Journal of Real-Time Image Processing
DOI 10.1007/s11554-013-0380-z

Wei Wang · Lili Chen · Ziyuan Liu · Kolja Kühnlenz · Darius Burschka

Textured/textureless object recognition and pose estimation using RGB-D image

Received: 13 May 2013/ Accepted: 29 October 2013

Abstract In this paper, we propose a novel global object descriptor, so called *Viewpoint oriented Color-Shape Histogram* (VCSH), which combines 3D object's color and shape features. The descriptor is efficiently used in a real-time textured/textureless object recognition and 6D pose estimation system, while also applied for object localization in a coherent semantic map. We build the object model firstly by registering from multi-view color point clouds, and generate partial-view object color point clouds from different synthetic viewpoints. Thereafter, the extracted color and shape features are correlated as a VCSH to represent the corresponding object patch data. For object recognition, the object can be identified and its initial pose is estimated through matching within our built database. Afterwards the object pose can be optimized by utilizing an iterative closest point strategy. Therefore, all the objects in the observed area are finally recognized and their corresponding accurate poses are retrieved. We validate our approach through a large number of experiments, including daily complex scenarios and indoor semantic mapping. Our method is proven to be efficient by guaranteeing high object recognition rate, accurate pose estimation result as well as exhibiting the capability of dealing with environmental illumination changes.

Keywords Real-time robotic vision · Object recognition and pose estimation · Viewpoint oriented color-shape histogram · Semantic map

W. Wang · L. Chen · D. Burschka
Institute of Robotics and Embedded Systems, Technische Universität München, D-85748 Garching bei München, Germany. E-mail: {wei.wang, lili.chen, burschka}@in.tum.de

Z. Liu
Institute of Automatic Control Engineering, Technische Universität München, D-80290 München, Germany. E-mail: ziyuan.liu@tum.de

K. Kühnlenz
Institute of Advanced Study, Technische Universität München, D-80333 München, Germany. E-mail: koku@tum.de

1 Introduction

To interact with autonomous mobile robots in unstructured environments, it is essential for a robot to successfully recognize objects, estimate its accurate pose and perform high-level tasks in real time. Therefore, object recognition and pose estimation plays a crucial role in a wide range of robotics applications, and is at the heart of high-level tasks such as object localization for semantic mapping. However, due to large invariance in respect to object size, position, and its viewpoint, heavily cluttered environment, occlusions in the scene, it is a greatly challenging problem [1–5].

Some previous approaches have been developed to address the challenges mentioned above. Among those approaches, an efficient object descriptor plays a most critical role. There is a large variety of object descriptors using diversified features. For 2D images, SIFT [6], SURF [7] and HOG [8] are the most popular features which can be extracted based on object's photometric properties (texture). Apart from the gray-scale features, color-based features are also widely proposed for object recognition [9–12]. However, the photometric features have the limitation of being not able to cover all potential poses in 3D space. While for 3D depth images, a wide variety of geometric quantities such as local patches [13], local moments [14], volume [15], polygon surface [16], spherical harmonics [17], contour [18] and edge [19] try to emulate comparable features, in order to be used for geometric descriptors. However, these geometric features only describe 3D object's shape primitives while ignoring the photometric information on the object surface.

With the massively increasing usage of new-released RGB-D sensors such as Kinect and stereo cameras, which can provide both photometric and geometrical information, multi-dimensional photometric and geometrical feature based object descriptor gets to be powerful alternative for object recognition and pose estimation by utilizing such sensors, as have been considered in the works [20–24]. Furthermore, an object should be recognized

whatever pose it is (scale and rotation invariant), thus the viewpoint component is also necessary to be integrated into the object descriptor building [25–30].

Inspired by above, we propose a novel object descriptor efficiently combining object's color and shape features within a textured/textureless object recognition and 6D pose estimation system. The main contributions of this paper thereby include:

- A novel object descriptor *Viewpoint oriented Color-Shape Histogram* combining color and shape features, as well as object viewpoint component;
- A real-time object recognition and pose estimation system which gives high recognition rate and accurate 6D pose recovery under various unstructured environment;
- 3D object recognition and localization for the coherent semantic mapping;
- Performance evaluation on object recognition rate, pose accuracy and stability analysis with respect to illumination changes;
- Live demonstrations and state-of-the-art comparisons;

Parts of our system have been previously described in [43], this paper presents more comprehensive descriptions as well as significant extensions and enhancements, then applying to various highly challenging scenarios as well as coherent semantic mapping.

The remainder of this paper is organized as follows: Section 2 reviews the state of the art and related work. Section 3 provides detailed description of VCSH, its integration within object recognition and pose estimation system and object localization in semantic map. The experimental results including the pose accuracy evaluation, stability analysis with illumination and runtime performance are presented in Section 4. Finally, Section 5 summarizes the paper and proposes future development roads.

2 Related Work

A large variety of approaches have been proposed for object recognition and pose estimation. Within those approaches, the key role - object descriptors could be mainly classified into two categories: global descriptor and local one. Global object descriptor extracts features from the well segmented and clustered object data [20, 29, 32]. The object needs to be well clustered and it is sensitive to partial occlusions. Instead, the local descriptor is based on pair-to-pair feature matching from real-scene data which causes high computational cost for final recognition and pose recovery [21, 23, 24, 31, 33].

More specifically, for global object descriptors, VFH [29] as an extension of FPFH [31], integrates the viewpoint variant component into the 3D geometrical features. However, it neither allows for object's full pose estimation nor considers texture or color feature. Wohlkinger et al. proposes a global 3D descriptor ESF (Ensemble of Shape Functions) [32], which creates the database by generating synthetic views through CAD object models. The combination of angle, point distance and area shape functions are applied on randomly selected point pairs, while local distribution features are accumulated into global descriptor. Nevertheless, ESF neglects the object's photometric information, thus being not able to accurately provide pose estimation. Tang et al. [20] directly uses the Naive Bayes matching method for object recognition and pose recovery. The object global hue value histogram is generated from the complete mesh object model as object's color feature. Combining with the extracted 3D SIFT from object's texture, the object can be recognized and its pose can be estimated. However, this approach needs the detailed mesh model for training and these objects are restricted to be fully textured.

For local object descriptors, SHOT (Signature of Histograms of Orientations) divides the spherical volume around one point into spherical grids based on the local reference frame [24]. Normal of each point falling into a certain grid is compared with normal of centering point. The angle relationship is counted and represented as histogram on each grid which are then concatenated as a descriptor. CSHOT as an extention of SHOT that adds color information to construct descriptor is presented in [33]. This method relies on a local reference frame, but the reference frame could not be stably estimated for object with rotational symmetry such as basketball. A real-time object recognition system [21] is proposed using ConVOSCH object descriptor which correlates the geometrical with visual RGB data, but the object's accurate pose is not possible to be recovered. Choi et al. [23] defines a local object color point pair feature descriptor, which is represented as a hash table combining the geometrical and HSV color information. However, the color information is only utilized for pruning potential false matches while not considered as a general object descriptor for recognition and pose estimation. Moreover this approach produces high computational cost and is sensitive to high dimensional parameter settings respective to different scenes.

3 Proposed Approach

In this section, we provide details on the design of our VCSH descriptor, and how it is integrated into a object recognition and pose estimation system, and finally show that it is efficiently applied for object localization in semantic mapping.

The framework of our proposed approach is illustrated in Figure 1. During the offline training phase, we first build the complete 3D object model by registering all the object's RGB-D data of different poses into a single coordinate frame. By using the centroid of object model as origin, we generate a sphere with a certain radius. On the surface of this sphere, a big mount of

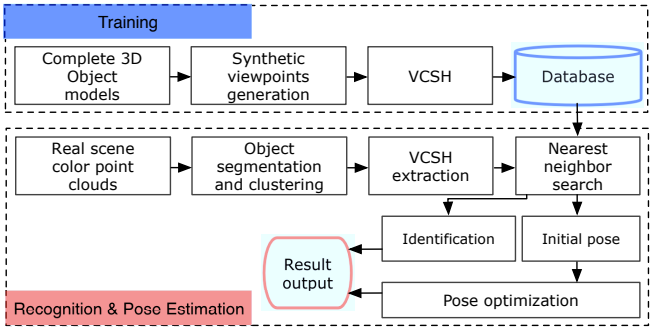


Fig. 1 Overview of real-time textured/textureless object recognition and pose estimation system using viewpoint oriented color-shape histogram descriptor.

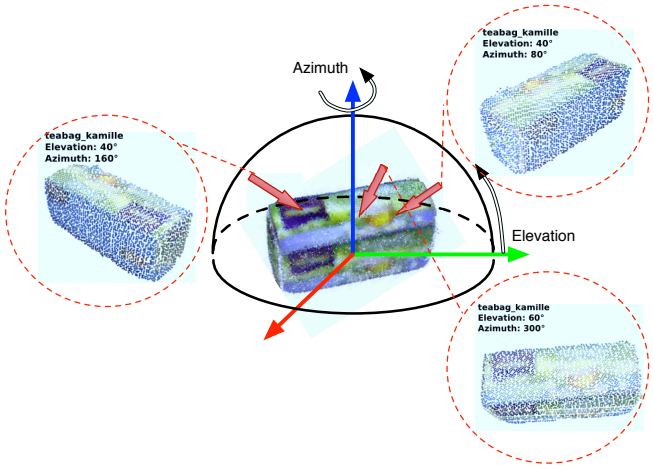


Fig. 2 Sampling the synthetic viewpoints in the upper hemisphere for object patch data generation: Red vertices represent the virtual camera viewpoints and the red circles illustrate some generated data from synthetic viewpoints.

viewpoints are homogeneously generated with their direction pointing to the sphere origin. Using each of these generated viewpoints, an object patch data which represents the object identification and the corresponding viewpoint pose, is generated. Subsequently VCSH can be computed as a global object descriptor for each object patch data, within which the color and shape information of all points is used for the descriptor generation. Consequently, an object is represented by the generated VCSH set and stored into the database. During the online recognition and pose estimation phase, the object data is segmented and clustered from the real world scene, and we compute its corresponding VCSH. Thereafter, the closest hypothesis is retrieved from our generated descriptor database by nearest neighbor searching, with outputting object identification and its initial pose. Finally, the recognized object's accurate 6D pose can be estimated through pose optimization and verification step. In addition, the object recognition and pose estimation system is applied into the coherent semantic map, for the robotic exploration in large-scale map and for further object manipulation. Next we explain the corresponding parts that are involved in greater details.

3.1 Building 3D Object Model

Our proposed object model building platform consists of a rotatable plane and a stationary Kinect sensor. After segmentation from the plane and Euclidean distance-based clustering, object color point cloud data $\{O_f\}$ for each single view and its transform $\{TF_f\}$ relative to the initial frame O_0 are captured, where $f = \{0 \cdots F\}$ is the frame index. By registering $\{O_f\}$ with $\{TF_f\}$ into a single object coordinate, the whole 3D model Ω then can be generated as a cluster of color point cloud,

$$\Omega = O_0 \cup TF_1^{-1} \cdot O_1 \cup \cdots \cup TF_F^{-1} \cdot O_F. \quad (1)$$

In order to eliminate noises, the Moving Least Squares (MLS) algorithm [36] is utilized to smooth the whole 3D model. Note that the detailed object mesh model and surface texture information are not necessary here.

3.2 Synthetic Viewpoints Generation

For each object model Ω_i, $i = \{1...I\}$, we generate J object patch data M_j with synthetic viewpoint VP_j where $j = \{1 \cdots J\}$. Note that the viewpoint is the sensor's view direction relative to the object. As the view direction should be considered to cover object's potential 6D poses, the synthetic viewpoints are therefore generated on a half sphere surface, with the origin being the object model's centroid. The synthetic viewpoint position is generated on the sphere surface in elevation and azimuth direction homogeneously. And its direction is pointing to the sphere's origin. With the generated synthetic viewpoint VP_j, object patch data M_j could be generated according to VP_j from the whole 3D object model Ω by using ray-casting method, as illustrated in Figure 2. A pseudocode is also given out in Algorithm 1.

It is necessary to mention that the object model Ω is not only restricted to the raw color point cloud model as in our platform, but also potential for CAD models.

Subsequently, a global object descriptor is needed to describe each M_j with its viewpoint VP_j for object recognition and 6D pose recovery.

3.3 Viewpoint oriented Color-Shape Histogram

For recognition and pose recovery for objects in our daily life, an object descriptor which consists of both color and shape information is prerequisite. In particular, this descriptor should be able to differentiate the objects which have same shape but different colors and also could deal with textured/textureless objects. In order to fulfill aforementioned requirements, a novel object descriptor *viewpoint oriented color-shape histogram* is proposed here based on both color and shape features of an object. During VCSH construction, firstly the color of each point p in

Algorithm 1 Object patch data generation using sampled synthetic viewpoint

Ω; //whole 3D object model
M; //generated object patch data
VP; //related synthetic viewpoint
ε; //threshold for point in a line
for$(iter = 0; iter < \Omega.size; iter++)$**{**
  $p = \Omega_{iter}$; //point in Ω
  $L = line3D(VP, p)$;
  $Flag(false)$; //flag of occluded
  for$(iter1 = 0; iter1 < \Omega.size; iter1++)$**{**
    if$(iter1 \neq iter)$
      $p^{\star} = \Omega_{iter1}$; //another point in Ω
      if$(dist(p^{\star}, L) < \varepsilon$ && $\|VP - p^{\star}\| < \|VP - p\|)$
    //point in line and closer to viewpoint (occluded);
        $Flag = true$;
        break;
  };
  if$(Flag == false)$*push p into M*;

object patch data M_j is smoothly ranged and color distributions for different ranges are estimated. Secondly, the shape features are estimated to describe each point's geometrical relationship with the viewpoint VP_j and the M_j's centroid c. Finally, the extracted color and shape feature are correlated and built as VCSH to describe each object patch data M_j.

3.3.1 Smoothed Color Ranging

To represent the uniqueness of color feature for each object patch data, the feature needs to be characterized and color distributions for different ranges should be estimated according to their color values. HSV color space is employed here for better characterizing each point's color feature, due to its robustness to illumination changes [10]. As shown in Figure 3, there are chromatic and achromatic areas in SV space, in which the chromatic area represents the true color space while achromatic area represents the gray scale space. That is, the histogram is divided into 8 regions as RE_u with the index of $u = \{0 \cdots 7\}$, in which six are for chromatic area, and the other two are for achromatic area [39].

To be more detailed, firstly, we consider the six true color histogram regions RE_0 to RE_5, which represents six typical colors CR_0 to CR_5. Each point's hue value then can be quantized into a certain color region CR. However, the hard quantization can not represent the true color correctly. To overcome this issue, a smoothed ranging method is proposed, by estimating two distributions w_H for two consecutive histogram regions RE in true color space. The detailed steps are presented as follows:

– Identify CR_n: red as $CR_0 = 0$, yellow as $CR_1 = 60$, green as $CR_2 = 120$, cyan as $CR_3 = 180$, blue as $CR_4 = 240$, purple as $CR_5 = 300$. Consequently, six histogram ranges are divided based on the color index CR, as $RE_u \rightarrow CR_n$ where $u = n = \{0 \cdots 5\}$.

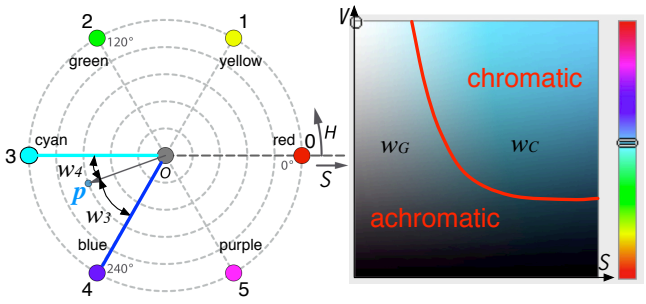


Fig. 3 Left: smoothed color range and estimation of contributions for adjacent regions in HS space. Right: illustration for the chromatic and achromatic areas in SV space.

– For each color point p, its hue value H is ranged into two consecutive histogram regions RE_u and RE_{u+1} as $u = \lfloor H/60 \rfloor$, if $u = 5$, the next histogram region RE_{u+1} would be reset to RE_0.
– Estimate color distributions w_{H_u}, $w_{H_{u+1}}$ according to the ranged adjacent regions RE_u, RE_{u+1} in true color space, based on the distance from hue value H to CR_n and CR_{n+1}:

$$w_{H_u} = (H - CR_{n+1})/60, \; w_{H_{u+1}} = 1 - w_{H_u}. \tag{2}$$

Secondly, we consider the achromatic area which consists of two histogram regions RE_6 and RE_7. When one of the saturation S and value V is near 0 in HSV space, the point color will be represented as gray scale. Since the color in achromatic space is highly sensitive to illumination changes, the previous estimated distributions w_{H_n} and $w_{H_{n+1}}$ in true color space should be redesigned according to the influence from S and V. In order to capture the nature color, a soft decision method [34] is employed and we update both chromatic and achromatic components of the histogram. The weight w_C of chromatic and w_G of achromatic component is determined by S, V, and their sum equals unity:

$$w_C = S^{r(1/V)^{r_1}}, \; w_G = 1 - w_C, \tag{3}$$

where $r, r_1 \in [0, 1]$. To give best precision on true color, $r = 0.14$ and $r_1 = 0.9$ are chosen empirically. Furthermore, V is quantized while distribution w_6 and w_7 are calculated for regions RE_6 and RE_7: $w_6 = w_G$ if $V < 0.5$, otherwise $w_6 = 0$; while the value of w_7 is converse.

We therefore update all the previous estimated color distributions as w_u and w_{u+1}, by considering the chromatic weight w_C's influence on true color representation.

$$w_u = w_{H_u} \cdot w_C, \; w_{u+1} = w_{H_{u+1}} \cdot w_C. \tag{4}$$

Finally, each point p with HSV color value is ranged into three histogram regions $\langle RE_u, RE_{u+1}, RE_6 | RE_7 \rangle$ with respective contributions being $\langle w_u, w_{u+1}, w_6 | w_7 \rangle$.

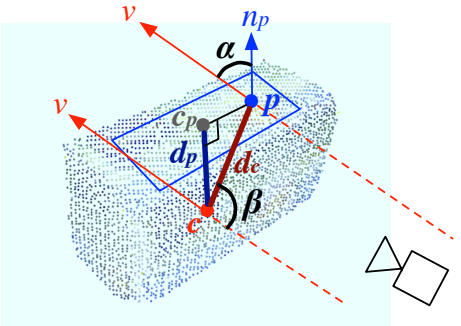


Fig. 4 Shape features of point p. c is the centroid of object patch data. n_p is the normal of p. v is the synthetic viewpoint direction. c_p is c's projection point on the tangent plane of p (blue rectangle frame). d_c and d_p are the distances from c to p and from c to c_p. α is the angle between v and n_p, and β is the angle between v and the line segment cp.

3.3.2 Shape Feature Extraction

After the color contributions have been estimated for the specific histogram regions, it is necessary to extract each object patch data M's shape features $F = \{f_0 \cdots f_m\}$ for the final histogram building, where m is the point number in M. With object patch data M representing the partial data of the object from viewpoint VP, each point p's geometrical information should be extracted in order to describe the object shape accurately and robustly. Partly inspired by the work in [35], we extract the shape features depending on point p's relationship with the centroid of M and viewpoint VP. As a global descriptor, the surface normal n_p of each point p in M and the centroid c of M are computed at first. The relationship of p and c represents the 3D shape of the object cluster. The relationship of p and VP indicates the rotation of the object cluster relative to the sensor direction. Note that VP and c are represented as the object's 6D pose.

As shown in Figure 4, the tangent plane of p is defined as a plane that is orthogonal to p's normal n_p. The centroid c is projected to this tangent plane as a point c_p. A four dimensional geometrical feature f consists of two distances and two angles components $\langle d_p, d_c, \alpha, \beta \rangle$, which are calculated as:

$$\begin{aligned} d_p = \|p - c\|\,,\ d_c = \|c_p - c\|\,, \\ \alpha = \arccos(n_p \cdot (p - c)),\ \beta = \arccos(v \cdot (p - c)). \end{aligned} \tag{5}$$

In object partial data M, each point p's feature f is calculated. Therefore, for single object model O which contains J object patch data, the final feature set is $F = \{f_0 \cdots f_m\}$ with m points, representing the object's shape from a certain viewpoint VP_j.

3.3.3 Color and Shape Feature Correlation

To describe an object patch data M with the viewpoint VP discriminatively and comprehensively as a histogram, the VCSH descriptor should be correlated with these two different features. In the smoothed color ranging phase, the whole histogram has been segmented into eight regions. Every component in each point's shape feature f has 30 bins, therefore each RE contains 120 bins inside. Each p's two distance components $\langle d_p, d_c \rangle$ are indexed as $\langle IN_{d_p}, IN_{d_c} \rangle$ by the quantization using their values scaling from minimum value $\langle d_{p_{min}}, d_{c_{min}} \rangle$ to maximum value $\langle d_{p_{max}}, d_{c_{max}} \rangle$. Two angle components $\langle \alpha, \beta \rangle$ are indexed as $\langle IN_\alpha, IN_\beta \rangle$ by the quantization using their values with the range of 0 to 90°as follows:

$$\begin{aligned} IN_{d_p} = \lfloor \frac{30 \cdot (d_p - d_{p_{min}})}{d_{p_{max}} - d_{p_{min}}} \rfloor, IN_{d_c} = \lfloor \frac{30 \cdot (d_c - d_{c_{min}})}{d_{c_{max}} - d_{c_{min}}} \rfloor, \\ IN_\alpha = \lfloor \frac{\alpha}{90} \cdot 30 \rfloor, IN_\beta = \lfloor \frac{\beta}{90} \cdot 30 \rfloor. \end{aligned} \tag{6}$$

During the object's color and shape features correlation step, each p's color contributions as $\langle w_u, w_{u+1}, w_6 | w_7 \rangle$ for three histogram regions $\langle RE_u,\ RE_{u+1},\ RE_6 | RE_7\ \rangle$ are incrementally added into $\langle INX_{d_p}, INX_{d_c}, INX_\alpha, INX_\beta \rangle$. The final certain bins index INX in VCSH regarding to each of these three $RE_m, m \in [u, u+1, 6, 7]$ are quantized as follows:

$$\begin{aligned} INX_{d_p} &= IN_{d_p} + 120 \cdot m, \\ INX_{d_c} &= IN_{d_c} + 120 \cdot m + 30, \\ INX_\alpha &= IN_\alpha + 120 \cdot m + 30 \cdot 2, \\ INX_\beta &= IN_\beta + 120 \cdot m + 30 \cdot 3. \end{aligned} \tag{7}$$

The whole histogram has incremental value corresponding to color contributions from all the points in M. During final object recognition phase, the object's descriptor should not change with varying distance at same view direction. However the histogram's absolute value of each bin will change according to the object cluster point number. To overcome this problem, the values of histogram are finally normalized with point number. Thus, VCSH could be viewed as a geometrical constrained color feature histogram. As shown in Figure 5, color contributions of all points in object patch data respected to different viewpoints are incrementally added into the certain indexes of whole VCSH, based on smoothed color ranging and shape feature extraction. An example of two picked points in object patch data for the final VCSH generation is illustrated within Figure 5, with the step of color-shape features extraction and correlation step. Each patch data of object could be represented as one VCSH. The final correlated histogram has $(6 + 2) \times (30 \times 4) = 960$ dimensions. The computational complexity of VCSH is $O(n)$, where n is the point number of object patch data M. Consequently, the final generated histogram gives the possibility for high successful object recognition rate, accurate pose estimation and real-time processing.

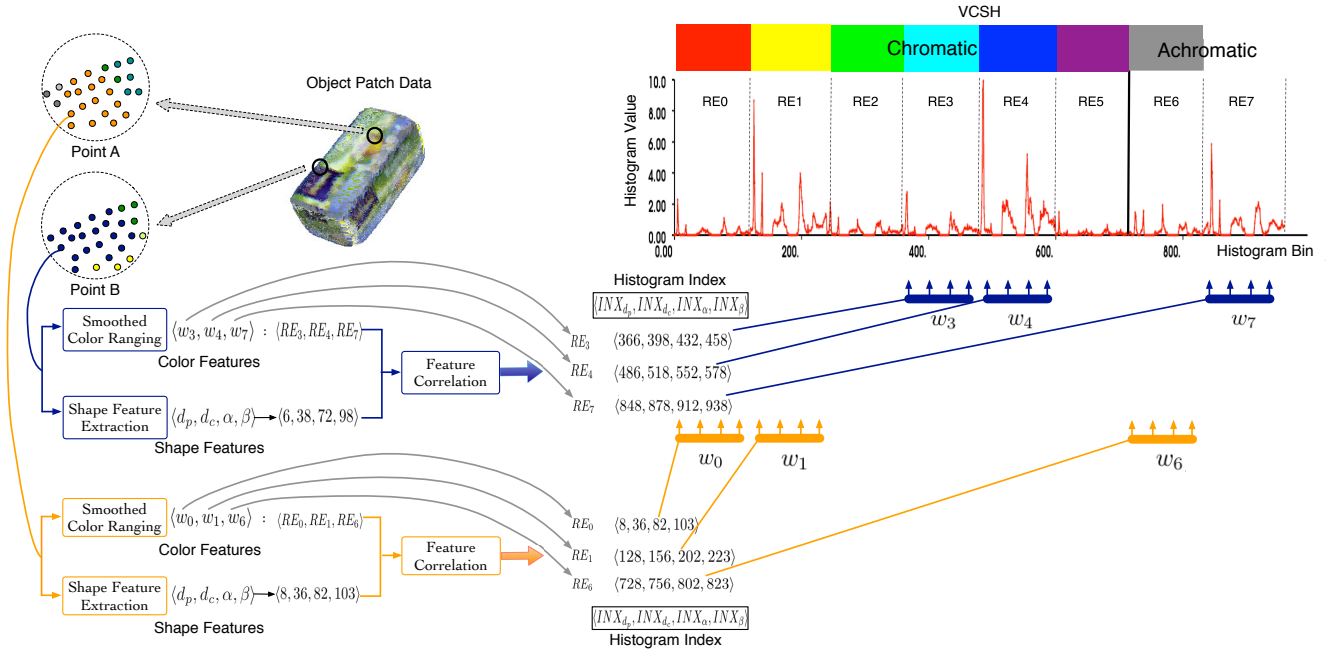


Fig. 5 VCSH training processing with smooth color ranging, shape feature extraction and feature correlation steps. All points' color contribution in object patch data respected to different viewpoints incremented into the whole VCSH's indexing, based on smoothed color ranging and geometrical features' indexing. VCSH is normalized with the points' size to deal with sensor-object distance scaling finally.

3.4 Object Recognition and Pose Retrieval

With the built object VCSH descriptors database, we are now going to get the object cluster's identification label L and its general pose P in real scene. Our system first segments and clusters the object cluster C from the background. Two frameworks of segmentation and clustering are proposed to accommodate different environments for object recognition and pose estimation:

Planar Background Environment The environment could be simplified as all the objects being with a planar background, for example a table surface as shown in Figure 8a. With the raw RGB-D image from Kinect sensor, the largest plane surface could be extracted by RANSAC [29], the object clusters C_k will be segmented from the plane surface and clustered by Euclidean distance [41].

Cluttered Background Environment The cluttered background environment is represented as a heavily cluttered background. It is difficult to constrain the objects' localization for segmentation and clustering, as the target objects have the possibilities of being with various pose as shown in Figure 8b. Aiming to solve that, the initial background image is trained in off-line phase based on Octree data structure [40]. With the extracted foreground data, the object clusters C_k will be segmented and clustered by Euclidean distance [42].

Based on object clusters C_k, the real scene objects' VCSH is calculated. The chi-squared distance χ^2 between the real scene object's VCSH value $H(C)$ and H_{ij} in the trained database is calculated for the best matching, through fast approximate K-Nearest Neighbors (KNN) method based on kd-trees [29]. $\langle \hat{L}, \hat{P} \rangle$ as the best matched object identification and corresponding pose could be extracted as:

$$\langle \hat{L}, \hat{P} \rangle = \underset{\langle L,P \rangle_{ij}}{\arg\min}\, \chi^2(H(C), H_{ij}). \tag{8}$$

Note that in VCSH definition, P in $\langle \hat{L}, \hat{P} \rangle$ represents the rotation of the object respect to the sensor's viewpoint. The centroid of the object cluster in real scene indicates the current position, which is used to update P as the object initial pose.

3.5 Object Pose Optimization and Verification

Due to the sampling rate of the synthetic viewpoints during VCSH database building, although the estimated pose P is recovered as the best matched pose from the built database, P may be not the real pose. Consequently, iterative closest point (ICP) method is employed to further optimize the estimated pose [37, 38], providing a transform T_{icp}. The sources for ICP are point cloud data of the best matched object patch data and object cluster in real scene. ICP's accuracy and iteration speed strongly rely on the given initial guess, which could be provided by our estimated pose P. The final pose of the object P_{final} is optimized according to the extracted initial pose

P and the ICP optimized transform T_{icp}. Therefore, the final updated object pose $P_{final} = T_{icp}^{-1} \cdot P$ is significantly accurate while the iteration speed is fast enough for real-time recognition and pose estimation.

Next pose verification is necessary to make sure that the optimized pose P_{final} is the correct estimation. The new object patch data M_{rec} will be generated by P_{final} and recognized objects 3D model Ω using Algorithm 1. Since the final pose is optimized, the detected object patch data M_{rec} might be not in the object model patch dataset that generated from the synthetic viewpoints during modeling. With the calculation for the difference between M_{rec} and the real object cluster data C_k, the given threshold composed by the photometric and geometrical difference could reject the false positives.

3.6 Object Localization in Semantic Map

Semantic mapping has attracted huge attention in the robotic applications, especially for wide-range navigation and exploration. Therefore, it is obvious that a coherent semantic map, which provides both semantic level understanding and metric representation of the environment, is very important for intelligent robot to successfully and efficiently perform daily tasks. To fulfill these requirements, our VCSH is an important component for this coherent semantic map building. VCSH could be utilized for the 3D object recognition and accurate pose estimation efficiently and successfully, which provides the possibility for the object localization in the large-scale semantic map. And these results could be retrieved in real time during robot mapping building process. Moreover, VCSH has no constrain for the object type. It can deal with texture and textureless objects using the object's color and shape information.

As our coherent mapping building stratage, the laser range data is firstly processed by a grid mapping algorithm to generate an occupancy grid map of the environment and to provide a coherent global coordinate system. In our work, we have used the GMapping [44] algorithm for this purpose. The resulting grid map is then used as input for the process of parametric environment abstraction which uses rectangular space units to approximate the geometry and the topology of the perceived environment. Within each space unit, unknown areas of the grid map are detected based on connected-components analysis. Such areas are considered as obstacles which can not be traversed by robots. More details on parametric environment abstraction can be found in [45]. On the other hand, 3D objects are localized in the global semantic map using our proposed object recognition and 6D pose estimation method. Finally, a coherent semantic map that captures the geometrical, topological and object information of the operating environment is generated by incorporating the 3D object information into the parametric environment model.

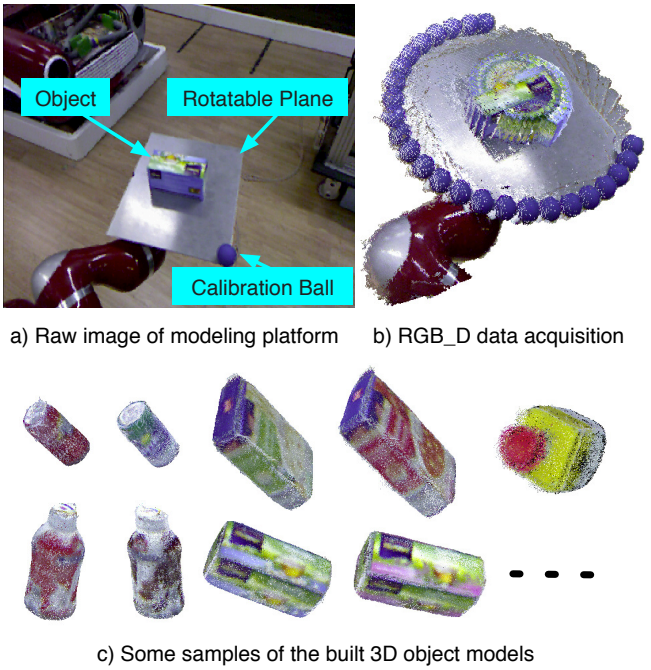


Fig. 6 Object 3D modeling, model data represents as the color point clouds.a) the platform for modeling; b) all frames object data from Kinect sensor; c) some objects models in our dataset.

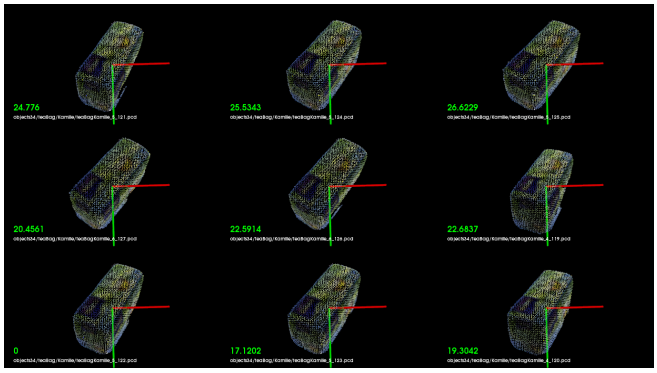


Fig. 7 Extract nine closest object VCSHs with relevant viewpoints in the dataset, after the recognition step using the simulation data. The green markers present the distances to the chosen VCSH as the target.

4 Experimental Results

We perform experiments where the goal is to evaluate our proposed *viewpoint oriented color-shape histogram* descriptor and the entire system. First, an object dataset consisting more than 25 objects is built, where some objects have the same shape but different color information on the surface. As shown in Figure 6a, the platform could be rotated by different angles using a KUKA arm end-effector controller. With a stationary Kinect sensor mounted on the robot, the color point cloud of the object can be captured with respect to the different rotating

Table 1 Map of the state-of-the-art methods on 3D object recognition and pose estimation: ConVOSH, CPPF and our VCSH could be applied for textured and textureless objects. ConVOSH can not retrieve 6D pose. CPPF as a local descriptor, faces the real-time challenge with high computational cost. Notice that the numerical values come from their respective papers, and in particular those numbers refer to their own datasets, therefore the results are illustrated for a rough comparison.

Name	Strategy	Type	Feature	Object	Dataset	Success	Pose Error	
	Recognition(R) Registration(RG) 6D Pose(P)	Local(L) Global(G)	Shape(S) Texture(T) Color(C)	Constrains	Size	Rate (%)	T (mm)	R (deg)
ConVOSCH [21]	R	L	S + C	No	63	72.2	NaN	NaN
LINEMOD [30]	R+P	L	S + C	Uniform Color	15	96.6	NaN	NaN
VFH [29]	R	G	S	Depth Only	60	98.1	NaN	NaN
CPPF [23]	RG +P	L	S + C	No	10	80	15	15
Tang [20]	R+P	G	S + T	Textured	35	90	50	10
Our VCSH	R+P	G	S + C	No	25	92	23.4	1.59

angles. Furthermore, a calibration ball is used to determine and optimize the final object model's coordination. In total, for each object, 25 frames of data with 10° as an angle step are captured at different poses. Some objects have the same shape but different color information such as COLA, SPRITE can. Some objects are textureless such as the emergency button (Figure 6c). During object model building, we assume that the object is standing on the table, its bottom part data is not in considered for the whole object model. During the object patch data generation, the viewpoints are sampled on the upper sphere surface around the object origin with radius of 0.8m. For every 10° in elevation range $[10°, 80°]$ and every 2° in azimuth range $[0°, 360°]$, a synthetic viewpoint and the relative object patch data are both generated. Therefore, $7 \times 180 = 1260$ synthetic views patch data for each object model are generated totally. In our database, each viewpoint object patch data contains around 1000-2000 color points. Consequently, each object is represented as 1260 VCSH descriptors respect to different viewpoints, which cover object's full potential poses.

To demonstrate our performance, we design multiple challenging scenarios. Some special objects are chosen to present VCSH's stability of recognition and also pose accuracy. There are some objects which have the same shape but the different visual information, some with texture or textureless surface. This challenge of common object recognition and accurate pose estimation with high speed, could not be solved by existing techniques [20–23,29,31]. We firstly use the object patch data within the database to perform closest VCSH retrieval accuracy. One VCSH in database is chosen to present the correction of its recognition and the relevant viewpoint retrieval. As shown in Figure 7, the green scores present these nine closest neighbors' VCSH distances with the target. The chosen object patch data's VCSH could be recognized correctly. All the object patch data in dataset can be correctly retrieved with 100% success rate.

Secondly, we demonstrate real time object recognition and 6D pose estimation using the real scene RGB-D data, which is captured from a single Kinect on an autonomous mobile robot. Because of Kinect's data acquisition range, the objects should be within the distance of 0.5m to 3.5m respect to the sensor. The recognized objects' 3D models are projected into the real scene with estimated 6D poses as shown in Figure 8. For the planar background scenario Figure 8a, we extracted the object cluster with the assumption that the planar surface where all the objects standing on should covers the 50% of the whole point clouds. Figure 8b illustrates the cluttered background scenario. The background should be trained at first, and all the objects have no geometrical constrains in real scene. The objects for the experiments include the textured (teabag box and milk bottle) and also the textureless (emergence button). Same shape and different color objects are also tested such as various teabag boxes to present the necessary for the object descriptor combined with color and shape features. All the trained objects could be correctly recognized and their estimated poses are highly accurate. Note that the works are partially based on Point Cloud Library [1].

Table 2 Runtime performances of our VCSH and Tang [20] on similar scenarios.

Single Object	Train	Feature Extract	Recognize	Pose Recovery
Our VCSH	2 min	5 ms	37 ms	0.83 s
Tang [20]	7 min	5 s	1 s	14 s

Our VCSH is a global object descriptor combining visual and shape features. Its geometrical feature is based on the centriod of the object cluster in real scene. With these limitations, all the objects for the demonstrations should be rigid and not be reflective or transparent. In addition, these objects should be well segmented from the environment background. As described in Sec. 3.4 and experimental results shown in Figure 8, our system could effectively deal with planar background and clut-

[1] `http://www.pointclouds.org`

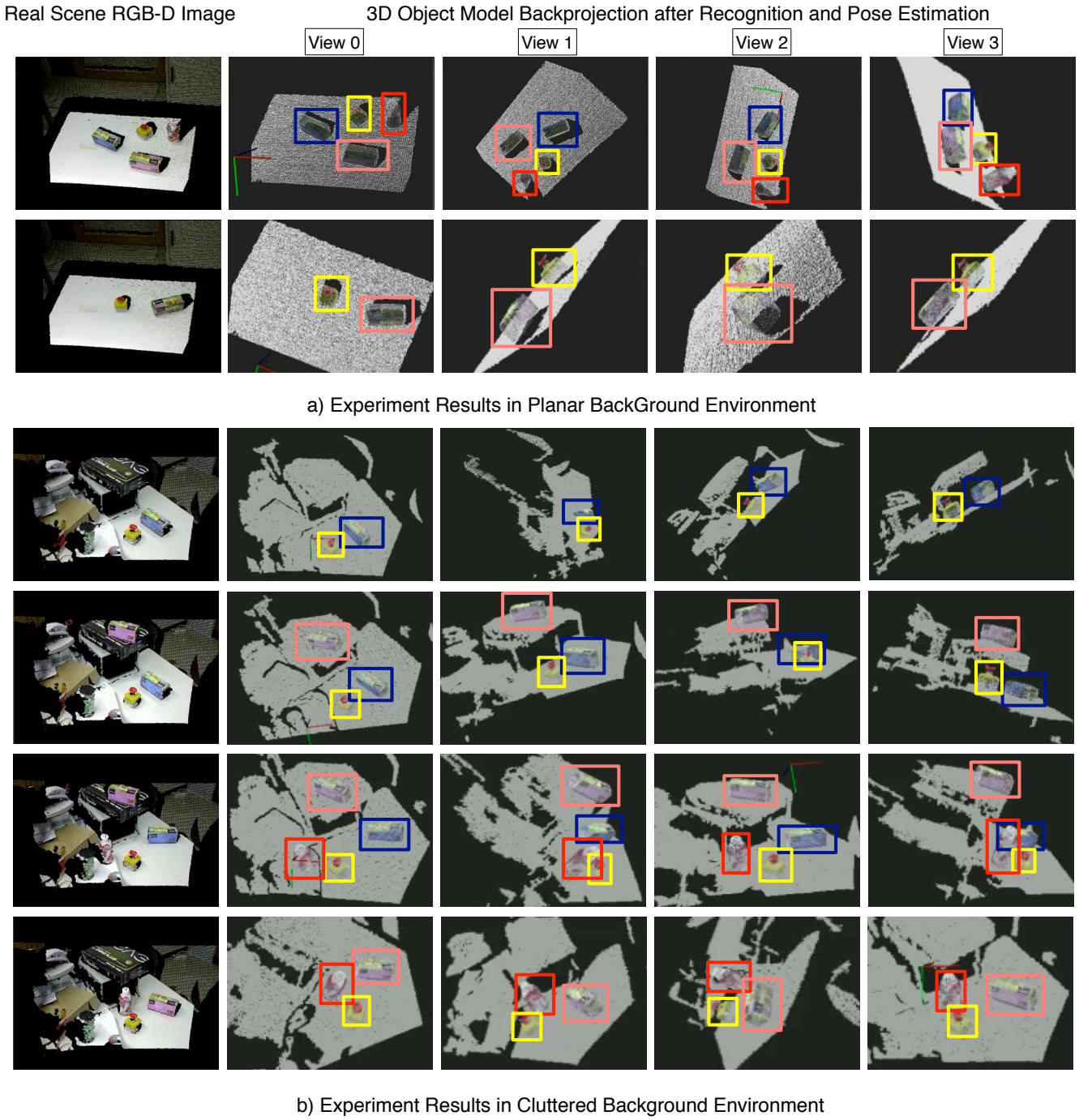


Fig. 8 Recognized objects's 3D models are projected onto the real scene with estimated 6D poses: a) within planar background environment; b) within cluttered background environment. Left one: real scene RGB-D data from the sensor. Right four: the different view results after object model backprojected into the scene data after recognition and pose estimation. Different color frames illustrate different objects.

tered environment background. To analyze the object occlusion's influence for the final results, we utilized multiple experiments for multiple objects with manual configurations for occlusion. During the experimental testing, if the object's occluded colored point clouds are less than 8% of the ideal whole object data, our VCSH provides stable and correct results for both of recognition and 6D pose estimation.

Furthermore, we apply our object recognition and 6D pose estimation method in semantic mapping for an indoor environment, as illustrated in Figure 9. As shown in Figure 9a, the resulting coherent semantic map cor-

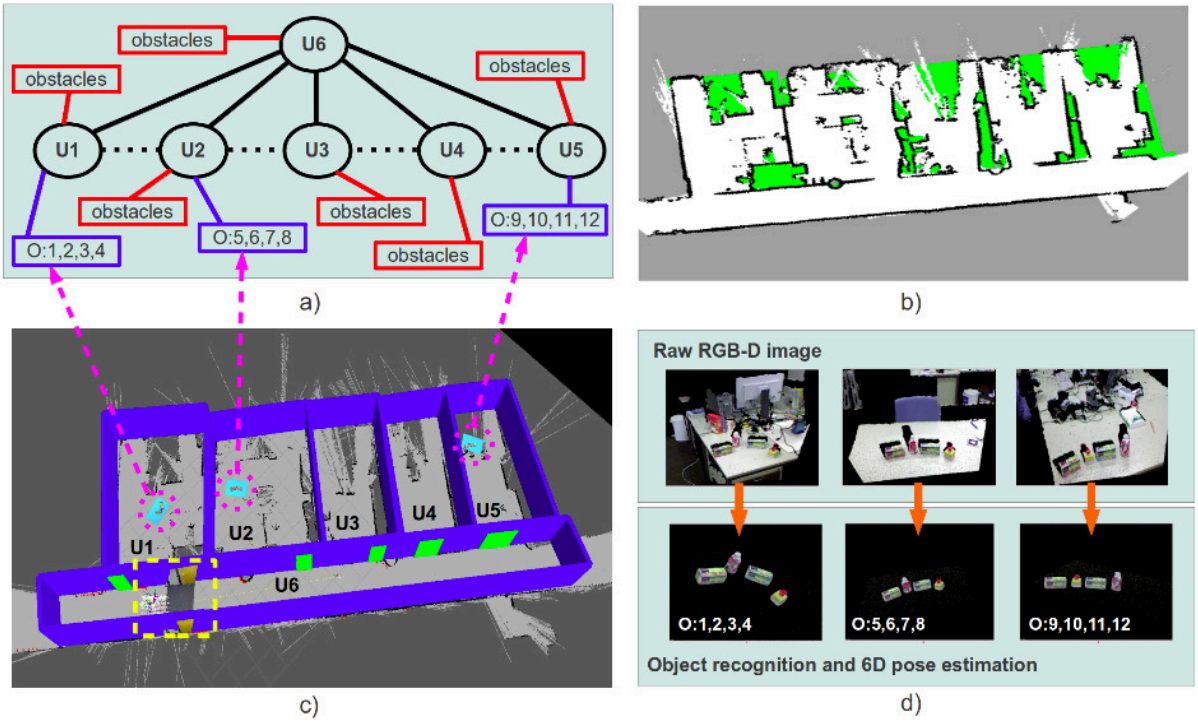
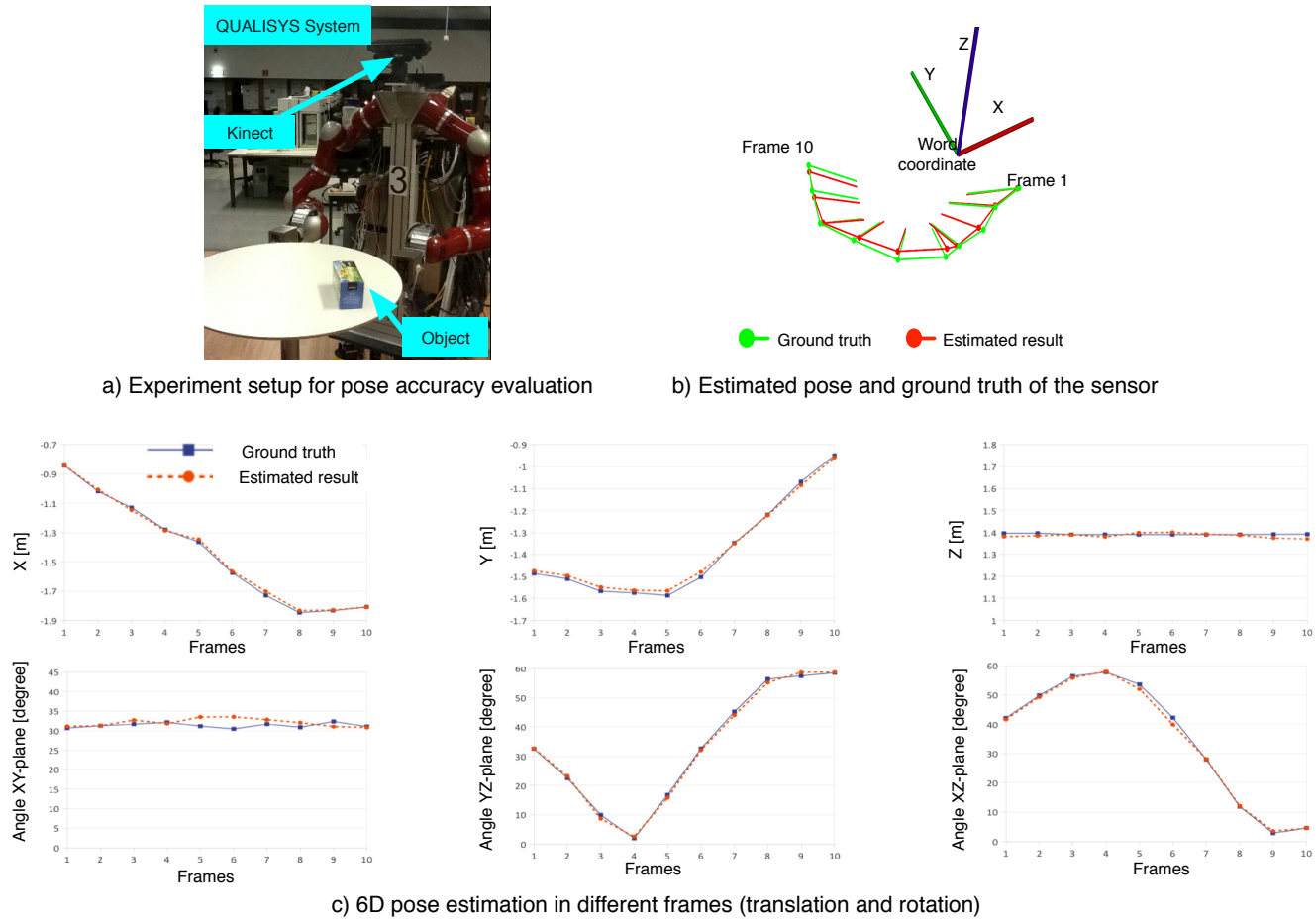


Fig. 9 Object localization in coherent semantic map. a) The abstract environment model. Black ellipses indicate space units. Solid black edges mean that two space units are connected by one or more doors. Dashed edges imply that two space units are adjacent but not connected by doors. Blue rectangles show the detected objects. Blue edges show the belongingness of these objects. b) The resulting grid map of the perceived environment. c) We plot the 3D semantic map directly onto the corresponding grid map (blue=walls, green=doors, cyan=detected tables with 3D objects). The current robot information including acquired RGBD data are highlighted by the dashed yellow rectangle. d) Details on 3D object localization. This semantic map includes identification and pose of each object in the global coordinate system and their belongingnesses.

rectly interprets the perceived environment with space units $U1, U2 \cdots U6$ and a corresponding topology (connectivity by doors and adjacency). In Figure 9b, these detected obstacles essentially represent the furniture of the perceived environment, such as tables and cabinets. 3D parametric model along with the detected 3D objects are shown in Figure 9c. Here the detected table planes and objects are back-projected in the map. Figure 9d depicts the details of object recognition and localization. In space units $U1, U2$ and $U5$, several 3D objects are recognized and localized regarding their 6D poses. By cell-wise checking of our parametric model and the input grid map, we measured an accuracy of 94.1% for geometry approximation. The mismatch of 5.9% is mainly due to some not-fully-explored areas of the input map.

Table 1 presents the state-of-the-art methods on the topic of object recognition and 6D pose estimation. There are mainly two types of descriptors including global and local. In particular, the local type is similar to the method of model registration. It could solve the problem when object data contains occlusions, using the pairwise matching with different features. However, this method requires high computational cost and is not suitable for real-time processing such as robotic applications. Furthermore, most of the local object descriptors must have the prior knowledge about the object' existence in real scene, such as CPPF [23]. Instead, in this paper, we introduce a new global object descriptor VCSH. Compared with other global methods as VFH [29] and Tang [20], we can retrieve accurate 6D pose, which cannot be solved in VFH [29]. Moreover, VFH [29] only uses shape features, thus cannot distinguish the objects with same shape but different visual appearance. The most similar work Tang [20] to ours, uses the SIFT feature based on the objects surface texture information. This method constrains the target objects be be textured with high quality and cannot deal with the textureless object, such as emergency button in our dataset. Our VCSH is based on object's color and shape features, thus has no object model type constraints, and could deal with the textured and textureless objects. Comparatively, most of the methods have their object model constrains, such as textured (Tang [20]), uniform color (LINEMOD [30]), depth only (VFH [29]). To the best of our knowledge, we are currently among the first to solve these problems with high recognition rate, accurate 6D pose estimation and low computational cost, without any object model constrains by combining the photometric and geometric information.



a) Experiment setup for pose accuracy evaluation

b) Estimated pose and ground truth of the sensor

c) 6D pose estimation in different frames (translation and rotation)

Fig. 10 Object pose accuracy evaluation in different frames with different robot positions: a) experiment setup for evaluation with omni-direction platform robot and QUALISYS tracker system; b) the estimated sensor trajectory with 10 frames; c) the estimated pose and ground truth in translation and rotation.

In general, our framework can reach the correct recognition and pose at 92%, correct recognition but wrong pose at 6% and 2% for wrong recognition over 1000 demonstrations. Computational cost and runtime performance are very important for applying our framework into the autonomous mobile robot's applications. The runtime performance for single object recognition and pose recovery is evaluated, we compare with the result from Tang et. al. [20] as shown in Table 2 on similar scenario setups. All experiments run on AMD X6 3.0 GHz with 8GB of RAM, while Tang et al. use 6-core 3.2GHz i7 with 24GB of RAM. For the single object recognition and pose recovery, it costs around 1s without any GPU speed-up architecture. Our framework's runtime outperformance brings the opportunities for the real time robotic applications, for instance, object grasping and manipulation based on perceptual system.

To further evaluate the pose accuracy using our proposed approach, QUALISYS motion capture system [2] is employed to capture the ground truth of the sensor pose. The robot with the Kinect senor moves around the stationary object. The camera pose is estimated with two methods for accuracy analysis: 1) recovered pose respect to the stationary object from our proposed method; 2) estimated pose using QUALISYS system as the ground truth. By transforming these data into the world coordinate, we compare the estimated pose with its ground truth to get the pose recovery accuracy, as shown in Figure 10. The root mean square error (RMSE) during the whole 10 frames is calculated for the pose accuracy analysis. From Table 1, the 6D pose error compared with the ground truth are 23.4 mm in translation and 1.59 degrees in rotation, while in work [20] are 50mm and 10 degrees respectively. Our VCSH outperforms Tang et. al.'s method both in translation and rotation accuracy with similar object models for the similar scenarios.

As the object's color information as photometric features is extracted for VCSH generation, the stability with illumination changes is a crucial aspect, therefore it needs to be analyzed. We utilize one light meter DT1309 to estimate the object's surrounding illumination intensity under an adjustable white LED array light. The sta-

[2] `http://www.qualisys.com/`

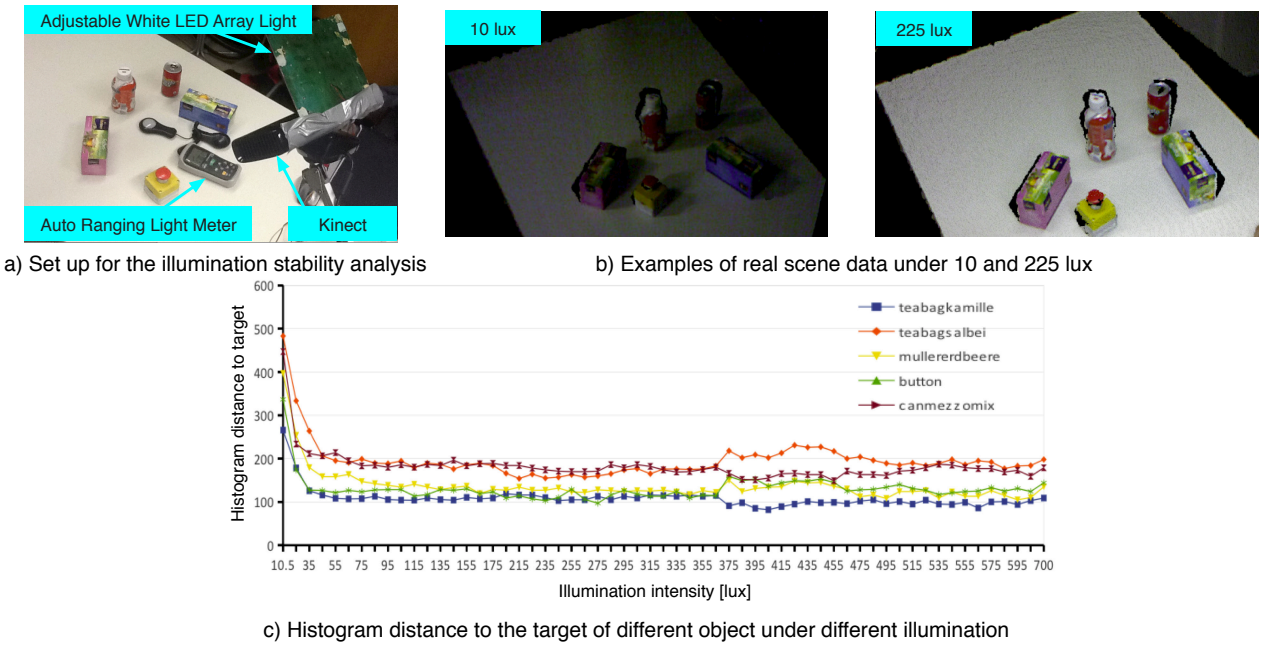


Fig. 11 Stability analysis with illumination change: a) experimental setup with adjustable white LED array and light meter to measure the illumination density; b) some real scene data recored under the different illumination densities; c) five objects's estimated VCSH distances to the relative targets under sixty different illumination densities from 10 lux to 700 lux.

bility is evaluated by the differences between the estimated objects' VCSH under various illumination conditions and their target VCSH (correct identification and pose) in database. As illustrated from Figure 11, when the illumination intensity exceeds 50 lux, all the objects' histogram differences remain under 220 and would be stable until 700 lux, which is the maximum illumination intensity. Note that the object modeling environment is under around 230 lux, while most of the common indoor and outdoor light condition is from 150 to 400 lux. From the result of stability analysis, our recognition and pose estimation framework, especially VCSH object descriptor is stable enough under varying illumination intensity.

From above experimental results, our proposed approach consisting of a novel object descriptor VCSH is efficient and robust. It guarantees high object recognition rate, fast and accurate pose estimation as well as exhibits the capability of dealing with illumination changes.

5 Conclusion and Future Work

In this paper, we presented a framework consisting of a global object descriptor *Viewpoint oriented Color-Shape Histogram*, which combines color and shape information for object recognition and 6D pose estimation. The proposed approach could be easily integrated into various robotic perception system for daily textured/textureless objects recognition and 6D pose estimation in real time. In addition, we successfully incorporated within a coherent semantic map, which could be used for robot exploration of objects in large-scale map.

Our approach achieves 92% success object recognition rate for both of correct object identification and pose retrieval. The estimation error of recognized object's 6D pose is under 24mm in translation and 1.6 degree in rotation. Our proposed framework has light computation cost. For a single object, it spends less than 1s to recognize and estimate it's accurate pose estimation after the pose optimization. Moreover, our VCSH is efficient and stable enough under varying illumination intensity in the common environment. Our experimental results demonstrate that the proposed approach is proven to be efficient by guaranteeing high object recognition rate, accurate pose estimation result. Moreover, it exhibits the capability of dealing with environmental illumination changes. Future work will focus on the system speed up based on GPU implementation and model building of wider-variety objects.

References

1. I. Biederman and P. C. Gerhardstein: Recognizing depth-rotated objects: evidence and conditions for three-dimensional viewpoint invariance. Journal of Experimental Psychology: Human Perception and Performance. **19**(6), 1162–1182 (1993)
2. I. Biederman and E. E. Cooper: Evidence for complete translational and reflectional invariance in visual object priming. Perception. **20**(5), 585–593 (1991)
3. S. Edelman and H. H. Bülthoff: Orientation dependence in the recognition of familiar and novel view of three-

dimensional objects. Vision Research. **32**(12), 2385–2400 (1992)
4. R. Ellis, D. A. Allport, G. W. Humphreys, and J. Collis: Varieties of object constancy. Quarterly Journal of Experimental Psychology. **41**(4), 775–796 (1989)
5. J. Fiser and I. Biederman: Size invariance in visual object priming of gray-scale images. Perception. **24**(7), 741-748 (1995)
6. D. G. Lowe: Distinctive image features from scale-invariant keypoints. IJCV. **60**(2), 91–110 (2004)
7. H. Bay, A. Ess, T. Tuytelaars, and L. Van Gool: Speeded-Up robust features. CVIU. **110**(3), 346–359 (2004)
8. N. Dalal and B. Triggs: Histograms of oriented gradients for human detection. In: Proceedings of CVPR, pp. 886–893 (2005)
9. A. E. Abdel-Hakim and A. A. Farag: CSIFT: a SIFT descriptor with color invariant characteristics. In: Proceedings of CVPR, pp. 1978–1983 (2006)
10. T. Gevers and A. W.M. Smeulders: Color-based object recognition. IJPR. **32**, 453–464 (1999)
11. T. Gevers: Robust histogram construction from color invariants for object recognition. TPAMI. **26**(1), 113–118 (2004)
12. D. Crandall and J. Luo: Robust color object detection using spatial-color joint probability functions. In: Proceedings of CVPR, pp. 1443–1453 (2004)
13. N. J. Mitra, L. J. Guibas, J. Giesen, and M. Pauly: Probabilistic fingerprints for shapes. In: Proceedings of Eurographics Symposium on Geometry Processing, pp. 121–130 (2006)
14. U. Clarenz, M. Rumpf, and A. Telea: Robust feature detection and local classification for surfaces based on moment analysis. IEEE Transactions on Visualization and Computer Graphics. **10**(5), 516–524 (2004)
15. N. Gelfand, N. J. Mitra, L. J. Guibas, and H. Pottmann: Robust global registration. In: Proceedings of Eurographics Symposium on Geometry Processing, no. 197 (2005)
16. K. Safronov, I. Tchouchenkov, and H. Wrn: Hierarchical iterative pattern recognition method for solving bin picking problem. In: Proceedings of Robotik, pp. 3–6 (2008)
17. P. Shilane and T. Funkhauser: Selecting distinctive 3D shape descriptors for similarity retrieval. In: Proceedings of IEEE Conference on Shape Modeling and Applications, pp. 18–27 (2006)
18. M. Pauly, R. Keiser, and M. Gross: Multi-scale feature extraction on point-sampled surfaces. Computer Graphics Forum. **22**(3), 281–289 (2003)
19. M. Kolomenkin, I. Shimshoni, and A. Tal: On edge detection on surfaces. In: Proceedings of CVPR, pp. 2767–2774 (2009)
20. J. Tang, S. Miller, A. Singh, and P. Abbeel: A textured object recognition pipeline for color and depth image data. In: Proceedings of ICRA, pp. 3467–3474 (2012)
21. A. Kanezaki, Z. Marton, D. Pangercic, T. Harada, Y. Kuniyoshi, and M. Beetz: Voxelized shape and color histograms for RGB-D. In: Proceedings of IROS, Workshop on Active Semantic Perception and Object Search in the Real World (2011)
22. A. Kanezaki, T. Harada, and Y. Kuniyoshi: Partial matching of real textured 3D objects using color cubic higher-order local auto-correlation features. Journal of the Visual Computer. **26**(10), 1269–1281 (2010)
23. C. Choi and H. I. Christensen: 3D pose estimation of daily objects using an RGB-D camera. In: Proceedings of IROS, pp. 3342-3349 (2012)
24. F. Tombari, S. Salti, and L. D. Stefano: A combined texture-shape descriptor for enhanced 3D feature matching. In: Proceedings of ICIP, pp. 809–812 (2011)
25. J. Liebelt, C. Schmid, and K. Schertler: Viewpoint-independent object class detection using 3D feature maps. In: Proceedings of CVPR, pp. 1–8 (2008)
26. Z. Fan and B. Lu: Fast recognition of multi-view faces with feature selection.In: Proceedings of ICCV, pp. 76–81 (2005)
27. A. Kushal, C. Schmid, and J. Ponce: Flexible object models for category-level 3D object recognition. In: Proceedings of CVPR, pp. 1–8 (2007)
28. A. Thomas, V. Ferrari, B. Leibe, T. Tuytelaars, B. Schiele, and L. V. Gool: Towards multi-view object class detection. In: Proceedings of CVPR, pp. 1589–1596 (2006)
29. R. B. Rusu, G. Bradski, R. Thibaux, and J. Hsu: Fast 3D recognition and pose using the viewpoint feature histogram. In: Proceedings of IROS, pp. 3467–3474 (2010)
30. S. Hinterstoisser, V. Lepetit, S. Ilic, S. Holzer, G. Bradski, K. Konolige, and N. Navab: Model based training, detection and pose estimation of texture-less 3D objects in heavily cluttered scenes. In: Proceedings of ACCV, pp. 548–562 (2012)
31. R. B. Rusu, N. Blodow, and M. Beetz: Fast point feature histograms (FPFH) for 3D registration. In: Proceedings of ICRA, pp. 3212–3217 (2009)
32. W. Wohlkinger and M. Vincze: Ensemble of shape functions for 3D object classification. In: Proceedings of ROBIO, pp. 2987–2992 (2011)
33. F. Tombari, S. Salti, and L. D. Stefano: Unique signatures of histograms for local surface description. In: Proceedings of ECCV, pp. 356–369 (2010)
34. A. Vadivel, A. K. Majumdar, and S. Sural: Perceptually smooth histogram generation from the HSV color space for content based image retrieval. In: Proceedings of Advances in Pattern Recognition, pp. 248–251 (2003)
35. C. B. Akgül, B. Sankur, F. Schmitt, and Y. Yemez: Multivariate density-based 3D shape descriptors. In: Proceedings of Shape Modeling International, pp. 3–12 (2007)
36. M. Alexa, J. Behr, D. Cohen-Or, S. Fleishman, D. Levin, and C. T. Silva: Computing and rendering point set surfaces. IEEE Transaction on Visualization and Computer Graphics. **9**(1), 3–15 (2003)
37. Z. Zhang: Iterative point matching for registration of free-form curves and surfaces. IJCV. **13**(2), 119–152 (1994)
38. K. S. Arun, T. S. Huang, and S. D. Blostein: Least-squares fitting of two 3-D point sets. TPAMI. **9**(5), 698–700 (1987)
39. S. Sural, G. Qian, and S. Pramanik: A histogram with perceptually smooth color transition for image retrieval. In: Proceedings of International Conference on Computer Vision, Pattern Recognition and Image Processing, pp. 664–667 (2002)
40. H. Eberhardt, V. Klumpp, U. D. Hanebeck: Density trees for efficient nonlinear state estimation. In: Proceedings of International Conference on Information Fusion, pp. 1–8 (2010)
41. W. Wang, V. Koropouli, Dongheui Lee, K. Khnlenz: Articulated object modeling based on visual and haptic observations. In: Proceedings of VISAPP, pp. 253–259 (2013)
42. W. Wang, D. Brščić, Z. He, S. Hirche, and K. Kühnlenz: Real-time human body motion estimation based on multi-layer laser scans. In: Proceedings of URAI, pp. 297–302 (2011)
43. W. Wang, S. Li, L. Chen, D. Chen, and K. Kühnlenz: Fast object recognition and 6D pose estimation using viewpoint oriented color-shape histogram. In: Proceedings of ICME, pp. 1–6 (2013)
44. G. Grisetti, C. Stachniss, and W. Burgard: Improved techniques for grid mapping with rao-blackwellized particle filters. IEEE Transactions on Robotics. **23**(1), 34–46 (2007)
45. Z. Liu, W. Wang, D. Chen, G. v. Wichert: A coherent semantic mapping system based on parametric environment abstraction and 3D object localization. In: Proceedings of European Conference on Mobile Robots (2013)