/w)\, P_w(w)}{P_x(I)} \qquad (2)$$

B. Designing Semantic Space

The concept detection only requires the largest posterior concept probability for a given image and it is possible to design a semantic space by retaining all posterior concept probabilities. A semantic representation of an image I_y, can be obtained by the vector of posterior probabilities, $\Pi^y = (\Pi_1^y, \ldots, \Pi_L^y)^T$ and Π_w^y denotes the probability $P_{w/x}(w|I)$. This vector is referred to as a Semantic Multinomial (SMN) and lies on a probability simplex S, referred to as the semantic space. In this way, the representation establishes a one-to-one correspondence between images and points Π^y in S.

Appearance-based object or concept recognition system can be used to produce the posterior probabilities in Π^y. These probabilities can even be produced by systems that do not learn appearance models explicitly. This is achieved by converting classifiers scores to a posterior probability distribution by using probability calibration

techniques. The distance from the decision hyper plane learned with Support Vector Machines (SVM) can be converted to a posterior probability using a simple sigmoid function.

C. Limitations of Appearance Based Models

The performance of strict appearance-based modeling is upper bounded by two limitations:

1) Contextually unrelated concepts can have similar appearance features

2) Strict appearance models cannot account for contextual relationships[23].

Image patches frequently have ambiguous interpretation that makes it compatible with many concepts if considered in isolation. Second, strictly appearance-based models lack information about the interdependence of the semantics of the patches that compose the images in a class.

In the first case, a patch can accidentally co-occur with multiple concepts a property usually referred to as polysemy in the text analysis. In the second, patches from multiple concepts typically co-occur in scenes of a given class the equivalent to synonymy for text. Only the co-occurrences of the second type are indicative of true contextual relationships and SMN distributions learned from appearance-based models capture both types of co-occurrences.

IV. Visual Recogniton By Context Based Models

The possibility to deal with the ambiguity of the semantic representation is to explicitly model contextual dependencies. This can be done by introducing constraints on the appearance representation by modeling constellations of parts or object relationships [25]. The introduction of such constraints increases complexity and reduces the invariance of the representation. A more robust alternative is to keep Bag-of-Features (BoF) and represent images at a higher level of abstraction there by ambiguity can be more easily detected. This is the strategy proposed in this work and the fact is that these two types of SMN co-occurrences have different stability to extract more reliable contextual features.

A. Semantics to Context

The basic idea is that, images from the same concepts are expected to exhibit similar contextual co-occurrences and this is not likely for ambiguity co-occurrences. By definition ambiguity co-occurrences are accidental and it is impossible to detect from a single image. Stable contextual co-occurrences should be detectable by joint inspection of all SMNs derived from the images of a concept. This is accomplished by extending concept modeling by one further layer of semantic representation. Each concept w is modeled by the probability distribution of the SMNs derived from all training images in its training set, Dw. This SMN distribution is referred as the contextual model for w. If Dw is large and diverse, this model is dominated by the stable properties of the features drawn from concept w. In this case, the features are SMNs and stable properties are the true contextual relationships of w. Concept models assign high probability to regions of the semantic space occupied by contextual co-occurrences and small probability to those of ambiguity co-occurrences. Considering that streets typically co-occur with buildings. The contextual model for street assigns high probability to SMNs that include both

concepts and street only co-occurs accidentally with bedroom, SMNs including this concept receive low probability. Representing images by using posterior distribution under contextual models emphasizes contextual co-occurrences. Posterior probabilities at this higher level of abstraction are referred as contextual features and the probability vector associated with each image as a contextual multinomial distribution.

B. Learning Contextual Concept Models

Extenuate the effects of ambiguity co-occurrences, contextual concept models are learned in the semantic space S, from the SMNs of all images that contain each concept. A concept w in L is shown to induce a sample of observations on the semantic space S. S is itself a probability simplex and it is assumed that this sample is drawn from a mixture of Dirichlet distributions by using the Equation (3) and (4).

$$P_{\Pi|W}(\Pi|w;\Omega^w) = \Sigma_k \beta_k^w \mathrm{Dir}(\Pi;\alpha_k^w) \quad (3)$$

The contextual model for concept w is characterized by a vector of parameters $\Omega^w = \{ \beta_k^w, \alpha_k^w \}$, here β_k is a probability mass function ($\Sigma_k \beta_k^w = 1$) and $\mathrm{Dir}(\Pi;\alpha)$ is a Dirichlet distribution of parameter $\alpha = \{\alpha_1, \ldots, \alpha L\}$ and r (.) is the Gamma function.

$$\mathrm{Dir}(\Pi;\alpha) = \frac{\Gamma\left(\Sigma_{i=1}^{L} \alpha_i\right) \Pi(\Pi i)^{\alpha i - 1}}{\Pi_{i=1} \Gamma(\alpha_i)} \quad 4)$$

The parameters Ω^w are learned from the SMNs Π_n of all images in D_w, the images annotated with the w^{th} concept. For this, rely on maximum likelihood estimation using the Generalized Expectation-Maximization (GEM) algorithm. GEM is an extension of the well known EM algorithm. It consists of two steps. The E-Step is identical to that of EM, computing the expected values of the component probability mass β_k. The generalized M-step estimates the parameters α_k.

C. Contextual Space

The contextual concept models $P_{\Pi|W}(\Pi|w)$ play in the semantic space S is a similar role to that of the appearance based models $P_{X|W}(x|w)$ in visual space χ by using the Equation(5). It follows the Minimum Probability of Error (MPE) concept detection on a test image I^y of SMN can be implemented with a Bayes decision rule based on the posterior concept probabilities,

$$P_{w|\Pi}(w|\Pi^y) = \frac{P_{\Pi|w}(\Pi^y|w)P_W(w)}{P_\Pi(\Pi^y)} \quad (5)$$

This is the semantic space equivalent of Equation (2) and once again assume a uniform concept prior $P_W(w)$.

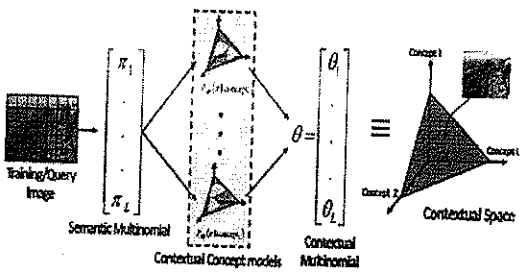

Figure 1 : Contextual Multinomial of an image

It is also possible to design a new semantic space by retaining all posterior concept probabilities $\theta_w = P_{W|\Pi}(W|\Pi^y)$. The vector $(\theta_1^y, \ldots, \theta_L^y)^T$ as the Contextual Multinomial (CMN) distribution of image I^y. Figure 1 shows CMN vector lie on a new probability simplex C, this is referred to as the contextual space.

V. Results And Discussion

A number of object recognition experiments were performed on N15 data set to evaluate the impact of parameters of the contextual representation on recognition performance. Recognition proceeds in two steps: Segmentation and Classification as shown in Figure 2.

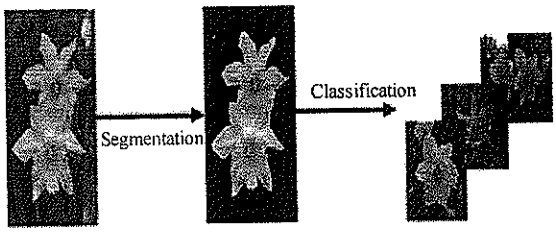

Figure 2 : Object Recognition Stages

The N15 dataset comprises of images from 15 natural scene categories. These images are used to learn concept densities and also used as test set. Recognition experiments are repeated 5 times with different randomly selected train and test images. The SMN and CMN vectors computed from each image. From the results the proposed CMNs are remarkably noise free for all semantic spaces considered. This captures the gist of the underlying scenes and assigning high probability only to truly contextual concepts. Using the SIFT descriptors and DCT the contextual modeling gives better accuracy results for scene classification and retrieval also the Bayesian rule plays an important role in this recognition. From the Figure 3 the performance of context based models is better than the existing appearance based models.

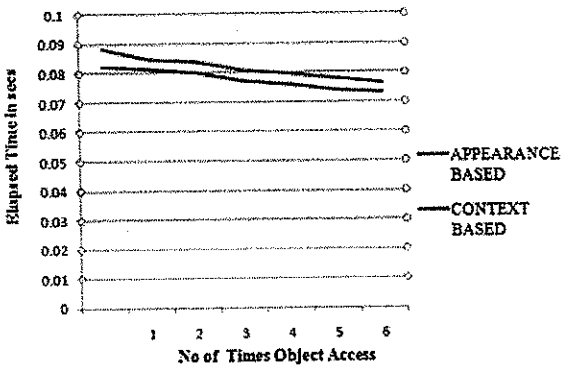

Figure 3 : Performance of Context Based Models

VI. Conclusion

The proposed modeling is quite simple and builds upon the availability of robust appearance classifiers. Images are represented by posterior probabilities with respect to a set of contextual models and built upon the bag-of features image representation through two layers of probabilistic modeling. The images are then represented by its posterior probabilities with respect to these distributions. The overall representation is similar to a topic model and topics are learned in a supervised manner. Supervised learning is a necessary condition for overcoming the semantic gap between the low-level patch representation and the higher level contextual relationships.

References

[1]. Biederman. I, Mezzanotte. R and Rabinowitz. R, *"Scene Perception: Detecting and Judging Objects Undergoing Relational Violations"*, Cognitive Psychology, vol. 14, pp. 143-77, 1982.

[2]. Carneiro. G, Chan. A, Moreno. P and Vasconcelos. N, *"Supervised Learning of Semantic Classes for Image Annotation and Retrieval"*, IEEE Trans.

Pattern Analysis and Machine Intelligence, vol. 29, no. 3, pp. 394-410, 2007.

[3]. Chan. A and Vasconcelos. N, *"Probabilistic Kernels for the Classification of Auto-Regressive Visual Processes"*, Proc. IEEE Conf. Computer Vision and Pattern Recognition, vol. 1, 2005.

[4]. Feng. S, Manmatha. R and Lavrenko. V, *"Multiple Bernoulli Relevance Models for Image and Video Annotation"*, Proc. IEEE Conf. Computer Vision and Pattern Recognition, 2004.

[5]. Galleguillos. C, Rabinovich. A and Belongie. S, *"Object Categorization Using Co-Occurrence Location and Appearance"*, Proc. IEEE Conf. Computer Vision and Pattern Recognition, 2008.

[6]. Heitz. G and Koller. D, *"Learning Spatial Context: Using Stuff to Find Things"*, Proc. 10th European Conf. Computer Vision, pp. 30-43, 2008.

[7]. Hofmann. T, *"Probabilistic Latent Semantic Indexing"*, Proc. ACM SIGIR Conf. Research and Development in Information Retrieval, 1999.

[8]. Kivinen. J, Sudderth. E and Jordan. M, *"Learning Multiscale Representations of Natural Scenes Using Dirichlet Processes"*, Proc. IEEE Int Conf. Computer Vision, 2007.

[9]. Lazebnik. S, Schmid. C and Ponce. J, *"Beyond Bags of Features: Spatial Pyramid Matching for Recognizing Natural Scene Categories"*, Proc. IEEE Conf. Computer Vision and Pattern Recognition, 2005.

[10]. Li. F.-.F and Perona. P, *"A Bayesian Hierarchical Model for Learning Natural Scene Categories"*, Proc. IEEE Conf. Computer Vision and Pattern Recognition, pp. 524-531, 2005.

[11]. Lim. J, Arbela´ez. P, GU. C and Malik. J, *"Context by Region Ancestry"*, Proc. IEEE Int Conf. Computer Vision, 2010.

[12]. Liu. J and Shah. M, *"Scene Modeling Using Co-clustering"*, Proc. IEEE Int Conf. Computer Vision, 2007.

[13]. Oliva. A and Torralba. A, *"Building the Gist of A Scene: The Role of Global Image Features in Recognition"*, Progress in Brain Research: Visual Perception, vol. 155, pp. 23-36, 2006.

[14]. Quelhas. P, Monay. F, Odobez. J, Gatica-Perez. D and Tuytelaars. T, *"A Thousand Words in a Scene"*, IEEE Trans. Pattern Analysis and Machine Intelligence, vol. 29, no. 9, pp. 1575-1589, 2007.

[15]. Rasiwasia.N and Vasconcelos.N, *"Scene Classification with Low- Dimensional Semantic Spaces and Weak Supervision"*, Proc. IEEE Conf. Computer Vision and Pattern Recognition, 2008.

[16]. Rasiwasia. N and Vasconcelos. N, *"Holistic Context Modeling Using Semantic Co-Occurrences"*, Proc. IEEE Conf. Computer Vision and Pattern Recognition, 2009.

[17]. Sivic. C, Russell. B, Efros. A, Zisserman. A and Freeman. W, *"Discovering Object Categories in Image Collections"*, Proc. IEEE Int Conf. Computer Vision, vol. 1, p. 65, 2005.

[18]. Torralba. A, *"Contextual Priming for Object Detection"*, Int J. Computer Vision, vol. 53, pp. 169-191, 2003.

[19]. Torralba. A, Murphy. K and Freeman. W, *"Contextual Models for Object Detection Using Boosted Random Fields"*, Proc. Advances in Neural Information Processing Systems, 2004.

[20]. Torresani. L, Szummer. M and Fitzgibbon. A, *"Efficient Object Category Recognition Using Classemes"*, Proc. 11th European Conf. Computer Vision, pp. 776-789, 2010.

[21]. Vasconcelos. N, *"Minimum Probability of Error Image Retrieval"*, IEEE Trans. Signal Processing, vol. 52, no. 8, pp. 2322-2336, 2004.

[22]. Vogel. J and Schiele. B, *"A Semantic Typicality Measure for Natural Scene Categorization"*, Proc. DAGM04 Ann. Pattern Recognition Symp, 2004.

[23]. Wang. G, Hoiem. D and Forsyth. D, *"Learning Image Similarity from Flickr Groups Using Stochastic Intersection Kernel Machines"*, Proc. IEEE Int Conf. Computer Vision, pp. 428-435, 2009.

[24]. Zhang. H, Berg. A, Maire. M and Malik. J, *"SVM-KNN: Discriminative Nearest Neighbor Classification for Visual Category Recognition"*, Proc. IEEE Conf. Computer Vision and Pattern Recognition, 2006.

[25]. Rasiwasia. N and Vasconcelos. N, *"Holistic Context Models For Visual Recognition"* IEEE Trans. Pattern Analysis and Machine Intelligence, vol. 34, no. 5, 2012.

Author's Biography

Dr. D. Chitra, is a Professor in the Department of Computer Science and Engineering at P. A. College of Engineering and Technology, Pollachi, Coimbatore. She received her M.E. Degree in CSE from Anna University, Chennai and PhD degree in CSE from Anna University of Technology, Coimbatore. Her resource interests include image analysis, Pattern recognition and Computer Vision. She is a life member of ISTE.

Ms. V. Suvetha, is an Assistant Professor in the Department of Information Technology at P. A. College of Engineering and Technology, Pollachi, Coimbatore. She received her B.E. Degree in CSE from Anna University, Coimbatore and M.E. degree in CSE from Anna University Chennai. Her research interest includes image processing and Analysis.