Similar Scene Classification Research Based on Dense Matching

Han Chao [1, a *], Hou Jianjun [1,b] and Xu Lingqing [1,c] and Bai Shuang [1,d]

[1] School of Electronic and Information Engineering, Beijing Jiaotong University, Beijing, China

[a]13120010@bjtu.edu.cn, [b]jjhou@bjtu.edu.cn, [c]12213055@bjtu.edu.cn, [d]shuangb@bjtu.edu.cn

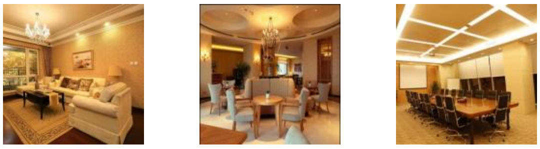
Keywords: Image representation; SIFT-Flow; displacement vector map; SVM; Scene classification.

Abstract. Scene classification is one of the important topics of computer vision, and the classification of similar scenes is even more challenging. This paper proposes a new method for image representation suitable for such a task. First, a displacement vector map of an input scene image can be obtained by utilizing SIFT-Flow. Then, the map is segmented into spatial blocks, so that information about the matching result can be used for creating a representation for the image. Finally, scene images can be classified by Supported Vector Machine (SVM). The proposed method outperforms state-of-the-art approaches for classifying similar scenes.

Introduction

Recently, along with the continuous improvement of computer technology and the industrial level, the intelligent technology has been developing rapidly. At present, scene classification is one of the most active research directions in the field of machine vision. Scene classification mainly focused on cognition and analysis of the scene as a whole. At the moment, there are many difficulties in these researches including the similar scene classification which is a major difficulty among scene classification.

A particular scene can generally contain several categories of objects. When there is great difference between scene images, a specific object will play a key role for scene understanding. During scene analysis, we can use key object detection information to assist the process of scene recognition. As shown in Fig. 1, office, cafe and conference room are three different scene categories, the types of tables and chairs in these three scenes are different. To identify a particular object is a kind of scene image classification methods. During the scene image classification, you first need to extract the image descriptors. Scene Images include not only the low-level features information such as color, texture and shape, but also the middle-level information which is based on global information, as well as the high-level information like semantic information. Representative operators of low-level features have HOG[1] and SIFT[2], representative operators of low-level features have bag-of-words (BoW)[3][4] and Spatial Pyramid Matching (SPM)[5].



Office Cafe Meeting room

Figure 1. Three similar scene images

At present, many researchers have conducted a lot of researches on scene classification which are based on different levels of features. For example, Wang proposed an improved bag-of-words model for scene classification[6], Zhang presented a classification algorithm which is based on multi-scale context semantic information of the scene image[7]. But in some similar scenes, such as restaurants and fast-food restaurants, the space layouts of the scene are similar. By contrast, the two scenes are lack of obvious typical features; even the human eyes can hardly recognize these scenes. In addition, the intra-class differences are unstable and the position that the representative items may appear can also be uncertain. The bag-of-words model would ignore some local information in the image, thus affecting the similar scene recognition.

For intra-class fuzzy within the scene images, we propose a new image representation method. For each scene category, we extract the representative picture as standard image thus to constitute a standard image group, any input image need to match with this group according to SIFT-Flow algorithm, after that, through computing the space match statistics of each descriptor, the whole match information can be eventually captured. Given a query image, compute its match information with the standard image group, and using support vector machine to predict the category of the query image.

A. SIFT feature extraction

SIFT is a local feature extraction operator, it's an operator that can not only remain invariant to rotation, scale and brightness changes, but also can maintain some stability and noise to the change of perspective and the affine transformation[2]. So we finally adopt SIFT descriptors for image matching.

B. Standard image group building

The classification problem of a query image can be treated as a similarity issue of a query scene with all kinds of scenes. This paper presents a new approach to image representation that the image features can be captured by matching the input image with a variety of standard scene images which have been constructed manually. When choosing the standard image from a set of images of a single class, we can use the Bag of Words[8] to get each image's histogram representation, after converting it to a vector, the distance between the image with the rest of the images can be calculated, if the distances between an image and other images are minimal, selecting this image as the standard image of the scene class.

In Bag of Words, we construct the word bags through K-means[9] clustering algorithm. The building process of standard image group has been shown in Fig. 2, in which each row represents a kind of scene. As shown in the figure, even within the same scene, the containing objects of different scenes and the space layouts are not the same. In this paper, we propose to get the most representative images of the scene by adopting the Bag of Words algorithm. As can be seen in the right of Fig. 2, standard images can be chosen through K-means.

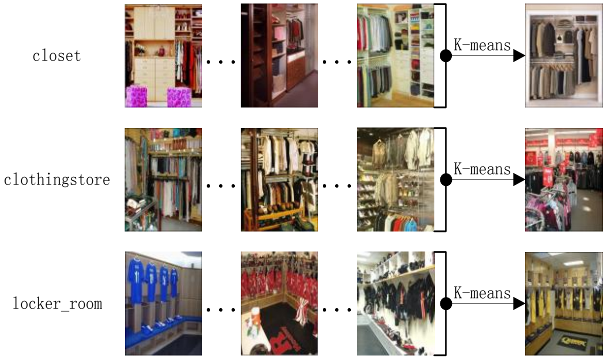


Figure 2. Standard image group building

C. SIFT-Flow dense matching

Through matching the input image with the standard image, can we obtain feature matching information of each SIFT point. SIFT-Flow feature matching is an algorithm which can still save space characteristics of images during matching[10]. After SIFT features extraction by dense-sampling, the displacement vector between each feature point in the input image and the corresponding points in standard images can be obtained through using SIFT-Flow algorithm. SIFT-Flow algorithm is shown in Eq. 1. E(w) is the energy function, let p = (x, y) be the coordinates of the image points, w(p) = (μ(p), ν(p)) be the displacement vector of point p, allow μ(p) and ν(p) be the predicted values, and let s_1, s_2 be two SIFT images that we want to match. $||\{.\}||_1$ denotes the Manhattan distance, t and d are the threshold values. SIFT-Flow adopts the pyramid principle, through using the method of Gaussian filtering, to construct image pyramids. Starting from the top of the pyramid, each layer uses the Eq. 1 to get the displacement vector of all feature points and as the initial input value of the next layer; finally, the resulting displacement vector between the input image and standard image can be gotten.

$$\begin{aligned} E(w) = {} & \sum\nolimits_p \min(||s_1(p) - s_2(p + w(p))||_1, t) + \\ & \sum\nolimits_p \eta(|\mu(p)| + |\nu(p)|) + \\ & \sum\nolimits_{(p,q)\epsilon\varepsilon} \min(\alpha|\mu(p) - \mu(q)|, d) + \min(\alpha|\nu(p) - \nu(q)|, d) \end{aligned} \quad (1)$$

E(w) contains a data term, small displacement term and smoothness term. These three terms mentioned above can be added up to the energy E(w). So the optimal solution of Eq. 1 is the ideal matching displacement vector we need.

Prepare image representation based on dense matching approach

A. Image Blocking

Matching information between input image and standard image is spatial information. If let the displacement matrix directly be the feature of the input image, the dimension of the feature vector will reach thousands of dimensions that will cause too much noise. Therefore the displacement matrix cannot be directly used, and displacement data needs to be dealt with later.

In terms of matching stability, we adopt blocking for image segmentation. In order to avoid the introduction of more noise, we adopt uniform blocking[11].

B. Scene image representation

After image segmenting, an image can be represented by the basic block unit. A query image needs to match with each standard image of all scenes, after getting the matching displacement matrix, displacement of the characteristic points need to be converted to matching values of feature points in the standard image. Each element in the matrix indicates the distance between that point and its matching point, therefore, we can then get all the coordinates of the matching points in standard images through displacement matrix. And through the statistics of the number of matching feature points, can we obtain the matching map of the input image. The matching process is shown in Fig. 3. Input image is a closet image and it need to match with standard images including closet, clothes shops and locker room. Displacement matrix can be gotten through SIFT-flow, and then by converting the match results to the matching map of the input image, In the map, the gray value of the square indicates the matching degree of block in the standard image. As shown in Fig. 3, there are three match maps in total, in the first figure, the gray value of all local blocks are close in some ways, which indicates the input image is matched to standard image of closet. In the second and the third figure, the top and the bottom are much darker than the middle part, which means the clothing part is relatively close to the same part in the standard image of clothing shop and locker room. From the perspective of matching degree, the query image is more likely to belong to the closet.

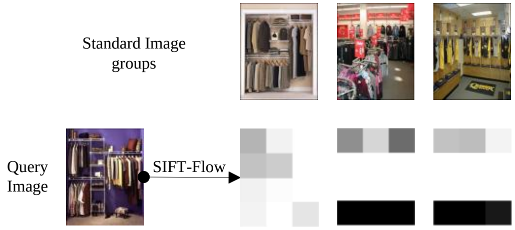


Figure3. Input-intensive image matching

Experiment and result

A. Dataset

Experiments are conducted on the 67 Indoor Scene[12] database. 67 Indoor Scene contains 67 indoor scenes, which comprises many similar scene combinations. As shown in Fig. 4, the Concert Hall, Auditorium and the movie theater, in which the space layouts and the styles of seats are almost the same, can constitute a set of similar scenes. Each category contains at least 100 images. In this paper, we choose similar scenes to constitute three similar scene groups. For each class, we use 80 images to train and 20 images for testing.

Concert Hall

Auditorium

Movie Theater

Figure4. Similar combinations

B. Experimental methods

The size of the image used for the experiments in this paper is 300×240, and the Bag of words algorithm is selected to construct standard image groups. Parameter setting includes SIFT features sampling step and the segmenting ways. In addition, we select the RBF kernel for support vector machine, which can be obtained by LIBSVM toolbox[13].

a) Results on 67 Indoor Scenes

There are three experimental groups of similar scenes, the first group contains a bedroom, children room and prison cell, the second group includes the closet, clothing store, and locker room, and the third group consists of buffet, dining room, fastfood restaurant, meeting room and restaurant. We have proved that when 4×3 is the mode of image segmentation, the classification result can achieve the best average accuracy rate.

The extraction of SIFT features requires dense-sampling for input image. Sampling steps have a several options. Table 1 shows the classification results under the different sampling steps. As we can seen in table 1, when the sampling steps are 2 grids, category rate can achieve an average of 85%, and the extracted features, which can keep the enough information of input images, can also eliminate the interference of individual features effectively. When the sampling step is 1 grid, the average accuracy can be 50%; this is because of the rising displacement information redundancy of adjacent points. And when the sampling steps rising, the accuracy of results will decrease oppositely, and this is because that the extracted features have lost too much space information.

Table 1. Classification accuracy at different sampling steps

Sampling steps Classification	1	2	3	4	5
Total（Group 1）	63%	86%	55%	58%	58%
Total（Group 2）	38%	73%	40%	38%	38%
Total（Group 3）	50%	90%	55%	59%	51%

b) Comparisons with other algorithms

Under the same test conditions, we have compared our algorithm with several recent classification algorithms and classical algorithms, such as GIST[14], OTC[15], BOW[16] and CENTRIST[17]. Comparison results are shown in table 2. As shown in the table, our proposed algorithm is superior to the list methods and the average classification rates achieved by our algorithm are more than 10%.

Table 2. Classification accuracy at different algorithms

Methods Classification	SIFT-Flow	BOW	GIST	CENTRIST	OTC
Total（Group 1）	63%	86%	55%	58%	58%
Total（Group 2）	38%	73%	40%	38%	38%
Total（Group 3）	50%	90%	55%	59%	51%

In the recognition of similar scenes, OTC and GIST have relatively poor robustness and all images can be easily recognized as a single category, extracted features are lack of discriminative information to accurately classify the scenes. And our image representation method can extract more image information and achieve high classification accuracy.

Conclusion

Our paper proposes a method for classification of similar scenes. By using SIFT-Flow, we can get the displacement vector map of all the characteristic points in the image. The image is divided into multiple partial blocks and the matching degree of each block should be integrated in a feature vector. We test our algorithm in 67 Indoor Scenes, and we can achieve good classification results compared to the classical algorithm and recent good classification algorithms we list.

Although our method has good classification results, but the stability of image matching still needs to be further improved, and how to enhance the entropy of matching maps needs further experimental and theoretical proof.

References

[1] Dalal N, Triggs B. Histograms of oriented gradients for human detection[J]. Proceedings. CVPR, IEEE Computer Society Conference on Computer Vision and Pattern Recognition. IEEE Computer Society Conference on Computer Vision and Pattern Recognition, 2005, 1(12):886-893.

[2] Lowe D G. Object Recognition from Local Scale-Invariant Features[C]. PROC. OF THE INTERNATIONAL CONFERENCE ON COMPUTER VISION, CORFU. 1999:1150.

[3] J Sivic. A Zisserman. Video Google: A text retrieval approach to object matching in videos[J]. Iccv, 2003, 2:1470.

[4] Shuang Bai, Tetsuya Matsumoto, Yoshinori Takeuchi, Hiroaki Kudo and Noboru Ohnishi,Informative Patches Sampling for Image Classification by Utilizing Bottom-up and Top-down Information, Machine Vision and Applications, Vol. 24, No.5, pp 959-970,2013.

[5] Lazebnik, S., Schmid, C., Ponce, J. Beyond bags of features: Spatial pyramidmatching for recognizing natural scene categories. In: CVPR.

[6] Wang YuXin, Guo He. Bag of Spatial Visual Words Model for Scene Classification[J]. Computer Science, 2011, 38(8):265-268.

[7] Zhang Ruijie,Li Bicheng,Wei Fushan. Image Scene Classification Based on Multi-Scale and Contextual Semantic Information[J]. Acta Electronica Sinica, 2014, 42(4):646-652.

[8] Fei-Fei L, Fergus R, Perona P. Learning generative visual models from few training examples : An incremental Bayesian approach tested on 101 object categories[J]. Computer Vision & Image Understanding, 2007, 106(1):178.

[9] Hartigan J A, Wong M A. Algorithm AS 136: A K-Means Clustering Algorithm[J]. Applied Statistics, 1979:100-108.

[10] Liu C, Yuen J, Torralba A, et al. SIFT Flow: Dense Correspondence across Different Scenes[C]. Proceedings of the 10th European Conference on Computer Vision: Part III. Springer-Verlag, 2008:28-42.

[11] Lazebnik S, Schmid C, Ponce J. Spatial Pyramid Matching[J]. Cambridge University Press, 2009.

[12] A. Quattoni, A. Torralba. Recognizing indoor scenes[C]. Computer Vision and Pattern Recognition, 2009. CVPR 2009. IEEE Conference on. IEEE, 2009:413 - 420.

[13] Chang C C, Lin C J. LIBSVM: A Library for Support Vector Machines[J]. Acm Transactions on Intelligent Systems & Technology, 2001, 2(3):389-396.

[14] Oliva A. Modeling The Shape Of The Scene: A Holistic Representation Of The Spatial Envelope[J]. International Journal of Computer Vision, 2001, 42(3):145-175.

[15] Ran M, Zelnik-Manor L, Tal A. OTC: A Novel Local Descriptor for Scene Classification[J]. Lecture Notes in Computer Science, 2014:377-391.

[16] Lazebnik S, Schmid C, Ponce J. Beyond Bags of Features: Spatial Pyramid Matching for Recognizing Natural Scene Categories[C].Proceedings of the 2006 IEEE Computer Society Conference on Computer Vision and Pattern Recognition - 2006:2169-2178.

[17] Wu J, Rehg JM.CENTRIST: A Visual Descriptor for Scene Categorization[C]. IEEE Trans Pattern Anal Mach Intell, 2011,33(8):1489-1501.