Object classification based on RGB-D features fusion and LLC coding

Zhang Hong1*

1College of Information Science and Engineering, Wuhan University of Science and Technology, Wuhan, 430081, China

Corresponding author’s e-mail: hzhang1998@163.com

Abstract. An object classification method based on RGB-D features fusion is proposed in this paper. Firstly, image features in terms of texture, size, shape, from the RGB image and the depth image are extracted; secondly, the improved K-means clustering method is used to cluster the different extracted features to obtain seven different visual dictionaries respectively; then, the LLC coding model is used to perform feature coding on the extracted visual features to obtain seven different image feature expressions; meanwhile, above seven features expression are fused using the linear cascade mode to obtain the fusion feature; finally, linear SVM performs classifier is used to realize the object classification based on fusion feature. Experimental results show that compared with some existing algorithm based on RGB-D, the accuracy of our proposed method is increased to 87.3%. The accuracy of object classification is significantly improved.

1. Introduction

Object classification has always been one of the research hotspots in the fields of computer vision, machine learning and robotics [1]. Object is susceptible to changes in illumination, viewing angle, and object scale, so that object classification still faces many challenges.

In the past few years, some researchers have directly extracted features of target objects from depth images to complete object classification. Tham [2] use Moment Invariants features with translation, rotation, and scaling invariance as features of depth images to achieve object classification. Bradski [3] used Viewpoint Feature Histogram as the feature of the depth image to complete object classification. These algorithms complete object classification by extracting corresponding feature descriptors from depth images, but these algorithms do not use some important features of RGB images, and they also have disadvantages such as high computational complexity and slow speed.

To solve the many problems caused by directly extracting the information of the depth image for object classification, some scholars have proposed a classification method that combines RGB image features and depth image features. Bo [4] proposed a deep kernel descriptor, and RGB features for object classification. Silbermann [5] extracted the features of SIFT in RGB images and depth images respectively, and they have good classification performance. Khan[6] used feature extraction techniques such as Color Autocorrelogram (Color Autocorrelogram), Wavelet Moments (Wavelet Moments), Local Binary Pattern (LBP), and KNN algorithm is used to realize object classification. The above proposed algorithm uses the information provided by the RGB-D image to improve the accuracy of object classification to a certain extent, but it still has certain limitations for target objects that are similar in color and spatial shape.

*Corresponding author’s e-mail: hzhang1998@163.com

Published under licence by IOP Publishing Ltd
For this reason, based on the above research and analysis, this paper proposes an object classification algorithm based on RGB-D fusion features.

2. Features extraction

2.1. SIFT feature and its improvement

SIFT [7] (Scale Invariant Feature Transform) is a local feature descriptor with good scale and rotation invariance, and at the same time, the influence of illumination and the change of the 3D camera viewpoint can remain partially unchanged.

(1) Construct image scale space

Identify the scale and position that can be identified from different views of the object or scene. This is achieved by using a Gaussian value of a scale space kernel. The definition of the comparative space is as follows:

\[ L(x, y, \sigma) = G(x, y, \sigma) * I(x, y) \]  

(1)

Where \( I(x, y) \) represents the spatial coordinates of the pixel; different image scales correspond to different parameters \( \sigma \).

At the same time, for more convenient description and calculation, the difference of Gaussian (Difference of Gaussian, DoG) operator is introduced to construct image representations in different scale spaces:

\[ D(x, y, \sigma) = G(x, y, k\sigma) - G(x, y, \sigma) * I(x, y) \]

\[ = L(x, y, k\sigma) - L(x, y, \sigma) \]  

(2)

(2) Detect extreme points in scale space (candidate key point detection)

Compare each pixel in the image with 26 pixels in the neighbouring pixels. If the point is the maximum or minimum in the neighbouring of 26 pixels, the pixel is determined as an extreme point detected in the corresponding scale space.

(3) Assign direction parameters for characteristic points

In the previous step, the feature points in each image are determined. In order to obtain a feature descriptor with rotation invariance, each feature point calculates a direction, use the gradient direction distribution of the pixels in the neighborhood of the feature point to specify the direction for each feature point. Where the gradient magnitude \( m(x, y) \) and gradient direction \( \theta(x, y) \) can be obtained from literature [7].

(4) Generate SIFT feature descriptor

The SIFT descriptor calculates the gradient vector of each pixel in the neighbouring feature points and establishes a normalized gradient direction histogram. A group of 16 histograms are aligned in a grid to obtain a 128-dimensional feature vector for each region of the target object. Then the feature vector of each region is normalized to unit length, and the part with small value is removed by the threshold. Finally, calculate the gradient size and direction of each sample point in the area around the key point, and create a SIFT descriptor.

During image classification, whether a large-scale and uniform feature point distribution can be obtained in the process of image local feature extraction is a key factor to ensure high accuracy of classification.

When the SIFT algorithm detects spatial extreme points (candidate key point detection), each key point must be related to 8 adjacent pixels in the same scale and 18 adjacent pixels in the upper and lower adjacent scales (total 8 +18=26 pixels) to compare to determine whether it is an extreme value. Since the SIFT algorithm has a small detection range in the candidate key point detection process, the number of detected feature points is large in a certain area, and the detected area is concentrated in the range of feature saliency, resulting in the generation of local extreme values. The LLC coding model has many dictionaries in a certain area of the image. The dictionary reflects the characteristics of certain areas in the image, and cannot reflect the overall characteristics of the image. In order to avoid such a situation, non-maximum suppression is used when obtaining the extreme points in DoG space.
Extremum detection is performed on a total of \((4R^2R+18)\) pixels in the area with the key point position as the center and \(R\) as the radius.

2.2. Kernel descriptor extraction

The kernel descriptor [4] avoids the discretization of pixel attributes by adopting the kernel view of the similarity of image patches (such as patch \(6 \times 6\)). The similarity between two image patches is based on a kernel function, called a matching kernel function, which averages the continuous similarities between all pixel attributes in two image patches. The matching kernel is very flexible, because the distance function between pixel attributes can be any positive definite kernel, such as the popular Gaussian kernel function.

The kernel descriptors of image classification involves the following steps: (1) designing the pixel attribute matching kernel; (2) using the kernel principle to analyze and learn compact basis vectors; (3) by projecting high-dimensional feature vectors to the learned The basis vector is used to construct the kernel descriptor. This paper describes how to apply the above steps to gradient, color, and shape pixel attributes, resulting in four effective kernel descriptors. Since the low-dimensional expression and feature vector mapping process of other kernel descriptors matching the kernel is roughly similar to the size kernel descriptor, only the size-kernel descriptor (Size-KD) is introduced. Spin kernel descriptor (Spin-KD), gradient kernel descriptor (Gradient-KD), LBP kernel descriptor (LBP-KD) references used in this article [4].

The Size-KD can obtain the physical size and shape information of the target object through the 3D point cloud converted from the depth image. In order to obtain the size kernel descriptor (size-KD), each pixel is first mapped to a corresponding three-dimensional coordinate vector, and the depth image is converted into a three-dimensional point cloud. In order to capture the size of the target, the distance between each point and the reference point of the point cloud is calculated. Specifically, let \(p\) denote a point cloud and \(p^*\) is the reference point of the point cloud.

\[
d_p = \|p - p^*\|_2
\]

In order to calculate the similarity between the distance attributes of the two point clouds \(P\) and \(Q\), a matching kernel is introduced

\[
k_{size}(P, Q) = \sum_{p \in P} \sum_{q \in Q} k_{size}(d_p, d_q) \quad (3)
\]

Among them, \(k_{size}(d_p, d_q) = \exp (-\gamma_s \|d_p - d_q\|_2) (\gamma_s > 0)\) is a Gaussian kernel function. It could be seen that the matching kernel \(k_{size}\) calculates the similarity of the two sets \(P\) and \(Q\) by aggregating all distance attribute pairs. Due to the introduction of the Gaussian kernel, the dimensionality of the feature vector on \(P\) is high-dimensional. According to the idea of literature [4], this high-dimensional feature vector is projected onto a set of finite basis vectors to obtain a finite-dimensional kernel descriptor:

\[
F^e_{size}(P) = \sum_{i=1}^{b_e} \alpha^e_i \sum_{p \in P} k_{size}(d_p, u_i) \quad (4)
\]

Where \(u_i\) is the basis vector extracted from the support area of the distance attribute, \(b_e\) is the number of basis vectors (set to 50 in this paper), \(\{\alpha^e_i\}_{i=1}^{E}\) is the top \(E\) feature vector calculated by the kernel principal component analysis.

3. Features coding

3.1. LLC feature coding

The above-obtained features are still low-level visual features and are generally not directly used for classification. These features need to be encoded to high-level visual features.

Local constraint linear coding [8] uses local constraints to project each feature descriptor into a local coordinate system. The LLC coding guidelines are as follows:
\[
\min_c \sum_{i=1}^{N} \|x_i - Bc_i\|^2 + \lambda \|d_i \otimes c_i\|^2 \\
\text{s.t. } \mathbf{1}^T c_i = 1, \forall i
\] (5)

Where \(x_i\) is the feature to be coded, \(B\) is the visual dictionary learned by the K-means clustering algorithm; \(c_i\) is the feature coding coefficient to be optimized; \(\lambda\) is the penalty factor item of the LLC, and the constraint condition \(\mathbf{1}^T c_i = 1\) guarantees the translation invariance of feature coding. \(d_i\) represents a locality adapter that assigns different degrees of freedom to each basis vector.

3.2. The visual dictionary construction based on improved K-means

After feature extraction, visual dictionary learning is required. Visual dictionary are generally required to be complete and representative. Generally, a clustering algorithm is used to learn the dictionary, and the common method is K-means clustering. In the traditional K-means: the selection of initial clustering centers are random, which causes the clustering results to be greatly affected. If the initial central point is not selected properly, the algorithm will fall into a local optimal solution. Therefore, the improved K-means [13] is adopted to build a visual dictionary for the extracted features. The kernel idea of the improved K-means algorithm is to make the Euclidean distance between them as large as possible when selecting the initial cluster centers.

4. Object classification based on fused features

Object classification is shown in Figure 1, which includes two stages: training phase and testing phase.

![Figure 1. Structure diagram of our proposed method.](image)

**Training phase**: image Feature sets such as SIFT, LBP-KD, Gradient-KD based on color image, and such as Gradient-KD, LBP-KD, Size-KD, Spin-KD are extracted based on depth image feature sets respectively. Then, the improved K-means [13] is used to obtain visual dictionary corresponding to each feature, and the corresponding feature are encoded by LLC feature. Finally, image expression sets are fused with a linear cascade mode and trained using a SVM classifier.

**Testing phase**: the extracted features are LLC-encoded according to visual dictionaries to obtain image expression set. Then a SVM classifier is used to realize image classification.

5. Experiment results and analysis

5.1. Experimental dataset and parameters

RGB-D Object image dataset is collected with kinect by Kevin Lai of the University of Washington in the United States [9]. This dataset includes 51 categories and more than 300 kinds of RGB-D images of daily household items under different lighting and viewing angle changes. The relevant parameters in our algorithm are set as follows: the sampling window of improved SIFT feature is set to 8×8, the sampling distance is 8 pixels; the kernel descriptor is set according to section 2.2; spatial pyramid layers of image is 2; the size of the visual dictionary is 256; liblinear toolkit is used for SVM
classification; for each type of target object, the training set and the test set are selected according to the ratio of 8:2.

5.2. Classification results comparison
The comparisons with other existing methods in the RGB-D object dataset, as shown in Table 1. Literature [4] extracts kernel descriptor from the original RGB-D image, and uses different classifiers method to achieve the final classification. Literature [10] extracts a variety of features from the RGB-D image, and then learns the optimal distance feature from these extracted features to realize the classification. Literature [11] extracts a set of local SURF features of RGB-D image, and then learn convolutional K-means features from local SURF feature, and final use convolution K-means features for classification. Based on literature [4], an improvement was made for the Size-KD, and then the kernel descriptors after with LLC feature encoding were concatenated and input SVM for classification in [12]. In this experiment, the algorithm proposed in this paper achieved an accuracy of 87.3%, which is an improvement of about 0.5% compared with the best result before. our method has better classification performance than above existing methods.

| Methods | Classification results (%) |
|---------|----------------------------|
| Kernel Descriptors + Linear SVM[4] | 81.9 |
| Kernel Descriptors + NonLinear SVM[4] | 83.8 |
| Kernel Descriptors + Random Forest[4] | 79.6 |
| IDL[10] | 85.4 |
| CKM desc[11] | 86.4 |
| Kernel Descriptors + LLC[12] | 86.8 |
| **ours** | **87.3** |

5.3. Comparison of features fusion
This experiment compares the impact of different features on the classification results. In table 2, the classification accuracy with RGB or Depth image features are 82.1% and 67.5%, respectively. The accuracy rate of ours based on fusion of RGB-D features and LLC encoding is 87.3%, which is higher than others using single-modal image features(only RGB or Depth features). Experimental results shows that combining depth information can effectively improve the classification accuracy.

| Features | Classification results (%) |
|----------|----------------------------|
| RGB-ALL | 82.1 |
| Depth-ALL | 67.5 |
| RGB-ALL+Depth-ALL | 87.3 |

5.4. Comparison of classification results before and after SIFT feature improvement
This experiment examines the impact of the improved SIFT algorithm on image classification results. 10,000, 20,000, or 30,000 images on the RGB-D object dataset are selected respectively for comparison experiments. As shown in Figure 2, based on different the number of experimental images, our improved SIFT features can get better classification results.
6. Conclusions
This paper adopts the technique of multi-feature fusion to improve object classification accuracy of RGB-D images. Based on RGB and depth image characteristics, texture, size, and shape features are extracted from RGB images and depth images separately. Then, visual dictionaries of above features are constructed by an improved K-means clustering method, clusters the different extracted features; above extracted features are coding using the LLC coding model respectively and then are fused to fusion feature by linear cascade mode; finally, linear SVM is adopted to classify the object based on fusion feature from RGB-D images. The experimental results of comparison with other methods on the RGB-D dataset show that this classification method is superior to the existing methods and has a higher classification accuracy.

References
[1] Zhu H, Weibel J B, Lu S. Discriminative Multi-modal Feature Fusion for RGBD Indoor Scene Classification[C]// IEEE Conference on Computer Vision & Pattern Classification. 2016: 2969-2976.
[2] Tham J S, Chen Y S, Fauzi M F A, et al. Depth image object classification using moment invariants[C]// IEEE International Conference on Consumer ElectronicsTaiwan2016: 1-2.
[3] BRADSKI G, RUSU R B, THIBAUX R, et al. Fast 3D classification and pose using the viewpoint feature histogram[C]. 2010 IEEE RSJ International Conference on Intelligent Robots and Systems(IROS). IEEE, 2010: 2155-2162.
[4] BO L, LAI K, REN X, et al. Object classification with hierarchica-l kernel descriptors[C]. 2011 IEEE Conference on Computer Vision and Pattern Classification(CVPR). IEEE, 2011: 1729-1736.
[5] Silberman N, Hoiem D, Kohli P, et al. Indoor Segmentation and Support Inference from RGB-D Images[C]. Proceedings of the 12th European Conference on Computer Vision. Springer Berlin, 2012: 746-760.
[6] Khan W, Phaisangittisagul E, Ali L, et al. Combining features for RGB-D object classification[C]//2017 International Conference on Electrical Engineering Congress (iEECON), 2017: 1-5.
[7] Wei M, Xiwei P. WLIB-SIFT: A Distinctive Local Image Feature Descriptor[C]// 2019 IEEE 2nd International Conference on Information Communication and Signal Processing (ICICSP), 2019: 379-383.
[8] Wang J, Yang J, Yu K, et al. Locality-constrained linear coding for image classification[C]//IEEE Conference on CVPR. San Francisco, 2010: 3360-3367.
[9] Herbst E, Ren X, Fox D. RGB-D object discovery via multi-scene analysis[C]//IEEE Conference on Intelligent Robots and Systems, San Francisco, 2011: 4850-4856.
[10] Lai K, Bo L, Ren X, et al. Sparse distance learning for object classification combining RGB and depth information[C]//Robotics and Automation (ICRA), 2011 IEEE International Conference on. IEEE, 2011: 4007-4013.

[11] Blum M, Springenberg J T, Wülffing J, et al. A learned feature descriptor for object classification in RGB-D data[C]//Robotics and Automation (ICRA), 2012 IEEE International Conference on. IEEE, 2012: 1298-1303.

[12] Luo J, Jiang M. Object classification method based on RGB-D image kernel descriptor. Journal of Computer Applications. 2017, 37(10): 255-261. (in Chinese)

[13] Arthur D, Vassilvitskii S. k-means++: The advantages of careful seeding[C]//Proceedings of the eighteenth annual ACM-SIAM symposium on Discrete algorithms. Society for Industrial and Applied Mathematics, 2007: 1027-1035.