IEICE TRANS. INF. & SYST., VOL.E96–D, NO.4 APRIL 2013

LETTER

Indoor Scene Classification Based on the Bag-of-Words Model of Local Feature Information Gain

Rong WANG[†,††a)], ***Student Member***, **Zhiliang WANG**[†††], ***and*** **Xirong MA**[††], ***Nonmembers***

SUMMARY For the problem of Indoor Home Scene Classification, this paper proposes the BOW Model of Local Feature Information Gain. The experimental results show that not only the performance is improved but also the computation is reduced. Consequently this method out performs the state-of-the-art approach.

key words: *information gain, Bag-of-Words model, indoor scene classification, visual attention*

1. Introduction

In recent years, the development of home service robot technology is increasingly mature. The visual system of home service robot as important input information has an effect on decision and action performance. So it is becoming the research hotspot that the home service robot could understand the home scene around correctly.

Scene classification, also called scene recognition, is a method that the image semantic, such as mountain, street, beach, is extracted based on the scene features. Because scene classification is decided by image semantic, it is not effective to depend solely on image low-level information. At present, many scene classification researches focus on modeling the scene semantic. There are three methods: Semantic Properties (SP) [1], Semantic Objects (SO) [2], Local Semantic Concepts (LSC) [3]. There are many indoor home scenes, such as living room, bedroom, kitchen, bathroom in the scene classification of home service robot, but the color and texture features of these scenes are not in evidence, so the method using low-level information or semantic properties is not effective [4]. The distribution of objects in indoor scene is complex. The marking work is taxing in the way of Semantic Objects, because it is not automatic. The image of indoor scene is complex. Meanwhile, the difference between classes is small. Therefore, for using local semantic concept to classify the scene, the key points of whole image are extracted. But the computation and its key features is not clear. Therefore, in this paper, based on image semantic concept, the information gain method is applied to BOW (Bag-of-Words). The important visual words of visual word histogram are computed based on information gain and they are enlarged. Thus, in this way, the difference of images which are in the same class is reduced, but it is enhanced in different classes. At the same time, the visual attention model is imported in the part of key point extraction to reduce data size to improve the computing speed. Because most of researches are based on indoor or outdoor scene, there is little database based on indoor scene. In this paper, the experiments use the home indoor scene database with 3200 images of 4 classes established by Ref. [5].

Manuscript received May 30, 2012.
Manuscript revised December 5, 2012.

[†]The author is with School of Automation and Electrical Engineering, University of Science & Technology Beijing, Beijing, 100083, China.
[††]The authors are with School of Computer & Information Engineering, Tianjin Normal University, Tianjin, 300387, China.
[†††]The author is with School of Computer & Communication Engineering, University of Science & Technology Beijing, Beijing, 100083, China.
a) E-mail: wangrong0806@163.com

DOI: 10.1587/transinf.E96.D.984

2. The BOW Model of Local Feature Information Gain

Bag-of-words (BOW) was discovered in the field of text analysis, then, is applied to the field of computer vision [6]. This method depicts images by clustering the visual features of images to classify. In this paper, the BOW model is improved and the BOW model based on information gain of local feature is put forward.

2.1 The Local Feature Extraction of Visual Attention

In the BOW model, because the features come from the whole image, it is difficult to deal with the large amount of data. Meanwhile, the foreground information rather than background information is useful. So, local foreground information as the feature is extracted to build BOW model to classify. The amount of data is reduced with image information maintaining. In this paper, the saliency regions are formed using visual attention model [7], and then in these regions the SIFT [8] features are extracted. And the image could be enlarge and reduce to produce 3 levels of pyramid. Afterward, 12 color feature maps, 6 brightness feature maps and 24 orientation feature maps come into being and normalized to be saliency map. The key points are computed in saliency map. The window with 16 pixels row and 16 pixels column around key point is divided into 16 parts. Every part is an image patch with 4 pixels row and 4 pixels column. Gradient direction histogram of 8 orientations in image patch is computed. So, the descriptor is a vector with 128 dimensions.

2.2 The Feature Selection of Visual Word Based on Information Gain

Information entropy is a parameter to measure uncertain degree of a random variable. Suppose X is a random variable. The more random variable X changes, the greater information quantity is. The information entropy of X is:

$$H(X) = -\sum_i p(x_i)\log_2(p(x_i)) \tag{1}$$

The information entropy about random variable X by observing random variable Y is:

$$H(X|Y) = -\sum_j p(y_i)\sum_i p(x_i|y_j)\log_2(p(x_i|y_j)) \tag{2}$$

Information gain is the difference of information entropy. It shows the information quantity after eliminating the uncertainty.

$$IG(X,Y) = H(X) - H(X|Y) \tag{3}$$

The larger information gain value is, the greater information quantity is.

In this paper, the information gain method is imported into feature selection of visual word. The image number of certain visual word in certain class appeared or not is computed to become information gain of this visual word toward this class. The definition is:

$$\begin{aligned} IG(t,c) \;=\; & p(t_i)p(c_j|t_i)\log_2 p(c_j|t_i) \\ & +p(\bar{t}_i)p(c_j|\bar{t}_i)\log_2 p(c_j|\bar{t}_i) \end{aligned} \tag{4}$$

In formula (4), t_i is the ith visual word, c_j is the jth kind of image. Accordingly, the information gain value of visual word with each class could be computed by visual word histogram of each training image. The visual word with great information gain has great contribution to this class. So, for each class, visual words are arranged according to information gain value from large to small. Then, the important visual words which are in the front are selected.

$$P(vw) = \frac{n(ivw)}{n(vw)} \tag{5}$$

In formula (5), n(ivw) is the munber of important visual words, n(vw) is the number of visual words. According to formula (5), when n(vw) is certain, the value of P(vw) is changes with different n(ivw). Thereby, the value of P(vw) with highest recognition rate is obtained by experiment. The important visual words of each image are enlarged in visual word histogram. The useful visual words are strengthened and useless visual words are weakened.

Finally, SVM classifier is trained making use of new visual word histograms of every image as training samples.

3. Experiments

At present, any image database in scene classification experiment always includes indoor images and outdoor images. It does not fit to this paper. Thus, Indoor Home Scene Database with four classes (bathroom, bedroom, kitchen, living room) in Ref. [5] is adopted by this paper. There are three parts in this database: (1) The image database in Ref. [9]. (2) Google search tool. (3) The images are taken using camera, as Fig. 1 shows. Then, after arranging, 1600 images are formed, 400 images for each class. The left part and right part of image could be exchanged, but upper part and lower part of image could not be exchanged, so flipping image horizontally does not affect result. Finally, the number of images is 3200, 4 classes, 800 images for each class. All images have a minimum resolution of 200 pixels in the smallest axis and a maximum resolution of 1000 pixels in the largest axis. At the same time, there are both colors images and gray images in this database.

In this experiment, firstly, the color images are transformed to gray image, then, the saliency regions are extracted as the Fig. 2 shows. After that, the SIFT features of saliency regions are produced. In the database of 3200 images, 2400 images are selected as training images, including 600 images for each class, and the last including 200 images for each class are the testing images. In training, using 800 images including 200 images for each class, the visual dictionary is created. And the number of the visual words in visual dictionary has a strong impact on the classification performance. If the number of the visual words is small, different features are able to be described as the same visual word, so that the definition of every kind of scene could be confused. Moreover, if the number of the visual words is too large, each feature is able to become the independent visual word to produce the information redundancy. Meanwhile, it will cause dimension disaster and reduce the computing speed. In this paper, because the SIFT features are extracted in saliency regions, the number of features is decreased greatly, and the number of visual words in visual dictionary is decreased correspondingly. After several experiments, as Fig. 3 shows, the result is that when the num-

Fig. 1 The sample images of indoor home scene database.

Fig. 2 The saliency regions of images.

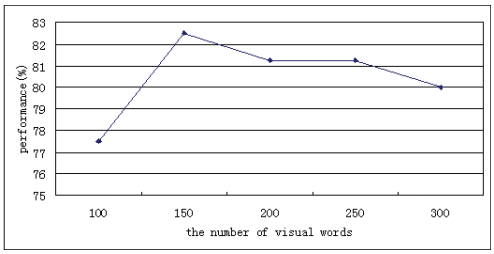

Fig. 3 The performance of difference number of visual words.

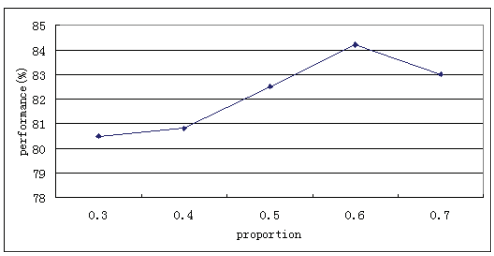

Fig. 4 The performance of difference proportion of important visual words.

ber of visual words in visual dictionary is 150, the recognition rate is the highest, which is 82.5%. When the number of visual words in visual dictionary is more than 150, the recognition rate is dropping.

The visual word histograms of training images are produced based on visual dictionary with the size of 150. According to the part of 2.2 in this paper, the important visual words of each image are selected based on the information gain. Because the number of visual words has a great effect on the classification result, useful visual words will be lost if it is too small and less important words will be enlarged if it is too large. So it is not good for classification. The experiment result of the proportion of important visual word is shown as Fig. 4. When the value is 0.6, that is the number important of visual word is 90, the recognition rate is the highest.

In order to strengthen the difference of every class, the important visual words of each image in visual word histogram are enlarged. The Table 1 shows results to be enlarged to differnet extents. When it is 2 times enhanced, the performance is 84.23 which is the best. Therefore, in this paper, the important visual words of each image are enlarged 2 times.

The Table 2 shows experiment result of AdaBoost classifier [10] and SVM classifier with the histogram proposed by this paper. There are 3 different AdaBoost schemes: Real AdaBoost (RAB), Gentle AdaBoost (GAB) andModest AdaBoost (MAB). And weak learner is constructed to Classification and Regression Trees (CART). When there are 200 iterations, the best result of AdaBoost classifier is 82.75% which belongs to MAB. Moreover, the result of SVM classifier is 84.23%. So, in this paper, the SVM classifier is chosen.

The Table 3 shows the result of method in this paper compared with BOW model. The global SIFT features of 2400 images are extracted, and the number of these is 6811944. Using the local saliency regions model proposed by this paper, the number of SIFT features is 1249593. So the data size is reduced to 18.34%. Meanwhile, using the visual word selection based on information gain, the recognition rate of each class is advanced, especially, the rate of the bathroom class and the kitchen class, of which the visual word features are not clear, are increased.

Table 1 The result of the times of important visual words englaged.

	1 times	2 times	3 times	4 times	5 times
performance (%)	83.4	84.23	83.4	73.4	73.4

Table 2 The performance of SVM classifier compared with AdaBoost classifier.

	GAB	RAB	MAB	SVM
200 interations (%)	74.38	79.81	80.78	84.23
100 interations (%)	75.57	79.05	82.75	

Table 3 The performance and Data size of this paper method compared with BOW model.

	The BOW Model	The method of this paper
bathroom (%)	75.4	83.2
bedroom (%)	80.6	82
kitchen (%)	73.2	86.3
living room (%)	83.5	85.4
Datasize	6811944	1249593
Datasize (%)	100	18.34

4. Conclusions

In this paper, for indoor home scene recognition of home service robot, the BOW model based on local feature information gain is proposed. Using information gain method, visual words in visual dictionary are selected, and the visual words with larger information are strengthened to enlarge the difference between classes and become easy to distinguish. At the same time, saliency regions are obtained by visual attention model, and in these regions, SIFT features are extracted. The computation is reduced. Making use of the database with 3200 images, 4 classes, the experiment of the method of this paper is compared to BOW model. The performance is improved.

Acknowledgments

This work was supported by the Major Special Project of the National Science and Technology of China (No. 2010ZX07102-006), the National Basic Research Program of China (973 Program) (No. 2011CB505402), the National Natural Science Foundation of China (No. 61170117, No. 60970060/F020508, No. 61103074), the Provincial Project of Production Teaching Research Integration (No. 2011A090200008), the Science and technology support plan key project (No. 09ZCKFGX00500), the Natural Science Foundation of Tianjin (No. 11JCYBJC00600),

the Youth Foundation of Tianjin Normal University (No. 52WU1106).

References

[1] A. Oliva and A. Torralba, "Modeling the shape of the scene: A holistic representation of the spatial envelope," Int. J. Comput. Vis., vol.42, no.3, pp.145–175, 2001.
[2] J. Luo, A.E. Savakis, and A. Singhal, "A Bayesian network-based framework for semantic image understanding," Pattern Recognit., vol.38, pp.919–934, 2005.
[3] A. Boseh, A. Zisserman, and X. Munoz, "Scene classification via PLSA," European Conference on Computer Vision, pp.517–530, April 2006.
[4] A. Quattoni and A. Torralba, "Recognizing indoor scenes," IEEE Conference on Computer Vision and Pattern Recognition, pp.413–420, 2009.
[5] Z.L. Wang, R. Wang, and X.R. Ma, "Indoor scene recognition based on the weighting spatial information fusion," 2012 International Conference on Intelligent Systems Design and Engineering Applications, pp.1040–1044, 2012.
[6] J. Yang, Y.G. Jiang, A.G. Hauptmann, and C.W. Ngo, "Evaluating bag-of-visual-words representations in scene classification," Proc. International Workshop on Multimedia Information Retrieval, pp.197–206, 2007.
[7] L. Itti and C. Koch, "Computational modeling of visual attention," Nature Reviews Neuroscience, vol.2, no.3, pp.194–203, 2001.
[8] D.G. Lowe, "Image features from scale-invariant keypoints," Int. J. Comput. Vis., vol.60, no.2, pp.91–110, 2004.
[9] S. Lazebnik, C. Schmid, and J. Ponce, "Beyond bags of features: Spatial pyramid matching for recognizing natural scene categories," Proc. IEEE Computer Society Conference on Computer Vision and Pattern Recognition, pp.2169–2178, 2006.
[10] S. Aoyagi, A. Kohama, Y. Inaura, M. Suzuki, and T. Takahashi, "Image-searching for office equipment using bag-of-keypoints and AdaBoost," J. Robotics and Mechatronics, vol.23, no.6, pp.1080–1090, 2011.