A Survey on Scene Recognition

Anu E, Anu K S

Abstract— **Scene recognition provides visual information from the level of objects and the relationship between them. The main objective of scene recognition is to reduce semantic gap between human beings and computers on scene understanding. For example, recognize the context of an input image and categorize it into scenes (forest, seashore, building etc). Some of the applications of scene recognition are object recognition, object detection, video text detection etc. Different methods are there for scene recognition and understanding. All are having positive and negative aspects. One of the main difficulties is to increase the accuracy. The negative aspects which affect the improvement in accuracy are, intrinsic relationship across different scales of the input image are not analyzed and impact of redundant features. So a new framework is suggested to overcome these limitations. The suggested framework combine multitask model and sparse feature selection based manifold regularization (SFSMR), and proposes the semi supervised learning method to identify the input scene more accurately.**

***Index Terms*—Multitask Model, Object Detection, SFSMR, Semi Supervised Learning**

I. INTRODUCTION

Image processing is the process of analysis and manipulation of a digitized image in order to improve its quality. Two principles of Image processing are improvement of pictorial information and processing of scene data. Recognizing the semantic category of complex scenes having content variations is a challenging task. A new technique is introduced to overcome the limitations in order to increase the accuracy. This technique includes two modules. First module is a multitask model to integrate different scales of the input image. The second module is a model of sparse feature selection based manifold regularization (SFSMR) to select the optimal information along with preserving the underlying manifold structure of input image. The suggested framework combine multitask model and SFSMR, and propose the semi supervised learning method to identify the input scene by reducing the two limitations. The main advantages of proposed method compared to existing methods are the following. First, the usage of multiresolution images generated from the same scene because these different scale images include same global spatial structures and different local features. So the different tasks can be analyzed in a joint framework. This helps to improve the performance in scene recognition. Second, optimal features can be chosen along with preserving the underlying manifold structure of each feature data. Third, the usage of $l2,1$ norm term [1] and the trace norm term helps to explore the correlation of different features at multiresolution and to transfer the information from different tasks among the generated multiple tasks.

***Manuscript received Dec , 2015*.**

***ANU E,** Department of Computer Science and Engineering, KMCT College of Engineering, Calicut University Calicut, India*

***ANU K.S,** Department of Information Technologyg, Calicut University, Calicut, India.*

II. LITERATURE SURVEY

Many scene recognition techniques try to build an intermediate semantic representation to reduce semantic gap. These methods focus on extracting low level features from single resolution image. It may fail to represent the entire scene completely. Some redundant features may reduce the accuracy of scene recognition. So a detailed survey is required to check whether the features are useful to recognize the semantic category of an input image.

Chang Cheng [3] introduced a novel outdoor scene image segmentation algorithm based on the background recognition and perceptual organization is introduced. The background objects such as the sky, the river is recognized based on the color and texture information. A perceptual organization model is introduced for structurally challenging objects. This model can capture the non accidental structural relationships in the constituent parts of the objects. Group them together and that is independent on a priori knowledge of the specific objects. Perceptual organization model (POM) for boundary detection is developed. The POM incorporates a list of Gestalt laws in a qualitative manner. So it is able to capture the nonaccidental structural relationships in the constituent parts of a structured object.

Under different outdoor environments, this model helps to identify the boundaries of various salient structured objects. A salient structured object is a structured object having an independent and visible physical boundary. The boundary of the object that not contained in another structured object is called an independent physical boundary. There are some limitations for the proposed idea. The geometric relationships between different object parts are the main basics for POM segmentation. So obtaining the geometric properties of object parts is a necessary task. The object parts may have homogenous surfaces, so the uniform regions in an image correspond to object parts. Another problem source is strong reflection. There exist some object classes with very complex structures and some parts of the objects may not strongly attach to other parts of the object. In this case, POM may not be able to piece the entire object together and these objects require higher level object specific knowledge to segment the entire objects.

Yanfei Zhong [4] suggests scene classification is an effective tool for semantic interpretation of high spatial resolution (HSR) remote sensing image. The probabilistic

ISSN: 2278 – 7798

topic model (PTM) can be applied to natural scenes by using a single feature but it is inadequate for HSR images due to its complex structure. So a technique called SAL PTM is introduced to combine multiple features. That is a semantic allocation level (SAL) multi feature fusion strategy based on PTM for HSR imagery is proposed. In SAL PTM the complementary spectral, texture, and scale invariant feature transform features are combined in an effective manner. By using *k* means clustering, the three features are extracted and quantized separately. This provides appropriate low level feature descriptions for the semantic representations. PTM captures the latent semantic allocations of the three features.

In the proposed SAL PTM, the LDA model and the pLSA model is used to capture the semantic information from the HSR images. It is performed on the basis of an adequate image representation generation strategy and a suitable latent semantic allocation mining procedure. The main contributions of this technique are; an effective feature description method for HSR Imagery, an appropriate image representation generation strategy for HSR Imagery and an adequate latent semantic allocation mining procedure for PTM Based HSR image scene classification. But in this method focus on normalization constraint of the PTM is required and structural features which are more appropriate for HSR images should be explored.

Yongzhen Huang [5] brought the idea of using genetic programming (GP) to generate composite operators and composite features from combinations of primitive operations and primitive features in object detection. The main reason for using GP is to overcome the human expert's limitations occurred in the feature synthesis. These limitations are the result of focusing only on conventional combinations of primitive image processing operations. In order to improve the efficiency of GP and a new fitness function is designed. It is based on minimum description length principle. This helps to incorporate both the pixel labeling error and the size of a composite operator into the designed fitness evaluation process. The other techniques added to improve the efficiency are, smart crossover, smart mutation and a public library ideas.

Lin [6] introduced EBIM. Biologically inspired model (BIM) is effective method for object recognition. However, BIM has some limitations. It includes very heavy computational cost as it needs dense input. There is a disputable pooling operation in modeling relations of the visual cortex. Another limitation is presence of blind feature selection in a feed forward framework. To overcome these problems, an enhanced BIM (EBIM) is introduced. EBIM is more effective and efficient than BIM. This method removes uninformative input by introducing sparsity constraints. A novel local weighted pooling operation is used with stronger physiological motivations. Then a feedback procedure is applied which helps to select effective features for combination.

Anna Bosch [7] introduced a hybrid discriminative approach. In this approach a set of labeled images of scenes is provided. The aim of this approach is to classify a new image into one of the categories (e.g., coast, forest, building, etc.). First discover latent topics using probabilistic Latent Semantic Analysis (pLSA). For each image a generative model from the statistical text literature is applied to a bag of visual words representation and training a multiway classifier on the topic distribution vector for each image. A novel vocabulary using dense color SIFT descriptors is introduced. The classification performance will be investigated under the changes in the size of the visual vocabulary, the number of latent topics learned, and the type of discriminative classifier used (k nearest neighbor or SVM)

Probabilistic Latent Semantic Analysis (pLSA) is a generative model which is used in text analysis to discover topics in a document using the bag of words document representation. Here images are considered as documents, and object categories as topics. So an image with instances of several objects is modeled as a mixture of topics. By using a visual analog of a word, the models are applied to images. pLSA is appropriate in the case of multiple object categories per image as it provides a correct statistical model for clustering. By using pLSA, the proposed scene classifier learns topics and their distributions in unlabeled training images and then uses their distribution in test images as a feature vector in a supervised discriminative classifier. But here, the images with a semantic transition between categories are not well clustered because no sufficient ambiguous images are there.

Jianxin Wu [8] suggests CENTRIST (CENsus TRansform hISTogram), a visual descriptor for recognizing scene categories (mainly for indoor environments). CENTRIST is a holistic representation. It captures the structural properties by modeling distribution of local structures. Rough geometrical information is captured by using a spatial CENTRIST representation. The desired properties of CENTRIST are the following. First one is holistic representation; the perceptual properties like degree of naturalness can be captured by the holistic representation Gist, and can be successfully used to recognize scene categories. Second one is capturing the structural properties; ability to capture general structural properties like rectangular shapes, tiles, etc. along with suppressing detailed textural information. Third one is rough geometry; strong geometrical constraints are very useful in object recognition. Fourth one is generalizability; the learned category concepts can be applied to new images. Either feature descriptors are compact within a category or they may be far apart when they belong to different categories. CENTRIST limitations are; it is sensitive to rotations so not suitable for multi view object recognition. Second limitation is, it is designed to recognize shape categories and it is not a shape descriptor. The third one is CENTRIST ignores color information.

Li Fei-Fei [9] examined a bayesian hierarchical model. In this model, an input image is represented as a collection of local patches. Each patch of the input image is represented using a codeword. Codeword is taken from a large vocabulary of codewords called codebook which is obtained

ISSN: 2278 – 7798

from 650 training examples from all 13 categories (around 50 images for each category). Image patches are detected using a sliding grid and random sampling of scales. The goal is to get a model that represents the distribution of these codewords in each category of scenes more accurately. In recognition phase, first identify all the codewords corresponding to unknown input image. Then determine the best category model that represents the distribution of the codewords of the particular image. The key idea is the usage of intermediate representations to improve performance as they are composed of mixtures codewords and to avoid usage of manually labeled or segmented images to train the system, if possible at all. Local regions are more robust to occlusions and spatial variations than global features.

In this framework, initially the local regions are clustered into different intermediate themes and then into categories. Probability distributions of the local regions and the intermediate themes are learnt in an automatic way. Different ways such as Evenly Sampled Grid, Random Sampling etc. are there for extracting local regions. Normalized 11 × 11 pixel gray values are used for describing a patch in the input image. By Using all the detected patches of training images including all categories, learn the codebook with the help of K means algorithm. The summary of the framework is; it provides a principled probabilistic framework for learning relevant intermediate representations of scenes automatically without supervision and helps to group categories of images into a sensible hierarchy.

Jian Yao [10] provided an approach to holistic scene understanding that reasons jointly about regions, location, class and spatial extent of objects, presence of a class in the image, as well as the scene type. Here in this approach learning and inference are efficient as it introduces auxiliary variables which decompose the inherent high order potentials into pair wise potentials between a few variables with small number of states. Inference is performed using convergent message passing algorithm. Unlike graph cuts inference, this algorithm has no sub modularity restrictions and potential specific moves are not required.

The aim of holistic scene understanding is recovering multiple related aspects of a scene to provide a deeper understanding of the scene as a whole. In this approach, the segments and super segments reason about the semantic class labels to be assigned for each pixel in an image. Super segments are employed to create longer range dependencies. It provides computational reasons, as they are fewer in number it is easy to connect them more densely to other parts of the model. The model can accept or reject the detections with the help of binary variables corresponding to each candidate bounding box generated by an object detector. The role of binary class variables is to reason about which classes are present in an image. But some of the main sources of error are bad unary potentials and false negative detections

Xiaodong Yu [11] introduced an active vision framework called active scene recognition is introduced for utilizing high level knowledge for scene recognition. The proposed approach consists of two modules. First one is a reasoning module, which is used to obtain higher level knowledge about scene and object relations, proposes instructions to the second module and draws conclusions about the scene contents. The second one is sensory module, which includes a set of visual operators. It is responsible for extracting features from images, detecting and localizing objects and actions. The sensory module does not passively process the image and it is guided by the reasoning module. The approach is based on an iterative process. The reasoning module decides what and where the sensory module should process next. Thus the sensory module focuses small number of objects at selected locations of the scene. This will provide faster and more accurate scene recognition procedure.

This scene recognizer classifies a scene by detecting the objects inside it in an iterative manner. In the k th iteration, the reasoning module provides an instruction to the sensory module for searching an object called O_k within a particular region L_k in the image. Then the sensory module runs the object detector and provides a response. This response includes the highest detection score d_k and the location of the object l_k. The reasoning module receives this response and analyses the same. After that the reasoning module will start a new iteration. This iteration is continued until any termination criteria are satisfied. To implement an active scene recognizer, following components are required. First one is a sensory module for object detection. Second one is a reasoning module for predicting the scene class based. Third one is a strategy for deciding which object and where in the scene the sensory module should process in the next iteration and the last one is a strategy for initializing and terminating the iteration.

Lorenzo Torresani [12] introduced a new descriptor is for images that allow the construction of efficient and compact classifiers with good accuracy on object category recognition. The descriptor is the result of a large number of weakly trained object category classifiers. The advantage of this descriptor is, by using efficient classifiers such as linear support vector machines, it allows object category queries to be made against image databases. The proposed system is a form of classifier combination because the components of the descriptor are the outputs of a group of predefined category specific classifiers applied to the image. The idea is that a novel category (called duck) is expressed in terms of the outputs of base classifiers (called Classemes). It describes either objects similar to ducks, or objects seen in conjunction with the ducks. Because these base classifier outputs provide a rich coding of the image, great accuracy will be achieved along with satisfying the requirements.

There are two distinct stages for the proposed method. ‘classeme learning’ and ‘using the classeme’. Distinct training sets for each of the two stages. In ‘classeme learning’ a set of category labels is drawn and for each category a set of training images is collected by querying on the category label to an image search engine. A classifier will be trained for each category. Provide the classeme vectors for all training images. It may desire to perform some sort of feature selection on the descriptors. The training images will be converted to classeme vectors. Any classifier can be

ISSN: 2278 – 7798

trained taking the classeme vectors as input. Simple and fast classifiers applied to the classemes can increase accuracy. But the problem is, it is not expected that these base categories will provide useful semantic labels, of the form grass, water, sky etc...

ANTONIO TORRALBA [13] suggested Contextual Priming for Object Detection. The main idea is that the context is a rich source of information about an object's identity, location and scale. A simple framework for modeling the relationship between context and object properties is introduced. The framework is based on the correlation between the statistics of low level features across the entire scene and the corresponding objects. The goal is to use context information in object representations and to show how it facilitates individual object detection. This approach uses the differences of the statistics of low level features in real world. A low dimensional holistic representation that encodes the structural scene properties is used. In this approach, one may expect to find a strong correlation between the objects present in the scene and the statistics of local low level features in the overall scene. This is because of the relationship between objects and context categories in real world scenes.

The main problem faced by the computational recognition approaches when including contextual information is the lack of simple representations of context and efficient algorithms for the extracting such information from the input. One way to define the context of an object is, define in terms of previously recognized objects within the scene. The drawback of this method is that it renders the complexity of context analysis. An alternative view of context is, use the entire scene information holistically. Another drawback is Color is not taken into account in this study

Xiaoqiang Lu [14] proposed a Semi-Supervised Multitask Learning for Scene Recognition Semantic gap between low features and human semantics is a major concern. Many scene recognition techniques try to build an intermediate semantic representation to reduce semantic gap. The main limitation for these methods is improvement in accuracy in scene recognition. The sources of this limitation are the following. Most scene recognition methods focus on extracting low level features from single resolution image. It cannot well represent the entire scene completely. Some redundant features may reduce the accuracy of scene recognition. One of the critical problems is that, whether the features are useful to recognize the semantic category of an input image. It is important to exploit the dense feature that is extracted from different resolution of the given input images in order to improve the performance of scene recognition. This is because dense feature extractions work better than interest point feature extractions. So a new idea is developed to overcome the above mentioned limitations.

First limitation is addressed by introducing a multitask model to integrate scene images of different resolution of the input image. The commonness and the differences among samples from input image is captured using a multitask technique and sparse feature selection and manifold regularization (SFSMR) technique. To address the second limitation, only reliable features are selected and the unreliable features are ignored by borrowing the knowledge from some other resolutions. It will preserve the underlying manifold structure of data. The proposed method can be designed by integrating the multitask model and SFSMR. It helps to overcome the limitations and to improve the accuracy of scene recognition. Advantages of proposed method are; it takes multiresolution images of given input image as multiple related tasks. Then use their common knowledge to learn the different tasks simultaneously. Second, the underlying manifold structure of each feature data is conserved and the optimal features are selected. Third, $l2,1$ norm term and the trace norm term helps to exploit the correlation of different features at multiresolutions.

III. SEMI SUPERVISED MULTITASK LEARNING FOR SCENE RECOGNITION

The proposed method mainly consists of two techniques.

- Multitask model
- sparse feature selection and manifold regularization

Multitask model:
A new model called multitask model is proposed to integrate scene images of different resolution of the input image and share the information and structure of different resolutions image in scene recognition. That is constructing multiresolution images and extracts their features. These features from resolution are considered as one task. To improve the accuracy in scene recognition, these different tasks are simultaneously learned in a joint framework. It is necessary to consider the manifold structure of the training data in this multitask model, since can result in a more accurate result. An iterative method is followed to optimize the object function.

Sparse feature selection and manifold regularization:
The commonness and the differences among samples from input image is captured using a multitask technique and sparse feature selection and manifold regularization (SFSMR) technique. The proposed semi supervised learning method improved the accuracy of scene recognition by integrating the multitask model and SFSMR.

Proposed method includes the following modules. First, different resolution images of input image are generated by downsampling the image. Second, SIFT features are densely sampled from different resolution images of the given input image. Third, a semi supervised learning method is proposed for incorporating multitask model. SFSMR model is introduced for selecting the optimal information, along with preserving the underlying manifold structure of data. Finally, Support Vector Machine classifier is used to perform Multiclass classification

ISSN: 2278 – 7798

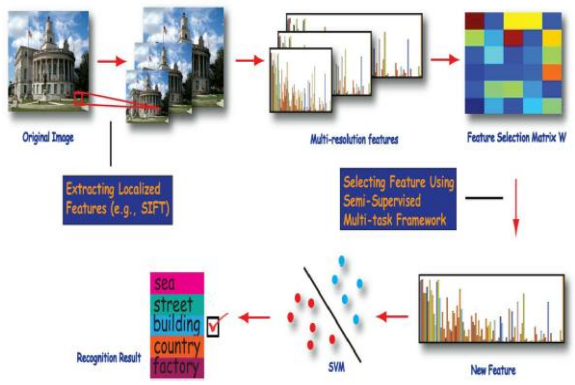

Fig. 1. Flowchart of proposed method. A process takes us from the SIFT feature to the semantic category.

IV. Conclusion

In this paper various scene recognition techniques are discussed and among them Xiaoqiang Lu [14] method which integrate the multitask model and SFSMR performs well for identifying the semantic category of input scene. A multitask model to integrate scene images of different resolution of the input image. The commonness and the differences among samples from input image is captured using a multitask technique and sparse feature selection and manifold regularization (SFSMR) technique. The proposed semi supervised learning method improved the accuracy of scene recognition by integrating the multitask model and SFSMR accuracy of scene recognition is increased.

Acknowledgment

I would like to thank my project guide Prof. Anu K.S and Head of the Department Prof. Mary Linda P.A for their guidance and support and also grateful to all the staff members of the Department of Information Technology and Computer Science & Engineering of KMCT College of Engineering and Technology, Calicut for providing all the important facilities like internet access and books, which were essential to carry out the survey.

References

[1] Zhao Lu ; Coll. Of Metropolitan Transp., Beijing Univ. Of Technol., Beijing, China ; Yanfeng Sun ; Yongli Hu ; Baocai Yin "Robust Face Recognition Based L21-Norm Sparse Representation" Digital Home (ICDH),2014 5th International Conference On Digital Home Nov 2014

[2] Yuanyuan Liu ; Key Lab. of Intell. Perception & Image Understanding of Minist. of Educ. of China, Xidian Univ., Xi'an, China ; Fanhua Shang ; Licheng Jiao ; Cheng, J. more authors "Trace Norm Regularized CANDECOMP/PARAFAC Decomposition With Missing Data" The scope of the IEEE Transactions on Cybernetics includes computational approaches to the field of cybernetics Volume:45

[3] Chang Cheng, Andreas Koschan, Member, IEEE, Chung-Hao Chen, David L. Page, and Mongi A. Abidi "Outdoor Scene Image Segmentation Based on Background Recognition and Perceptual Organization" IEEE TRANSACTIONS ON IMAGE PROCESSING, VOL. 21, NO. 3, MARCH 2012

[4] Yanfei Zhong, Senior Member, IEEE, Qiqi Zhu, Student Member, IEEE, and Liangpei Zhang, Senior Member, IEEE "Scene Classification Based on the Multifeature Fusion Probabilistic Topic Model High Spatial Resolution Remote Sensing Imagery" IEEE TRANSACTIONS ON GEOSCIENCE AND REMOTE SENSING, VOL. 53, NO. 11, NOVEMBER 2015

[5] Yongzhen Huang,Nat Lab of Pattern Recognition, Chinese Acad of Sci Beijing ; Kaiqi Huang LiangshengWang DachengTao "Enhanced biologically inspired model" Computer Vision and Pattern Recognition, 2008. CVPR 2008. IEEE Conference june 2008

[6] Lin, Yingqiang ; Center for Res. in Intelligent Syst., Univ. of California, Riverside, CA, USA ; Bhanu, B." Object detection via feature synthesis using MDL-based genetic programming" Systems, Man, and Cybernetics, Part B: Cybernetics, IEEE Transactions on Systems, Man, and Cybernetics VOL:35 ,

[7] Annabosch,&andrewzisserma"scene classification using a hybrid generative approach" ieee transactions on pattern analysis and machine intelligence, vol. 30, no. 4, april 2008

[8] Jianxin Wu, and James M. Rehg, Member, IEEE "CENTRIST: A Visual Descriptor for Scene Categorization" J. Wu and J.M. Rehg are with the Georgia Institute of Technology.

[9] Li Fei-Fei California Institute of Technology Electrical Engineering Dept. Pietro Perona California Institute of Technology Electrical Engineering Dept "A Bayesian Hierarchical Model for Learning Natural Scene Categories" Pasadena, CA 91125, USA

[10] Jian Yao TTI Chicago, Sanja Fidler University of Toronto, Raquel Urtasun TTI Chicago "Describing the Scene as a Whole: Joint Object Detection, Scene Classification and Semantic Segmentation"

[11] Xiaodong Yu_, Cornelia Ferm¨ullery, Ching Lik Teoz, Yezhou Yangz, Yiannis Aloimonosz "Active Scene Recognition with Vision and Language" Computer Vision Lab, University of Maryland, College Park, MD 20742, USA

[12] Lorenzo Torresani Darthmouth College, Hanover, NH, USA, Martin Szummer Microsoft Research, Cambridge, United Kingdom, and Andrew Fitzgibbon Microsoft Research, Cambridge, United Kingdom "Efficient Object Category Recognition Using Classemes"

[13] ANTONIO TORRALBA Artificial Intelligence Laboratory, Massachusetts Institute of Technology, Cambridge, MA 02139, USA "Contextual Priming for Object Detection" Received August 15, 2001; Revised October 21, 2002; Accepted January 15, 2003

[14] Xiaoqiang Lu, Xuelong Li, Fellow, IEEE, and Lichao Mou "Semi-Supervised Multitask Learning for Scene Recognition" IEEE TRANSACTIONS ON CYBERNETICS, VOL. 45, NO. 9, SEPTEMBER 2015

Anu E is pursuing her M.Tech degree in Computer Science and Engineering from KMCT College of Engineering, Calicut University. She obtained her B.Tech Degree in Computer Science and Engineering from Amrita School of Engineering, in 2011.

Anu K.S. is Assistant Professor, Department of Information Technology, KMCT College of Engineering, Calicut University. Her research focuses on image processing and data mining. She obtained AMEI in Computer Science & Engineering from IEI in 2007. She completed her M.Tech degree in Computer Science & Engineering from NMAMIT college, Nitte in 2010.

ISSN: 2278 – 7798