A Method for Categorization of Scene through Subscene Recognition using Prototypes

Ashwini Fulkar
ME IInd year
PG dept of Computer Science
SGBAU Amravati

V. M. Thakre, PhD
HOD
PG dept of Computer Science
SGBAU Amravati

ABSTRACT

The conventional scene categorization methods ignore spatial information within a scene and are not able to discern categories that share similar subscenes but different in layout; or categories that are ambiguous by nature. To address this issue, in this paper a method is proposed to incorporate subscene attributes within global descriptions to improve categorization performance, especially in ambiguity cases. This is done by encoding subscenes with layout prototypes that capture the geometric essence of scenes more accurately and flexibly. The proposed method improves categorization accuracy. the proposed method can detect and evaluate ambiguity images more accurately.In this paper, scene categorization method is proposed by including subscene attributes to global descriptors. The use of prototypes is proposed to model the geometric configuration of subscenes. These prototypes are more accurate at capturing the layout and simple in training compared to the shape element-based approaches. Incorporating subscene descriptors can enhance the scene categorization result. Having been capable to detect ambiguity, the proposed method offers better understanding of the scene.

General Terms

Image processing, object recognition

Keywords

Scene recognition, subscene categorization

1. INTRODUCTION

The common trend in scene categorization is to adopt a holistic approach, which samples low-level features and then get the global statistical distribution to represent the image. Based on it, a discriminative learning technique is used to infer the global category directly. Conventional scene categorization methods tend to generalize representation of the scene via a holistic approach to calculate a distribution of visual words observed in the image. Scene categorization aims at labeling scenes according to known categories. When applied to background extracted from an image or video, it helps to discover the context from which the foreground objects or events or activities are detected. In fact, the knowledge of scene category can provide contextual cues for many vision tasks, such as semantic labelling, event detection, visual surveillance and image retrieval. Bag-of-words (BOW) scheme densely sample SIFT feature from evenly divided patches of a scene image and learn the visual words in the codebook by performing *k*-means clustering. From the visual words, a distribution histogram is generated to represent the scene. Although the BOW approach is simple and efficient, it completely disregards any spatial or contextual information inherent in the scene image. The discriminative power among scene categories is thus abated.

In this paper a method is proposed to incorporate subscene attributes within global descriptions to improve categorization performance, especially in ambiguity cases. This is done by encoding subscenes with layout prototypes that capture the geometric essence of scenes more accurately. The proposed method improves scene categorization accuracy. The proposed method can detect and evaluate ambiguity images more accurately.

2. BACKGROUND

Object recognition using computer vision methods has gone through considerable progress during the last decade. These include the methods based on low level features e.g., Scale-Invariant Feature Transform (SIFT), Speeded Up Robust Feature (SURF), pyramid Histogram of Oriented Gradients (pHOG), and Self Similarity, and specific designed machine learning techniques. In one-shot recognition only one training sample is available for each object category in the target test domain, with the help of apriori knowledge data from the source domain. A novel unsupervised hierarchical feature learning framework is used to learn a feature pyramid from the prior-knowledge domain[1]. A novel work in 3d scene understanding using structured learning is used to simultaneously reason about many aspects of scenes[2]. A novel method based on Spectral Regression (SR) for efficient scene recognition is also used [3]. A new SR approach, called Extended Spectral Regression (ESR), is proposed to perform various learning on a huge number of data samples. An efficient Bag-of-Words (BOW) based method is developed which employs ESR to summarize local visual features with the semantic, spatial, scale, and orientation information of data samples for scene recognition. A method is presented to improve the flexibility of descriptor matching for image recognition by using local multiresolution pyramids in feature space [4]. Image patches are represented at multiple levels of descriptor detail and that these levels are defined in terms of local spatial pooling resolution.

3. PREVIOUS WORK DONE

The conventional object recognition systems require a large number of labelled training images. In contrast human cognitive systems could perform recognition tasks well provided only one or a few labelled training samples. One-shot recognition is a visual classification task. In one-shot recognition only one training sample is available for each object category in the target test domain, with the help of apriori knowledge data from the source domain. In Zhenyu Guo *et al.* solved the one-shot recognition problem under a more exciting setting[1]. Only unlabeled images are used as prior knowledge, which requires less labelling effort. A novel unsupervised hierarchical feature learning framework is proposed to learn a feature pyramid from the prior-knowledge domain. This feature learning method could also be applied across multiple feature spaces.

A long standing goal of computer vision is to build a system that can automatically understand a 3D scene from a single image. This requires extracting semantic concepts and 3D information from 2D images which can depict an enormous variety of environments that comprise our visual world. Jianxiong Xiao *et al.* described the richly annotated SUN database which is a collection of annotated images spanning 908 different scene categories with object, attribute, and geometric labels form any scenes [2]. This database allows user to systematically study the space of scenes and to establish a benchmark for scene and object recognition. An integrated system to extract the 3D structure of the scene and objects depicted in an image is presented.

Liyuan Li *et al.* proposed a novel method based on Spectral Regression (SR) for efficient scene recognition [3]. First, to perform manifold learning on a huge number of data samples a new SR approach, called Extended Spectral Regression (ESR) is proposed. Then, an efficient Bag-of-Words (BOW) based method is developed which employs ESR to encapsulate local visual features with their semantic, spatial, scale, and orientation information for scene recognition. In many applications, such as image classification and multimedia analysis, there are a huge number of low-level feature samples in a training set. In the ESR-based scene recognition, an enhanced low level feature representation is proposed which combines the scale, orientation, spatial position, and local appearance of a local feature. Then, ESR is applied to insert enhanced low-level image features. The ESR-based feature embedding generates a low dimension feature representation. Also it integrates various aspects of low-level features into the compact representation.

Lorenzo Seidenari *et al.* presented a novel method to improve the flexibility of descriptor matching for image recognition by using local multiresolution pyramids in feature space [4]. Image patches are represented at multiple levels of descriptor detail and that these levels are defined in terms of local spatial pooling resolution. Preserving multiple levels of detail in local descriptors is useful for prevarication on which levels will most relevant for matching during learning and recognition. The Pyramid SIFT (P-SIFT) descriptor introduced. Its use in four state-of-the-art image recognition pipelines improves accuracy and yields state-of-the-art results. This technique is applicable independently of spatial pyramid matching. Spatial pyramids can be combined with local pyramids to obtain more improvement. This technique is efficient and is extremely easy to integrate into image recognition pipelines.

Traditional scene categorization methods tend to generalize representation of the scene via a holistic approach to calculate a distribution of visual words observed in the image. They disregard spatial information within a scene and are not able to discern categories that share similar subscenes but different in layout; or categories that are ambiguous by nature. Shan-shan Zhu *et al.* addressed this issue. Subscene attributes within global descriptions are used to improve categorization performance, especially in ambiguity cases [5]. This is done by encoding subscenes with layout prototypes that capture the geometric essence of scenes more accurately and flexibly. It is also observed that the proposed method is more accurate at detecting and evaluating ambiguity images.

Minguang Song and Ping Guo proposed Spatial Image Representation by Combining Local Difference Binary Pattern (LDBP) with Local Neighbour Binary Pattern (LNBP) [6]. LDBP is based on the comparisons between center pixel and its neighbouring pixels, but the relationship among neighbour pixels is not considered. LNBP is combined with LDBP to construct a spatial representation for scene recognition, because that LNBP can provide complementary information regarding neighbouring pixels for LDBP.

4. EXISTING METHODOLOGY

A subscene should be recognized by its label as well as its spatial layout. The spatial layout should be approximated more accurate and flexible yet simpler in training than the shape element-based approaches. A subscenes capture local attributes efficiently and by combining subscenes with global-based feature distribution will improve the scene categorization performance. In the existing methodology to combine subscenes attributes with global features in scene categorization is proposed [5]. Prototypes are generated directly from subscenes to approximate the spatial configurations more accurately. The prototypes are proved to be highly representative and yet flexible. Subscenes encoded by prototypes combined with the global feature vector from a BOW approach are jointly trained for the scene categorization. The proposed method has been tested on a widely used scene categorization dataset and compared with the state-of-the-art methods. the result of this method confirms that combining the subscenes can enhance scene categorization performance. Also it helps to eliminate ambiguity in scene images by identifying the spatial layout structure of subscenes.

5. ANALYSIS AND DISCUSSION

In the existing method the Contextual BOW approach is followed to generate visual words to describe the image in a global statistical representation approach. Each image is evenly divided into grids according to five different scales. SIFT and color features are sampled from the grids of each scale. The features are then concatenated with features from its neighboring patches and from coarser scale to build the contextual feature. In training, the contextual features are clustered to construct the codebook. Using the codebook, all the features are represented as visual words. Then, all the visual words form the histogram on the codebook to represent the image globally. Evaluating features using the prototypes can better adapt to the layout changes. The subscene predictions correctly label each subscene. Combining the encoded subscenes successfully categorizes them correctly while solely based on global based approach the predictions are all wrong. By recognizing distinguishable subscenes and studying their layout relationships, this method is more robust to differentiate these less typical scenes. The other group of scenes that can be improved is the ambiguity scene. Ambiguity scenes are those that contain more than one category in one image. In this case, the subscene that dominates the scene is crucial to the decision.

6. PROPOSED METHODOLOGY

In the proposed method, the image is denoted by *I*. The distribution π_1 on the codebook is derived from the visual words captured from the image. *N* instances of subscenes **s** = $\{s_1, s_2, . . . , s_N\}$ labeled by semantic classes are assumed, and there are *M* trained subscene prototypes **Pt** = $\{Pt_1, Pt_2, . . . , Pt_M\}$. These prototypes are used to capture the spatial layout of subscenes. Then the subscenes are encoded by the group of prototypes and produces representation π_2. Combining π_1 and π_2 together, the category of the image is predicted. The category of image is denoted by $c = \{1, . . . ,C\}$, where *C* is the total number of categories.

Understanding the category *c* based on the features from the image *I* and the group of trained prototypes **Pt** can be formulated as recognition based on the representation π_1 of features from global scale and π_2 from subscene scale:

$$\hat{c} = \text{argmax}\ p(c|\pi_1, \pi_2) \qquad (1)$$

As the prior distribution *p(c)* is always assumed to be as a uniform distribution, the prior probability of category takes a fixed value of 1/*C*. Thus, the classification problem is to solve:

$$\hat{c} = \text{arg max}\ p(\pi_1, \pi_2|c) \qquad (2)$$

The problem of inferring category is modelled as a SVM problem. The features π_1 and π_2 are concatenated together to represent the image, *c* is the target category. The *C*-class classification problem is solved by a set of 2-class classification problems using a one-against-all strategy.

6.1 Generation of Features of Subscenes

Features of visual words and subscenes are generated from a given image from two independent approaches. The visual words are generated by a global-based scene categorization approach. Each image is evenly divided into patches, or a pyramid of patches at different spatial resolutions and features are sampled from each grid. After the features are obtained, a codebook is formed by clustering the features into visual words. Then, all features are represented by the visual words from the codebook. Then, the distribution histogram π_1 on the codebook is built to represent the image.

A layout prototype is a mask on the image indicating where a subscene is. The proposed prototypes of each semantic class are generated in training and are used jointly in testing. The prototypes are generated from bottom-up. For each semantic label, any training image contained the subscene s which takes the label is considered as an instance, and use the mask image of them as prototypes to form the leaves of the prototype tree. Every two leaves are evaluated and the most similar pairs are merged into a new node in the upper layer of the tree; the generated nodes continue merging in the same manner, and the tree is built up until no more pairs are to be merged to generate a new layer. The similarity score is within 0 and 1. When two prototypes have comparable features at the same spot its value is near to 1.

Some prototypes can be pruned because they are less representative at each layer of the layout prototype tree. Instances are used to evaluate at each layer, the most representative prototypes is found from this layer. The statistical results of the prototypes are calculated. The prototypes which jointly cover over the majority number of the instances from training set are kept, and the other prototypes are pruned from this layer. Prototypes are selected from one or more layers and are added to the group of subscene prototypes.

It can be seen that the subscene layout prototypes are highly representative and they capture the variation within each class. For example, the different spatial layouts of buildings are in response to different viewpoint changes. The probabilistic representation enables the prototype to be more accurate and flexible than rigid shape-based approaches. However, the subscene with few instances is less accurate due to insufficient training, such as the beach prototypes

6.2 Encoding Subscene with Prototypes

After the layout prototypes have been generated, the image can be represented by subscenes with spatial configurations using the prototypes. Subscenes are first generated by semantic labelling. The subscene representation is then encoded with every prototype from the library in the form of a series of histograms. Finally, the distribution histograms of all the prototypes are concatenated to represent the image.

The flowchart of proposed method is shown in figure 1.

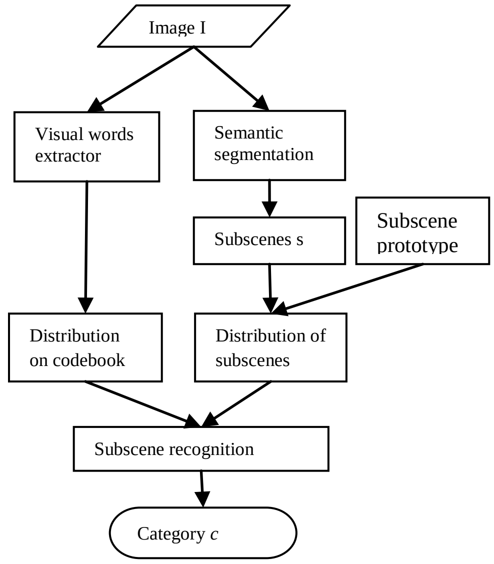

Fig 1: flowchart for the proposed method

7. POSSIBLE OUTCOMES AND RESULTS

The scenes have varied appearances with uncommon viewpoints, which make it difficult to classify them correctly based on rigid global feature statistics. The subscene prototypes are more capable to capture subscenes at different positions. Evaluating features using the prototypes can better adapt to the layout changes. Combining the encoded subscenes successfully categorizes them correctly while solely based on global-based approach the predictions are all wrong. By recognizing distinguishable subscenes and studying their layout relationships, the proposed method is more robust to differentiate these less typical scenes. The other group of scenes that can be improved is the ambiguity scene.

8. CONCLUSION

The subscene predictions correctly label each subscene. Combining the encoded subscenes successfully categorizes them correctly. By recognizing distinguishable subscenes and studying their layout relationships, the proposed method is more robust to differentiate the less typical scenes. The differences in layout of subscenes are useful to better recognize between categories even at global scale they contain similar statistics of features. The proposed methodology improves scene categorization by incorporating subscene attributes to global descriptors. The prototypes are used to form the geometric configuration of subscenes. These prototypes are more accurate at capturing the layout and simple in training compared to the shape element based approaches. Including subscene descriptors can improve the scene categorization result. Subscenes should be also better at detecting ambiguity in scenes. The subscene that dominates the scene is crucial to the decision. From the analysis, it is found that they are more robust to variations in the scene and could evaluate the ambiguity more accurately. Having been able to detect ambiguity, the proposed method generally offers better understanding of the scene. There may be scenes that have not been improved by introducing subscene recognition, or even misclassified.

9. FUTURE SCOPE

The subscene prototypes are more capable to capture subscenes at different positions. Evaluating features using the prototypes can better adapt to the layout changes. So the use of prototypes can be enhanced in future.

10. REFERENCES

[1] Zhenyu Guo and Z. Jane Wang, "An Unsupervised Hierarchical Feature Learning Framework for One-Shot Image Recognition", IEEE TRANSACTIONS ON MULTIMEDIA, VOL. 15, NO. 3,pp. 621-632, APRIL 2013

[2] Jianxiong Xiao, James Hays, Bryan C. Russell, Genevieve Patterson, Krista A. Ehinger, Antonio Torralba and Aude Oliva, "Basic level scene understanding: categories, attributes and structures", Frontiers in Psychology: Perception Science, VOL 4, pp. 1-10, August 2013

[3] Liyuan Li, Weixun Goh, Joo Hwee Lim, Sinno Jialin Pan, "Extended Spectral Regression for efficient scene recognition", Pattern Recognition (Elsevier), Vol 47, Issue no. 9 pp. 2940-2951, March 2014

[4] Lorenzo Seidenari, Giuseppe Serra, Andrew D. Bagdanov, and Alberto Del Bimbo, "Local Pyramidal Descriptors for Image Recognition", IEEE TRANSACTIONS ON PATTERN ANALYSIS AND MACHINE INTELLIGENCE, VOL. 36, NO. 5, pp. 1033- 1040, MAY 2014

[5] Shan-shan Zhu and Nelson H. C. Yung, "Improve scene categorization via subscene recognition", Machine Vision and Applications (Springer), VOL 25, ISSUE. NO. 6, pp. 1561–1572, June 2014.

[6] Minguang Song and Ping Guo, "Combining Local Binary Patterns for Scene Recognition", JOURNAL OF SOFTWARE (ACADEMY PUBLISHER), VOL. 9, NO. 1, pp. 203-210, JANUARY 2014