GENERIC OBJECT RECOGNITION BASED ON FEATURE FUSION IN ROBOT PERCEPTION

Xinde Li,* Chaomin Luo,** Jean Dezert,*** and Yingzi Tan*

Abstract

A new generic object recognition (GOR) method for robot perception is proposed in this paper, based on multi-feature fusion of two-dimensional (2D) and 3D scale invariant feature transform descriptors drawn from 2D images and 3D point clouds. The trained support vector machine is utilized to construct multi-category classifiers that recognize the objects. According to our results, this new GOR approach achieves higher recognition rates than classical methods tested, even when one has large intra-class variations, or high inter-class similarities of the objects. Simulation results demonstrate the effectiveness and efficiency of the proposed GOR approach.

Key Words

Generic object recognition, Point cloud, SIFT, Feature fusion, SVM, belief functions

1. Introduction

Finding the object for robots according to the natural language description depends on the ability of generic object recognition (GOR) in real-world applications, which plays a significant role in computer vision and artificial intelligence with applications in intelligent monitoring, robotics, medical image processing, *etc.* [1]–[3]. Unlike specific object recognition (SOR) [4], [5], GOR is much more difficult to be accomplished. The generic features of objects express the common properties in the same class, but help to make difference between classes, which need to be found out, instead of defining characteristics of particular category as used in SOR methods. Currently, most techniques for GOR are focused on local feature extraction algorithms on 2D images, typically the 2D SIFT (scale invariant feature transform) descriptors [6], [7]. However, 2D images have loss of the 3D information of the objects and are susceptible to change due to various external illumination conditions. To overcome this drawback, 3D SIFT descriptors based on volumes [3], [8]–[12], and point cloud model [13]–[15] have been proposed recently by a few researchers. Point cloud model of object obtained from the depth images depends on the geometry of the objects, but has nothing to do with the brightness and reflection features of the objects. Therefore, it is very important for 3D SIFT descriptors based on point cloud model to recognize generic object when a robot moves in unknown environments.

In this paper, to further increase the rate of recognition and improve the performance of GOR, we propose a new method for GOR on the basis of feature fusion of 2D and 3D SIFT descriptors. Its main novelties include: (1) the 3D SIFT feature descriptor from the point cloud representation model is improved; (2) the improved 3D SIFT feature descriptor is applied to GOR; (3) 2D and 3D SIFT features are fused together to accomplish the algorithm of GOR.

The rest of paper is organized as follows. The recognition algorithm is described in detail in Section 2 and its program procedure is given in Section 3. Section 4 evaluates the performance of this new method on real data sets. Conclusions with perspectives are summarized in Section 5.

2. New GOR Method

This new method of GOR that consists of three main steps (features extraction and representation, features fusion, and classifier design) is presented in detail in this section.

2.1 Features Extraction and Representation

2.1.1 2D SIFT Descriptor

In 1999, Lowe [6] presented for the first time a new method to extract keypoints of objects in images and described their local features to make GOR, *e.g.*, in computer vision applications. The method was improved in [7] and extended to 3D by other researchers (see next paragraph).

* Key Laboratory of Measurement and Control of CSE of Ministry of Education, School of Automation, Southeast University, Nanjing, 210096, China; e-mail: {xindeli, tanyz}@seu.edu.cn

** ECE Department, University of Detroit Mercy, Detroit, MI, USA; e-mail: luoch@udmercy.edu

*** The French Aerospace Lab, Palaiseau, F-91761, France; e-mail: jean.dezert@onera.fr

The feature description of the object drawn from a training image is then used to identify the presence (if any) of the object in real (usually cluttered) observed scene. To obtain good object recognition performance, a 2D SIFT [16], [17] was proposed to warranty that the features extracted (*i.e.*, the keypoints) from the training image are detectable under changes in image orientation, scale, noise, and illumination, and even if partial object occlusions occur in the observed scene. This is because Lowe's SIFT feature descriptor is invariant to uniform scaling, orientation, and partially invariant to illumination changes and robust to local geometric (affine) distortion. The stable keypoints locations of SIFT are provided by the detection of scale space extrema in the difference-of-Gaussian (DoG) function $D(x, y, \sigma)$ convolved with the image $I(x, y)$, more precisely [7].

The local extreme points of $D(x, y, \sigma)$ functions (DoG images) define the set of keypoint candidates (the SIFT descriptor). Once all the keypoints are determined, one must assign a consistent orientation based on local image properties, from which the keypoint descriptor can be represented, hence achieving invariance to image rotation. For this, the scale of the keypoint is used to choose the Gaussian-blurred image L with the closest scale. The keypoint descriptor is created by computing at first the gradient magnitude $m(x, y)$ and its orientation $\theta(x, y)$ at each pixel (x, y) in the region around the keypoint in this Gaussian-blurred image L as follows [7]:

$$\begin{cases} m(x,y) = \sqrt{L_x^2 + L_y^2} \\ \theta(x,y) = \tan^{-1}\left(\frac{L_y}{L_x}\right) \end{cases} \tag{1}$$

with $L_x \triangleq L(x+1, y) - L(x-1, y)$ and $L_y \triangleq L(x, y+1) - L(x, y-1)$. A set of orientation histograms is created on 4×4 pixel neighbourhoods with eight directions (bins) each. These histograms are computed from magnitude and orientation values of samples in a 16×16 region around the keypoint such that each histogram contains samples from a 4×4 sub-region of the original neighbourhood region. The magnitudes are weighted by a Gaussian function with σ equal to one half the width of the descriptor window. The descriptor then becomes a 128D feature vector because there are $4 \times 4 = 16$ histograms each with eight directions. This vector is therefore normalized to unit length to enhance invariance to affine changes in illumination. Additionally, a threshold of 0.2 is applied to reduce the effects of non-linear illumination, and the vector is again normalized. The simplest method to find the best candidate match for each keypoint would consist in identifying its nearest neighbour based on Euclidean distance metric in the database of keypoints from training images.

2.1.2 3D SIFT Descriptor

The previous 2D SIFT descriptor working with pixels has been extended to 3D using volumes in different manners [3], [8]–[12]. In this paper, we adopt the 3D SIFT for point cloud inspired by [8], [15]. However, all the methods require same functional steps as for 2D SIFT, that is (1) keypoints detection; (2) keypoints orientation; and (3) descriptor representation. We present these steps in detail in the next subsections.

Keypoint detection

The scale space of a 3D input point cloud is defined as a 4D function $L(x, y, z, \sigma) = G(x, y, z, k\sigma) * P(x, y, z)$ obtained by the convolution of a 3D variable-scale centred Gaussian kernel $G(x, y, z, \sigma)$, $\sigma = \sqrt[3]{2}$, with the input point $P(x, y, z)$, where:

$$G(x,y,z,\sigma) = \frac{1}{\left(\sqrt{2\pi}\sigma\right)^3} e^{-(x^2+y^2+z^2)/2\sigma^2} \tag{2}$$

Extending Lowe's approach [7], the candidate keypoints in 4D scale space are taken as the local extrema (maxima or minima) of the multi-scale DoG defined by:

$$D(x,y,z,k^i\sigma) = L(x,y,z,k^{i+1}\sigma) - L(x,y,z,k^i\sigma) \tag{3}$$

To find extrema of the multi-scale DoG function, each sample point is compared to its $27 + 26 + 27 = 80$ neighbours, where 26 neighbours belong to the current scale and 27 neighbours in the scale above and below. A keypoint is chosen only if it is larger than all of its neighbours or smaller than all of them. To eliminate the bad candidate keypoints having low contrast, one uses a thresholding method to remove the erroneous points. A contrast threshold is applied on $D(x, y, z, k^i\sigma)$ to eliminate all the candidate keypoints below a chosen threshold value $\tau (\tau = 0.5$ in our simulations).

Keypoint Orientations

Similarly, for 2D SIFT, once all the keypoints are determined in 3D, one may assign a consistent orientation based on local points properties, from which the keypoint descriptor can be represented, hence achieving invariance to object rotation. For this, the 2D histogram is calculated by gathering statistics of the angles between the neighbouring points and their centre. The keypoint descriptor is created by computing at first the gradient magnitude $m(x, y, z)$ and its orientations $\theta(x, y, z)$ (azimuth angle) and $\varphi(x, y, z)$ (elevation angle) between each point (x, y, z) in the region around the keypoint and their centre (x_c, y_c, z_c) as follows:

$$\begin{cases} m(x,y,z) = \sqrt{(x-x_c)^2 + (y-y_c)^2 + (z-z_c)^2} \\ \theta(x,y,z) = \tan^{-1}\left((y-y_c)/(x-x_c)\right) \\ \varphi(x,y,z) = \sin^{-1}\left((z-z_c)/m(x,y,z)\right) \end{cases} \tag{4}$$

In 3D point cloud, each point has two values representing the orientation of the region, as shown in Fig. 1, whereas in 2D case each pixel had only one direction of the gradient.

Extending Lowe's approach in 3D case, to find the keypoint orientations, we construct a weighted histogram for the 3D neighbourhood around each candidate keypoint, which can be achieved by multiple methods. In this work, a 2D histogram is produced by grouping the angles in bins which divide θ and φ into 10 deg angular bins. A regional Gaussian weighting of $e^{-(2d/R_{\max})^2}$ for the points whose

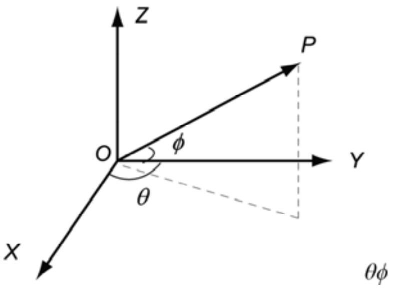

Figure 1. $\theta(x, y, z)$ (azimuth angle) and $\phi(x, y, z)$ (elevation angle).

magnitude is d is applied to the histogram, where $R_{\max}$ represents the max distance from the centre. The sample points at a distance greater than $R_{\max}$ are ignored. The histogram is smoothed using a Gaussian filter to limit the effect of noise. The dominant azimuth α and elevation β of the keypoint are determined by the peaks of the 2D histogram. To enhance robustness, peaks in the histogram within 80% of the largest peak are also retained as possible secondary orientations.

Descriptor Representation

Each keypoint p is described by its location $\mathbf{p}[x_p, y_p, z_p]^t$, scale σ_p, and orientation angles α_p and β_p. The descriptor representation associated with a keypoint p is based on the local spatial characteristics around it to describe its features. To ensure rotation invariance of the descriptor, the k-points $p_i (i = (1, \ldots, k))$ of coordinates $\mathbf{p}_i[x_i, y_i, z_i]^t$ around the keypoint of interest p are transformed (rotated) in the dominant orientation of p by the following transformation:

$$\mathbf{p}'_i = \begin{bmatrix} \cos\alpha_p \cos\beta_p & -\sin\alpha_p & -\cos\alpha_p \sin\beta_p \\ \sin\alpha_p \cos\beta_p & \cos\alpha_p & -\sin\alpha_p \sin\beta_p \\ \sin\beta_p & 0 & \cos\beta_p \end{bmatrix} \cdot \mathbf{p}_i \quad (5)$$

Before the normal vector $\mathbf{n}$ at the keypoint in the k-points neighbourhood is calculated according to the routine available in the open point cloud library [18]. For each (rotated) point $\mathbf{p}'_i (i = (1, \ldots, k))$ in the k-points neighbourhood of the keypoint p, we calculate the vector $\mathbf{pp}'_i$, the magnitude m, angles θ and φ according to (6). The angle δ between $\mathbf{n}$ and $\mathbf{pp}'_i$ is given by:

$$\delta = \cos^{-1}\left(\frac{\mathbf{pp}'_i \cdot \mathbf{n}}{|\mathbf{pp}'_i| \cdot |\mathbf{n}|}\right) \quad (6)$$

Therefore, a keypoint p with its neighbour p_i is represented by the 4-tuple $(m, \theta, \varphi, \delta)$. To reduce the computational effort, instead of dividing the neighbourhood into $n \times n$ sub-regions (with $n = 4$ as in Lowe's 2D SIFT descriptor), we take directly the entire neighbourhood, which means that we have $n = 1$. The histogram used to generate the 3D descriptor at the keypoint p is derived by splitting $(\theta, \varphi, \delta)$ space into 45 deg bins, and adding up the number of points with the Gaussian weighting of $e^{-(2m/R_{\max})^2}$. As a result, the dimension of our 3D SIFT descriptor is $n \times n \times 4 \times 4 \times 8 = 128$ (as for the 2D SIFT descriptor described previously), Due to (1), $n = 1$; (2) the azimuth angle $\theta \in [0, 360]$ deg which is split into 8 bins of 45 deg; (3) the elevation angle $\varphi \in [-90, 90]$ deg which is split into 4 bins of 45 deg; (4) $\delta \in [0, 180]$ deg which is also split into 4 bins of 45 deg. Each 3D SIFT descriptor is normalized to unity.

The 2D and 3D SIFT descriptors summarize efficiently the useful information contained in 2D images and 3D point clouds. Instead of working directly with whole images and point clouds, it is usually more interesting (in terms of computational burden reduction) to work directly with 2D and 3D SIFT descriptors, particularly, if real-time object recognition is necessary. Generally, the objects characterized by 2D and 3D SIFT descriptors have different number of keypoints, which makes the feature fusion problem for object recognition very challenging. For example, for a simple object like an apple, we can obtain 45 keypoints using 3D SIFT descriptor and 38 keypoints using 2D SIFT descriptor. To overcome this problem, we adopt the bag of words (BoW) model [19] to gather the statistics of the 2D and 3D SIFT descriptors to describe the objects.

2.1.3 BoW Model for Features Vector

In the BoW feature model, the feature descriptors of all the interest points are quantized by clustering them into a pre-specified number of clusters (*i.e.*, $K = 300$). Instead of using k-means algorithm as in [2], we use the k-means++ method [20] which selects more effectively the initial cluster centres to complete this step. The resultant cluster centres are now called visual words, while the collection of these cluster centres is referred to as the visual word vocabulary. Once our vocabulary is computed, the descriptors are matched to each visual word based on the Euclidean distance and the frequency of the visual words in image and in point cloud is accumulated into a histogram, which is the BoW feature vector of the image and of the point cloud. So each object in 2D image and in 3D point cloud is described by a 1×300 BoW-based feature vector denoted respectively $\mathbf{BoW}_{2\mathrm{D}}$ and $\mathbf{BoW}_{3\mathrm{D}}$. These two BoW-based feature vectors will be used for feeding the trained support vector machine (SVM) classifiers to get the final object recognition.

2.2 Classifier Design

Once the object description is completed, SVMs are trained to learn object categories and to perform the object classification. SVM is a supervised and discriminative machine learning method providing usually good performance. Through offline training of pre-limited samples, we seek a compromise between model complexity and learning ability, to obtain a satisfied discriminant function [21]. Linear SVM classifier is applied for its efficiency, which is a typical classifier for two categories problems. In plenty of real-world applications, we are face to multi-category classification problems and use trained 1V1 SVMs between classes to set up a multi-category classifier. The training

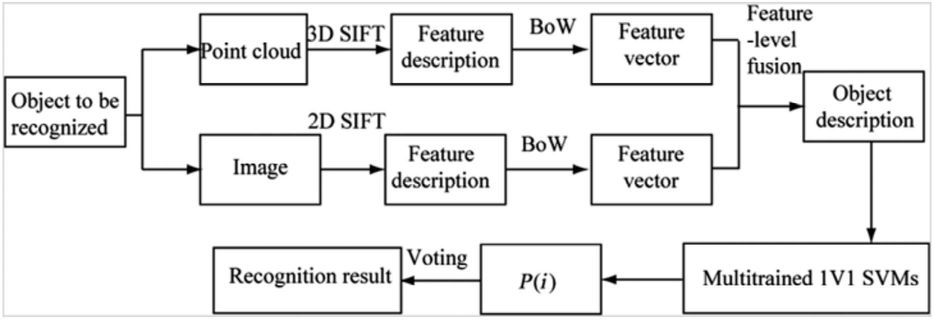

Figure 2. Direct feature-level fusion strategy.

process is accomplished as follows: for training samples belonging to the ith category, we make a pairwise SVM training with respect to all the other classes. Thus, we achieve $C_n^2 = n(n-1)/2$ 1V1 SVM classifiers for training samples of n categories.

2.3 Features Fusion Strategies

When the two BoW-based feature vectors of the object to be recognized have been computed from 2D and 3D SIFT descriptors, we have to use them to achieve the object recognition thanks to the trained SVM classifiers from the BoW-based features vectors of known objects of our database. In this paper, we mainly present the direct feature-level fusion strategy. This feature-level fusion is for feeding SVM classifiers in training phase before making object recognition in testing phase. With this strategy we combine (fuse) directly the two BoW-based feature vectors $\mathbf{BoW}_{\mathrm{2D}}$ and $\mathbf{BoW}_{\mathrm{3D}}$ to get a 1×600 vector $\mathbf{BoW}_{\mathrm{2D},3D} \triangleq [\mathbf{BoW}_{\mathrm{2D}}, \mathbf{BoW}_{\mathrm{3D}}]$, and we feed the trained (global) SVM classifiers with the fused vector to get the final recognition. The principle of our method based on this strategy is summarized in Fig. 2.

3. The Procedure of Our New GOR Method

This section focuses on the algorithm procedure of our new GOR method. The GOR process mainly includes the training phase and the testing phase.

3.1 The Training Phase

This training phase consists of the following steps:

a. For any object to be trained, we extract its 2D and 3D SIFT descriptors associated with each keypoint. N_{2D} 2D SIFT descriptors of size 1×128 are available, if one has extracted N_{2D} keypoints from the 2D image under training, and we get N_{3D} 3D SIFT descriptors of size 1×128 if one has extracted N_{3D} keypoints from the 3D point cloud under training.

b. The visual words vocabulary of 2D SIFT descriptors $\mathit{center}_{\mathrm{2D}} = \{\mathit{center}_{\mathrm{2D}l},\ \ l = 1, 2, \ldots, K\}$ is obtained by k-means++ clustering algorithm [20] from the N_{2D} 2D SIFT descriptors. Similarly, we can obtain the visual words vocabulary of 3D SIFT $\mathit{center}_{\mathrm{3D}} = \{\mathit{center}_{\mathrm{3D}l},\ l = 1,\ 2,\ \ldots,\ K\}$, K representing the number of cluster centres.

c. For each 2D SIFT descriptor with 128 components, we calculate the Euclidean distance between the descriptor and the visual word vocabulary $\mathit{center}_{\mathrm{2D}}$, and the frequency of the visual words in image is accumulated into a histogram, which is the BoW feature vector of the image. Similarly, the BoW feature vector of the point cloud is obtained. So each object in 2D image and in 3D point cloud is described by a 1×300 BoW-based feature vector denoted respectively $\mathbf{BoW}_{\mathrm{2D}}$ and $\mathbf{BoW}_{\mathrm{3D}}$.

d. The direct feature-level fusion and decision-level fusion are different in the fourth step. For the decision-level fusion, the BoW-based feature vectors $\mathbf{BoW}_{\mathrm{2D}}$ and $\mathbf{BoW}_{\mathrm{3D}}$ are used to train the 1V1 SVM classifiers respectively. And for the direct feature-level fusion, $\mathbf{BoW}_{\mathrm{2D}}$ and $\mathbf{BoW}_{\mathrm{3D}}$ are combined to gather a 1×600 vector $\mathbf{BoW}_{\mathrm{2D,3D}}[\mathbf{BoW}_{\mathrm{2D}}, \mathbf{BoW}_{\mathrm{3D}}]$. Then the $\mathbf{BoW}_{\mathrm{2D,3D}}$ is used to train the 1V1 SVM classifiers.

3.2 The Testing Phase

This testing phase consists of the following steps:

a. For any object to classify, we extract its 2D and 3D SIFT descriptors associated with each keypoint.

b. From the N_{2D} 2D SIFT descriptors of size 1×128, we compute 1×300 BoW feature vectors $\mathbf{BoW}_{\mathrm{2D}}$, and from the N_{3D} 3D SIFT descriptors of size 1×128, we compute 1×300 BoW feature vectors $\mathbf{BoW}_{\mathrm{3D}}$ thanks to the BoW model representation [19].

c. The direct feature-level fusion is done by stacking the BoW-based feature vectors $\mathbf{BoW}_{\mathrm{2D}}$ and $\mathbf{BoW}_{\mathrm{3D}}$ to get a 1×600 vector $\mathbf{BoW}_{\mathrm{2D,3D}}[\mathbf{BoW}_{\mathrm{2D}}, \mathbf{BoW}_{\mathrm{3D}}]$.

d. The direct feature-level fusion/the decision-level fusion is carried out.

e. The object is associated with the category (or class) having the largest probability, *i.e.*:

$$\mathrm{Class(Object)} = \arg \max_{1 \le i \le n} \{P(i)\} \tag{7}$$

4. Simulation Results

4.1 The Experimental Set-up

We evaluate the proposed recognition algorithm on a large-scale multi-view object data set collected using an RGB-D camera [22]. This data set contains colour, depth images, and point clouds of 300 physically distinct everyday objects taken from different viewpoints. The objects belong to 1 of 51 categories and contain three viewpoints. To test the recognition ability of our features, we test category recognition on objects that were not present in the training set. At each trial, we randomly choose one test object from each category and train classifiers on the remaining objects. Randomly choose 100 training samples and 60 test samples for each category. The object recognition rate (ORR) is calculated using (8):

$$ORR = n_r/N \tag{8}$$

where n_r is the number of objects correctly recognized and N is the total number of test samples.

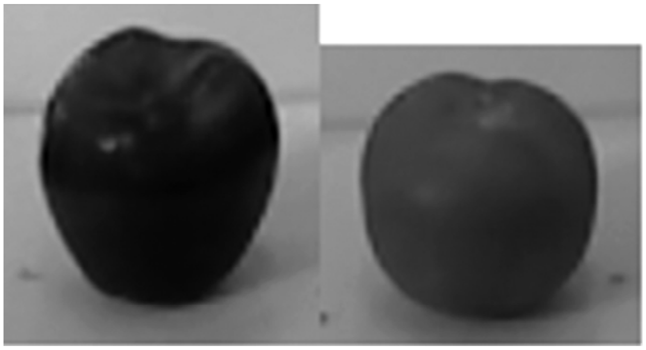

Figure 3. Apple and tomato.

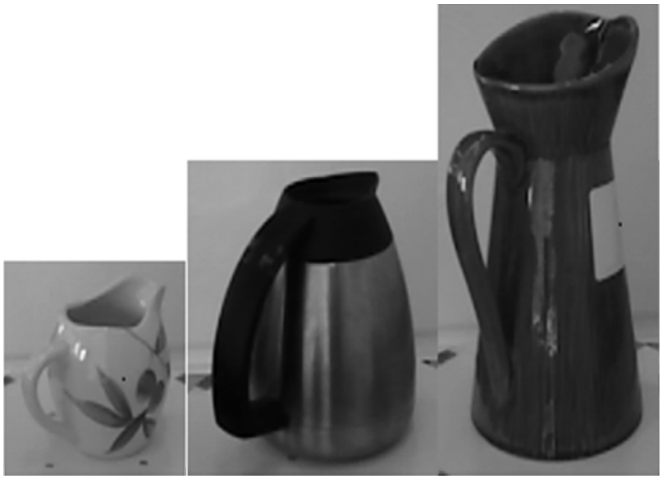

Figure 4. Pitchers.

4.2 Experiment Results and Analysis

4.2.1 Robustness to Intra-class Variation and Inter-class Similarities

In this study, we compare the ORR performances in different classes having high similarity (*e.g.*, apple and tomato) and in the same class but having strong variation (*e.g.*, pitcher object) as in Figs. 3 and 4. We evaluate the accuracy of point feature histograms RGB (PFHRGB), 2D SIFT, 3D SIFT, and the feature-level fusion of 2D and 3D SIFT under the same conditions. Training and testing samples are the same as in the first experiment. Our simulation results are given in Table 1.

Summarized in Table 1, using 3D SIFT increases the ORR of 3.05% with respect to 2D SIFT. This shows that introduction of the depth information improves the quality of object recognition. Three different objects of the pitcher class are illustrated in Fig. 4, where there are great differences within such class. 3D SIFT achieves ORR with 96.67% accuracy, much superior to the 70% obtained by PFHRGB. *Apple* and *tomato* displayed in Fig. 3 look highly similar even if they belong to two distinct classes. 3D SIFT provides much better ORR than other descriptors. As given in Table 1, our GOR method based on feature-level fusion of 2D and 3D SIFT offers better robustness to intra-class variations and inter-class similarities, and 3D SIFT reaches higher accuracy than other single descriptors.

4.2.2 Robustness to Angle of View

In this experiment, we evaluate the performance of our GOR method when applied under different observation conditions, more precisely when the objects are observed under three very distinct angles of view (30, 45 and 60 deg). Training samples are the same as Experiment 4.2.1. By randomly selecting 60 objects from each view to be as the test samples, for each view, there are 360 test samples from six categories. The experimental results are illustrated in Fig. 5.

From Fig. 5, it is observed that ORR with 3D SIFT is relatively accurate and stable in comparison with PFHRGB descriptor. The direct feature-level fusion strategy (with $ORR > 90\%$) offers far better ORR than using the best

Table 1
ORR (in %) of Different Classes

Feature Descriptor	PFHRGB	2D SIFT	3D SIFT	2D+3D SIFT
ORR (apple)	61.67	53.33	71.67	65.00
ORR (tomato)	100	98.33	91.67	100
ORR (banana)	91.67	93.33	93.33	100
ORR (pitcher)	70.00	95.00	96.67	98.33
ORR (cereal_box)	91.67	98.33	95.00	95.00
ORR (kleenex)	90.00	90.00	100	100
Averaged ORR	84.17	88.05	91.39	93.06

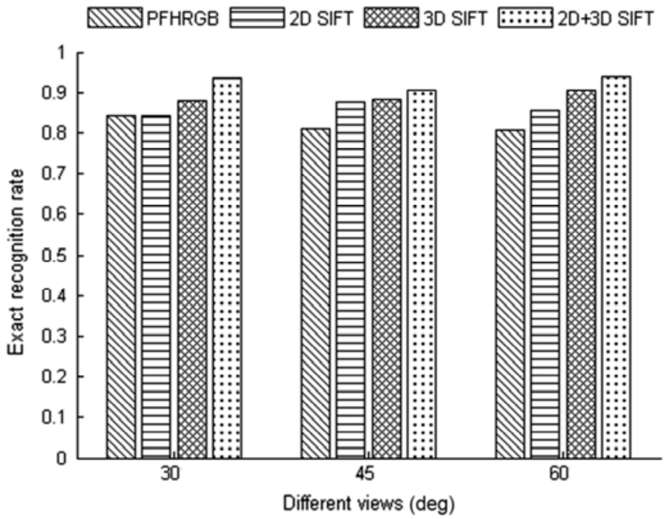

Figure 5. ORR performances under three angles of view.

Table 2
Averaged ORR (in %) for Different Zoomings

Feature Descriptor	PFHRGB	2D SIFT	3D SIFT	2D+3D SIFT
ORR (no zoom)	84.17	88.06	91.11	93.06
ORR (zoom = 1/2)	74.44	77.50	76.67	82.78
ORR (zoom = 1/3)	63.33	64.17	65.28	68.89
ORR (zoom = 1/4)	61.39	46.94	61.67	63.05

single descriptor, indicating that the combination of 2D and 3D SIFT is effective and robust for category recognition even under largely distinct angles of view.

4.2.3 Robustness to Size Scaling

The training samples are the same as in the first experiment. To evaluate the robustness of our method to size scaling (zooming), the test samples are zoomed out to 1/2, 1/3, and 1/4, which are respectively listed in Table 2.

It demonstrates in Table 2 that our GOR method with fusion is superior to the algorithm based on single descriptor. However, the ORR of each feature descriptor decreases. Especially when zoomed to 1/4, the accuracy of ORR with 2D SIFT is only 46.94%. The main reason is that part of the images, such as *apple* (whose original size is only 84×82) after scaling, reduces the number of useful keypoints. The feature-level fusion algorithm still provides an averaged ORR of 63.05%.

5. Conclusion

In this paper we have proposed a new GOR method for robot perception based on 2D and 3D SIFT descriptors that calculate multiple feature vectors combined with different strategies, and feed SVM classifier for making object recognition. Our performance evaluation based on open real data sets has demonstrated the superiority of our new 3D SIFT descriptor adapted for point cloud with respect to the existing 3D features such as PFHRGB. Our GOR method based on feature fusion of 2D and 3D SIFT functions better than the one using best single feature. As future research work, we would like to reduce the computational time needed for feature extraction and description by realizing GPU-based implementation while allowing still good recognition rate and explore more feature fusion strategies to improve recognition performances.

Acknowledgement

This work was supported by NNSF of China (61175091, 61573097), Qing Lan Project of Jiangsu Province, Aeronautical Science Foundation of China (20140169002), and Six Major Top-talent Plan of Jiangsu Province.

References

[1] Y. Lei, M. Bennamoun, M. Hayat, and Y. Guo An efficient 3D face recognition approach using local geometrical signatures, *Pattern Recognition*, *47*(2), 2014, 509–524.

[2] X.D. Li, X. Zhang, B. Zhu, and X.Z. Dai, A visual navigation method for robot based on a GOR and GPU algorithm, *Robot*, *34*(4), 2012, 466–475 (in Chinese).

[3] S. Allaire, J.J. Kim, S.L. Breen, D.A. Jaffray, *et al.* Full orientation invariance and improved feature selectivity of 3D SIFT with application to medical image analysis, *Proc. IEEE CVPR Workshops*, Anchorage, AK, 2008, 23–28.

[4] Y. Wakuda, K. Sekiyama, and T. Fukuda, Dynamic event interpretation and description from visual scene based on cognitive ontology for recognition by a robot, *International Journal of Robotics & Automation*, *24*(3), 2009, 263–279.

[5] Z. Hamici, Real-time pattern recognition using circular cross-correlation: A robot vision system, *International Journal of Robotics & Automation*, *21*(3), 2006, 174–183.

[6] D.G. Lowe, Object recognition from local scale-invariant features, *Proceedings of the IEEE CCV Conference*, *2*(September), 1999, 1150–1157.

[7] D.G. Lowe, Distinctive image features from scale-invariant key points, *International Journal of Computer Vision*, *60*(2), 2004, 91–110.

[8] P. Scovanner, S. Ali, and M. Shah, A 3-dimensional SIFT descriptor and its application to action recognition, *Proc. 15th ACM MM Conf.*, Augsburg, Germany, 2007, 357–360.

[9] W. Cheung and G. Hamarneh, N-SIFT: N-dimensional scale invariant feature transform for matching medical images, *Proc. 4th IEEE Int. Symp. on Biomedical Imaging*, Arlington, VA, 2007, 720–723.

[10] R.N. Dalvi, I. Hacihaliloglu, and R. Abugharbieh, 3D ultrasound volume stitching using phase symmetry and Harris corner detection for orthopaedic applications, *Proc. SPIE*, Vol. 7623 (Medical Imaging, 2010), San Diego, CA, 2010.

[11] M. Niemeijer, *et al.*, Registration of 3D spectral OCT volumes using 3D SIFT feature point matching, *Proc. SPIE*, Vol. 7259 (Medical Imaging, 2009), Lake Buena Vista, FL, 2009.

[12] G.T. Flitton, T.P. Breckon, and N. Megherbi, Object recognition using 3D SIFT in complex CT volumes, *Proc. BMV Conf.*, Aberystwyth, UK, 2010, 1–12.

[13] R.B. Rusu, N. Blodow, Z.C. Marton, and M. Beetz, Aligning point cloud views using persistent feature histograms, *Proc. IEEE/RSJ Int. Conf. on Intelligent Robots and Systems*, Nice, France, 2008, 3384–3391.

[14] R.B. Rusu, N. Blodow, and M. Beetz, Fast point feature histograms (FPFH) for 3D registration, *Proc. IEEE Int. Conf. on Robotics and Automation*, Kobe, Japan, 2009, 3212–3217.

[15] S. Lazebnik, C. Schmid, and J. Ponce, A sparse texture representation using local affine regions, *IEEE Transactions on PAMI*, *27*(8), 2005, 1265–1278.

[16] SIFT demo program (Version 4, July 2005), Beijing, China, [Online], http://www.cs.ubc.ca/lowe/keypoints/.

[17] R. Hess, An open source SIFT library, *ACM MM*, 2010 [Online], http://robwhess.github.io/opensift/

[18] R.B. Rusu and S. Cousins, 3D is here: Point cloud library (PCL), *Proc. IEEE Int. Conf. on Robotics and Automation*, Shanghai, China, 2011, 1–4.

[19] J. Sivic and A. Zisserman, Video Google: A text retrieval approach to objects matching in videos, *Proc. 9th CCV Conf.*, 2003, 1470–1478.

[20] D. Arthur and S. Vassilvitskii, k-means++: The advantages of careful seeding, *Proc. SODA '07*, Philadelphia, PA, USA, 2007, 1027–1035.

[21] B.E. Boser, I.M. Guyon, and V.N. Vapnik, A training algorithm for optimal margin classifiers, *Proc. the 5th ACM Workshop on Computer Learning Theory*, Pittsburgh, PA, 1992, 144–152.

[22] K. Lai, L.-F. Bo, X.F. Ren, and D. Fox, A large-scale hierarchical multi-view RGB-D object dataset, *Proc. IEEE Int. Conf. on Robotics and Automation*, Shanghai, China, 2011, 1817–1824.

Biographies

Xinde Li earned his Ph.D. in Control Theory and Control Engineering, from Department of Control Science and Engineering, Huazhong University of Science and Technology (HUST), Wuhan, China, in 2007. Afterwards, he joined School of Automation, Southeast University, Nanjing, Jiangsu, China, where he is currently an Associate Professor and Ph.D. Supervisor. His research interests include information fusion, object recognition, computer vision, intelligent robot, and human–robot interaction. He has published more than 60 SCI and EI papers and holds 10 national patents. He was granted a "Talent of Qing Lan Project" award of Jiangsu province and a "Six Major Top-talent Plan" award of Jiangsu province, China.

Chaomin Luo received Ph.D. in Electrical and Computer Engineering from University of Waterloo, Canada, in 2008, where he was awarded Postgraduate Scholarship (PGS) from Natural Sciences and Engineering Research Council (NSERC) of Canada; received the Best Student Paper Presentation Award at 2007 SWORD'2007, and was a recipient of 2003–2005 Graduate Incentive Award and 2005–2006 President's Graduate Scholarship. He earned M.Sc. in Engineering Systems and Computing at University of Guelph, Canada, and B.Eng. in Radio Engineering from Southeast University, China. After receiving Ph.D., he was Assistant Professor in Graduate Institute of Electrical Engineering, National Taipei University, in 2008. He is currently an Associate Professor in Advanced Mobility Lab, Department of Electrical and Computer Engineering, at University of Detroit Mercy, Michigan, USA. His research interests include robotics and automation, intelligent systems, computational intelligence, mechatronics, VLSI, and embedded systems. Dr. Luo's extensive industry experience in embedded systems, intelligent instrument, automation, control and mechatronics includes working as an Electronics Engineer, Hardware Designer and a Director of the Embedded Systems and Intelligent Instrument Lab. Dr. Luo was Publicity Chair in 2011 IEEE International Conference on Automation and Logistics. He was on Conference Committee in 2012 International Conference on Information and Automation and International Symposium on Biomedical Engineering and Publicity Chair in 2012 IEEE International Conference on Automation and Logistics. He is currently Chair of IEEE SEM – Computational Intelligence Chapter and Chair of Education Committee of IEEE SEM, and Vice Chair of IEEE SEM – Robotics and Automation Chapter. He serves as Editorial Board Member of International Journal of Complex Systems – Computing, Sensing and Control. He has organized and chaired several special sessions on topics of Intelligent Vehicle Systems and Bio-inspired Intelligence for Robotics in IEEE reputed international conferences such as IEEE-IJCNN, IEEE-SSCI, and IEEE-ICIA.

Jean Dezert was born in l'Hay les Roses, France, in August 1962. He received his doctor degree from the University Paris XI, Orsay, France, in 1990, in Automatic Control and Signal Processing. Since 1993, he is senior research scientist in the Information Modeling and Processing Dept. at the French Aerospace Lab. His current research interests include estimation theory, and information fusion (IF) and plausible reasoning with applications to MS-MTT, defence and security, robotics, and risk assessment. He has published three books and more than a hundred of papers in conferences and journals on tracking and information fusion. He is the co-founder with Florentin Smarandache of DSmT (Dezert-Smarandache Theory) of information fusion based on belief functions.

Yingzi Tan earned her Ph.D. in Thermal Engineering, from Department of Power Engineering, Southeast University, Nanjing, Jiangsu, China, in 2002. Afterwards, she joined School of Automation, Southeast University, where she is currently an Associate Professor and Master Supervisor. Her research interests include intelligent robot, object recognition, multi-agent system, and biped robot.