FAST OBJECT RECOGNITION AND 6D POSE ESTIMATION USING VIEWPOINT ORIENTED COLOR-SHAPE HISTOGRAM

Wei Wang[1], *Shile Li*[1], *Lili Chen*[1], *Dongming Chen*[2], *Kolja Kühnlenz*[13]

[1]Technische Universität München, D-80290 München, Germany.
{wei.wang, lili.chen}@tum.de li.shile@mytum.de
[2]Ecole Centrale de Lyon, LIRIS UMR 5205, F-69134, Lyon, France. dongming.chen@ec-lyon.fr
[3]Bayerisches Landesamt für Maß und Gewicht, D-80638 München, Germany. koku@tum.de

ABSTRACT

In this paper, we present an object recognition and pose estimation framework consisting of a novel global object descriptor, so called *Viewpoint oriented Color-Shape Histogram* (VCSH), which combines object's color and shape information. During the phase of object modeling and feature extraction, the whole object's color point cloud model is built by registration from multi-view color point clouds. VCSH is trained using partial-view object color point clouds generated from different synthetic viewpoints. During the recognition phase, the object is identified and the closest viewpoint is extracted using the built feature database and object's features from real scene. The estimated closest viewpoint provides a good initialization for object pose estimation optimization using the iterative closest point strategy. Finally, objects in real scene are recognized and their accurate poses are retrieved. A set of experiments is realized where our proposed approach is proven to outperform other existing methods by guaranteeing highly accurate object recognition, fast and accurate pose estimation as well as exhibiting the capability of dealing with environmental illumination changes.

Index Terms— Object recognition, 6D pose estimation, viewpoint oriented color-shape histogram

1. INTRODUCTION

Object recognition and 6D pose estimation plays a crucial role in a wide range of robotic applications, such as object grasping and manipulator occlusion handling. More specifically, successful object recognition, highly accurate pose estimation and near real time operation are necessary capabilities but also tough challenges for a robot perception system.

A variety of object descriptors using different features of the objects have been proposed to solve the problems mentioned above. The most popular features are currently the SIFT [1] and SURF [2], both are extracted based on object's texture information. Fast Point Feature Histogram (FPFH) [3] and Viewpoint Feature Histogram (VFH) [4] are geometry-based shape descriptors. However, these descriptors are restricted to some objects which are fully textured or distinctive through their shape. These disadvantages make above object descriptors to be restricted useful, since some objects in real world are textureless and may have the same shape but different visual information. An autonomous robot perception system should be able to recognize the objects with aforementioned case and accurately estimate their poses.

With the massively increased usage of new-released RGB-D sensors, which can provide geometrical and visual information about the real scene. Object descriptor could use multi-dimensional color and geometrical features for object recognition and pose estimation by using such a depth sensor. With the real scene data, the object needs to be recognized with different poses, thus the viewpoint component could be integrated into the object descriptor building. For this aim, a novel framework and object descriptor for object recognition and pose estimation are proposed in this paper, which provide the following main contributions: 1) A novel object descriptor *Viewpoint oriented Color-Shape Histogram* combined with color and shape features, including object viewpoint component; 2) A framework which gives highly object recognition rate and its accurate 6D pose estimation; 3) Object pose accuracy evaluation and stability quantitive analysis with respect to the illumination changes; 4) Live demonstrations and comparisons with existing methods.

This remainder of the paper is organized as follows: Section 2 provides the proposed framework and the detailed description of proposed VCSH object descriptor. The experimental results including the pose accuracy evaluation, stability analysis with illumination and running time performance are presented in Section 3. Finally, Section 4 summarizes the paper and proposes future development roads.

2. PROPOSED APPROACH

The framework of our proposed object recognition and 6D pose estimation system is illustrated in Figure 1. In the training phase, we first build the whole 3D object model by registering all the object's data with different poses into a sin-

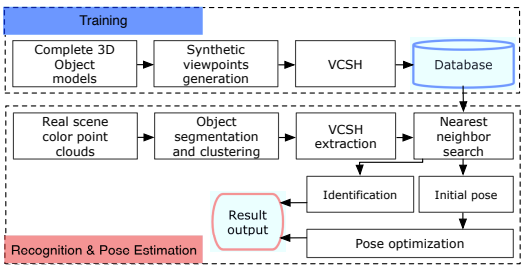

Fig. 1. Framework of object recognition and pose estimation.

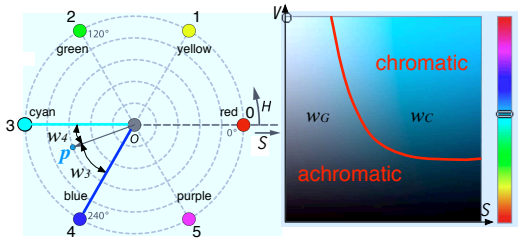

Fig. 2. Left: smoothed color range and estimate the contributions for neighbor regions in HS space. Right: illustrate the chromatic and achromatic area in SV space.

gle coordinate frame. Follows with a large amount of object patch data generated from the 3D object model, according to synthetic viewpoint with known pose. The generated object patch data includes potential object label and corresponding viewpoint pose. Subsequently VCSH can be computed as a global object descriptor from each object patch data, which then is stored into our database. In the 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 label and its initial pose. Finally, the highly accurate pose can be recovered through optimization and verification.

2.1. Synthetic Viewpoints Generation

The 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 $\{S_f\}$ and its poses $\{PO_f\}$ are captured where f is frame index. By registering $\{S_f\}$ with $\{PO_f\}$ into a single object coordinate, the whole 3D model O then can be generated as a cluster of color point cloud. In order to eliminate noises, the moving least squares (MLS) algorithm is utilized to smooth the whole 3D model.

For each object model O_i, where $i = 1...I$, we generate J object patch data M_j with synthetic viewpoint VP_j where $j = 1...J$. These generated synthetic viewpoints could be taken as the sensor's view direction to the object, which also illustrates the object's rotation respect to the sensor. Aiming at the object full pose estimation, all the potential view directions should be considered. For that, the synthetic viewpoints are generated on a half sphere surface, with the center of the object model O's centroid and a certain radius. The synthetic viewpoint position is generated on sphere surface in elevation and azimuth direction with certain angle step, and its direction is point to the object's centroid. With these generated synthetic viewpoints VP_j, object patch data M_j is generated according to VP_j using similar ray-casting method from the whole 3D object model O. The object model O is not restricted as the raw color point cloud model using our proposed modeling platform, but also applicable for the invented color CAD models. Then, a global object descriptor is in need to describe each M_j with its viewpoint VP_j for final object recognition and its pose recovery.

2.2. Viewpoint oriented Color-Shape Histogram

For recognition and pose recovery for common objects, an object descriptor which consists both color and geometrical information is prerequisite. In particular, this descriptor could differentiate these objects which have same shape but different colors and also same color but different shapes. For these requirements, a novel object descriptor *viewpoint oriented color-shape histogram* combines color and shape features is proposed. During VCSH construction, firstly, the color of each point p_t in object patch data M_j is smoothed ranged and given distributions for different color ranges, where $t = \{1 \cdots T\}$ is the point index. Secondly, object's shape features are estimated, which describe each point's geometrical relationship with the viewpoint VP_j and the centroid c of M_j. Finally, these extracted color and shape feature are correlated as VCSH for each object patch data M_j.

2.2.1. Smoothed Color Ranging

To represent the uniqueness of color feature for each object patch data M, its color need to be characterized and the distributions for different color ranges need to be estimated by their color values. To be more robust to illumination changes, the point cloud's RGB value is convert to HSV color space (Hue, Saturation and Value) as shown in Figure 2. The Hue component H is represented with 360 degrees angular dimension for different color. The saturation $S \in [0, 1]$ indicates the colorfulness and value $V \in [0, 1]$ describes the brightness.

Compared with the work [5] which only using the Hue histogram for the color feature representation, in our proposed VCSH, the HSV values are used for the points' color ranging not only in true color space, but also in gray scale. As shown in Figure 2, there are chromatic and achromatic areas in SV space, in which the chromatic area could be considered as the true color space, achromatic area represents the gray scale space. To this consider, eight histogram regions RE with index $u = \{0 \cdots 7\}$ are divided for the whole VCSH building, in which six are for true color space (chromatic) and the other

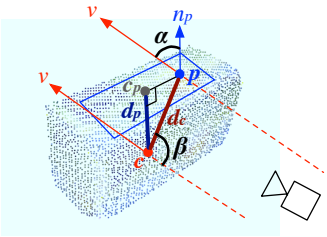

Fig. 3. Point p's shape features extraction.

two for grey scale space (achromatic).

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

- 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 $u = n = \{0 \cdots 5\}$.
- For 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 two color distributions $[w_{H_u}, w_{H_{u+1}}]$ respect to the neighbor histogram regions $[RE_u, RE_{u+1}]$ in true color space, based on the Hue distances to CR_n and CR_{n+1} where $u = n$:

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

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 in gray scale. In particular, if $S = 0$, color changes from black to white when V increases from 0 to 1, and if $V = 0$, color changes from gray to the pure hue color when S increases from 0 to 1. Since the color in achromatic space has high sensitive hue value with 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 [6] 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 are summed to be equal unity and determined by S and V as:

$$w_C = S^{r(1/V)^{r_1}}, w_G = 1 - w_C, \quad (2)$$

where $r, r_1 \in [0, 1]$. For the best precision of the true color, $r = 0.14$ and $r_1 = 0.9$ are chosen empirically. In particular, in the achromatic area which consists of two histogram regions RE_6 and RE_7, V is quantized and these distributions are calculated for these two regions: if $V < 0.5$, $w_6 = w_G$ and $w_7 = 0$, otherwise $w_6 = 0$ and $w_7 = w_G$. The final distributions w_u and w_{u+1} considering whole true color and gray space then have to be updated as:

$$w_u = w_{H_u} \times w_C, w_{u+1} = w_{H_{u+1}} \times w_C. \quad (3)$$

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 $\langle w_u, w_{u+1}, w_6 | w_7 \rangle$.

2.2.2. *Shape Feature Extraction*

After the color contributions have been estimated for the specific histogram regions, we are now 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 represents the partial data of the object from viewpoint VP with direction v, each point p's geometrical feature should be extracted in order to describe the object shape accurately and robustly. Partly inspired by [7], we extracted these features depends on the 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 respect to the sensor direction. The VP and the centroid c could be transformed as the 6D pose of the object.

As shown in Figure 3, the tangent plane of p is defined as a plane that is orthogonal to p's normal v. 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:

$$\begin{aligned} d_p &= \|p - c\|, d_c = \|c_p - c\| \\ \alpha &= \arccos(n_p \cdot (p - c)), \beta = \arccos(v \cdot (p - c)). \end{aligned} \quad (4)$$

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

2.2.3. *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 is segmented into eight regions. Every component in each point's geometrical 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 by the quantization using their values scaling from M's minimum to maximum value. Each p's two angle components $\langle \alpha, \beta \rangle$ are indexed by the

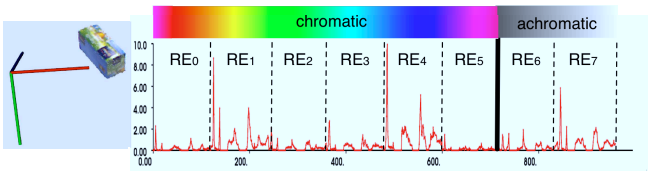

Fig. 4. Left: object's patch data from a certain viewpoint. Right: generated viewpoint oriented color-shape histogram.

quantization using their values with the range of 0 to 90°. As the correlation step, each p's color contributions for three histogram regions $\langle RE_u, RE_{u+1}, RE_6|RE_7\rangle$ are added into the geometrical certain bins $\langle IN_{d_p}, IN_{d_c}, IN_\alpha, IN_\beta\rangle$ in each of these three RE. 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 following with the object cluster point number. To overcome this problem, the values of histogram are normalized with point number finally. The VCSH could correctly indicates the certain view object's color and shape features, no matter with the distance from sensor to object. Thus, VCSH could be viewed as a geometrical constrained color feature histogram. As shown in Figure 4, the sampled object has a blue rectangle region on the top surface. These points in this region has significant large histogram value in the bins of RE_4, because of the similar color and geometrical features.

Consequently, the final correlated histogram has $(6+2)\times(30\times4) = 960$ dimensions. The computational complexity of VCSH is $O(n)$, where n is the point number of single viewpoint object patch data M. This dimension size and computational complexity makes VCSH feasible for real-time application. Furthermore, the final generated histogram could represent the object's point color and shape with high accuracy, which gives the possibility for the highly successful object recognition and accurate pose estimation.

2.3. Object Recognition and Pose Retrieval

With the built object VCSH descriptors database, we are now going to get the real scene potential object cluster's identification label L as recognition result and its general pose P. Our system first segments and clusters the object cluster C from the background. The largest plane surface could be extracted by RANSAC [4], since all the objects are assumed that standing on a table or a planar background. All the object clusters C_k will be segmented from the plane surface and clustered by Euclidean distance. Based on C_k, the real scene objects' VCSH is calculated. The chi-squared distance between the real scene object's VCSH value $Hist(C)$ and each $Hist_{ij}$ in the trained database is calculated for the best matching. In the database, each object model O contains J number VCSHs as the object descriptors from different viewpoints. The fast approximate K-Nearest Neighbors (KNN) method is employed

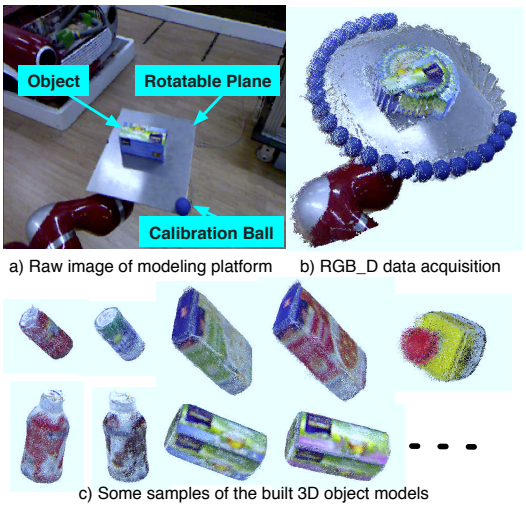

Fig. 5. Whole object 3D modeling building, final data represents as the color point clouds.

for best matching based on kd-trees [4]. The best matched object identification and the relative viewpoint pose $\langle \hat{L}, \hat{P}\rangle$ could be extracted as:

$$\langle \hat{L}, \hat{P}\rangle = \underset{\langle L,P\rangle_{ij}}{\arg\min}\, \chi^2(Hist(C), Hist_{ij}). \quad (5)$$

Have to mention here, 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 in the real scene.

2.4. Object Pose Optimization and Verification

As the estimated pose P is recovered as the best matched pose from the built database, however, because of the sampling rate of the synthetic viewpoints during the database building, the P could be not the correct pose of object. Consequently, iterative closest point (ICP) method is employed for the accurate pose optimization [8]. ICP's accuracy and iteration speed are strongly judged by the given initial guess. Our method could estimate the general pose of object by extracting the closest viewpoint in the object database. The final pose of the object P_{final} is optimized with the extracted initial pose from the recognition step and the ICP estimated transform T_{icp}, which is computed by the closed object patch data and real extracted object cluster in real scene. After ICP, the final updated object pose $P_{final} = P \cdot T_{icp}$ is significant accurate and the iteration speed is fast enough for the real-time recognition and pose estimation scenarios.

The pose verification is necessary to guarantee the recognized object with estimated pose P_{final} is the correct hypothesis in the object database. The patch object data M_{rec} is extracted by the estimated P_{final} view. By the comparison with the estimated nearest object patch data $\hat{M}$ in the recognition step, the incorrect recognition or the error pose will be rejected when the distance beyonds a given threshold.

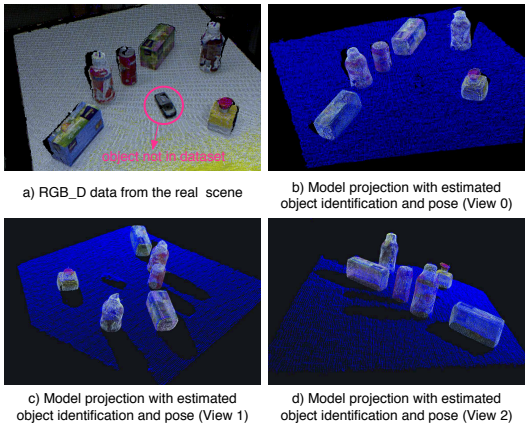

Fig. 6. Recognized objects's 3D models are projected into the real scene with estimated 6D poses.

Table 1. Running time performance of proposed method

Single Object	Train	Feature Extract	Recognize	Pose Estimate
VCSH	5 min	20 ms	70 ms	1.7 s
Tang 2012	7 min	5 s	1 s	14 s

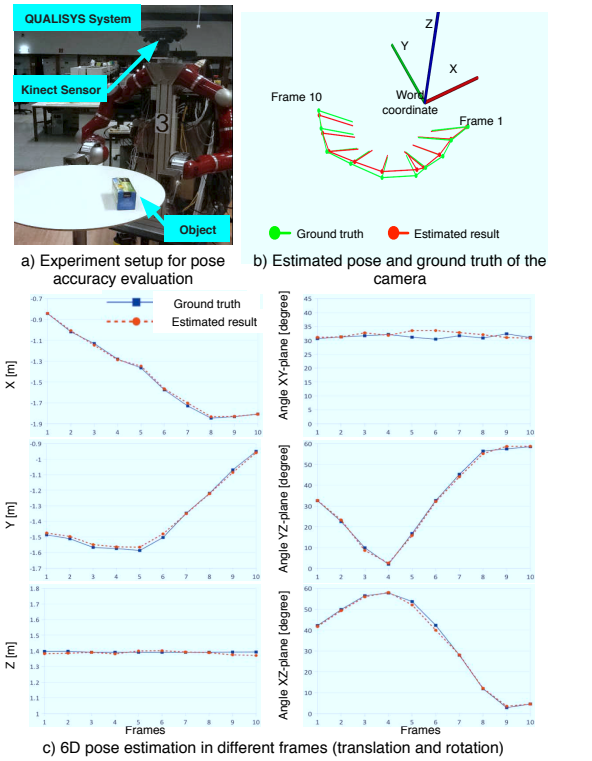

Fig. 7. Object pose accuracy evaluation in different frames with different robot positions.

3. EXPERIMENTAL RESULTS

We perform experiments where the goal is to evaluate our proposed *viewpoint oriented color-shape histogram* descriptor and the system architecture. First, an object dataset consisting more than 20 objects is built, where some objects have the same shape but different color information on the surface. As shown in Figure 5, 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 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 the cola and sprite tan and the different taste tea bags, see Figure 5c.

During object model building, note that, as 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 half sphere surface around the object with radius of 0.8m. For every $10°$ in elevation and every $2°$ in azimuth, 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 full potential poses of object. VCSH combines object's visual and geometrical features, so that it gives the maximum capability for object recognition and accurate pose estimation.

To demonstrate our superior performance compared to state-of-the-art, we design multiple challenging scenarios. Some special objects are chosen for the demonstrations to show our 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 [3, 4, 5, 9, 8]. The recognized objects' 3D models are projected into the real scene with their estimated 6D poses as shown in Figure 6. Notice that the cellphone is not recognized since it has not been built in our model database. All the trained objects could be correctly recognized and their estimated poses are highly accurate. These works are partially based on Point Cloud Library [1].

Our framework using VCSH can reach the correct recognition and pose as 92%, correct recognition but wrong pose as 6% and 2% for wrong recognition over 100 demonstrations. For the running time performance evaluation, we compare with the result from [5] as shown in Table 1. Our testing results run on AMD X6 3.0 GHz with 8GB of RAM, while [5] uses 6-core 3.2GHz i7 with 24GB of RAM.

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

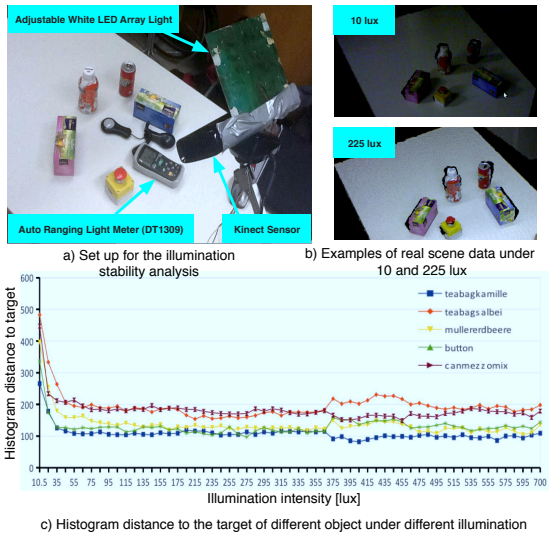

Fig. 8. Stability analysis with illumination change.

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 and estimates the object's pose. With these data transformed into the world coordinate, we compare the estimated pose with the ground truth to get the pose recovery accuracy, as shown in Figure 7. The root mean square error during the whole 10 frames are 23.4 mm in translation and 1.59 degrees in rotation, while in work [5] are 50mm and 10 degrees respectively.

As color information is extracted for VCSH generation, the stability with illumination changes is a crucial aspect, therefore 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 stability is evaluated by the differences between the estimated objects' VCSH under various illumination conditions and their target VCSH (correct object and pose) in database. As illustrated from Figure 8, 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. Mention 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.

4. CONCLUSION AND FUTURE WORK

In this paper, we presented a novel framework consisting of a global object descriptor *Viewpoint oriented Color-Shape Histogram*, which combines color and shape information for both object recognition and highly accurate object's pose retrieval. The proposed approach could be easily integrated into various robotic perception system for common objects fast recognition and 6D pose estimation, where no matter these objects are texture or textureless. A set of experiments is realized where our proposed approach is proven to outperform recent state-of-the-art methods by guaranteeing highly accurate object recognition, fast and accurate pose estimation as well as exhibiting the capability of dealing with environmental illumination changes. Future work will focus on the pose optimization and model building of wider-variety objects.

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

5. REFERENCES

[1] D. G. Lowe, "Distinctive image features from scale-invariant keypoints," *Int. J. of Computer Vision*, vol. 60, no. 2, pp. 91–110, 2004.

[2] H. Bay, A. Ess, T. Tuytelaars, and L. Van Gool, "Speeded-up robust features (surf)," *Computer Vision and Image Understanding*, vol. 110, no. 3, pp. 346–359, 2004.

[3] R. B. Rusu, N. Blodow, and M. Beetz, "Fast point feature histograms (fpfh) for 3d registration," in *in Proc. IEEE ICRA*, 2009.

[4] R. B. Rusu, G. Bradski, R. Thibaux, and J. Hsu, "Fast 3d recognition and pose using the viewpoint feature histogram," in *In Proc. of the Int. Conf. on Intelligent Robot Systems (IROS)*, 2010, pp. 3467–3474.

[5] J. Tang, S. Miller, A. Singh, and P. Abbeel, "A textured object recognition pipeline for color and depth image data," in *In Proc. of the Int. Conf. on Robots and Automation (ICRA)*, 2012.

[6] A.Vadivel, A.K.Majumdar, and S. Sural, "Perceptually smooth histogram generation from the hsv color space for content based image retrieval," in *Int. Conf. on Advances in Pattern Recognition*, 2003, pp. 248–251.

[7] C. B. Akgul, B. Sankur, F. Schmitt, and Y. Yemez, "Multivariate density-based 3d shape descriptors," in *IEEE Int. Conf. on Shape Modeling and Applications*, 2007, pp. 3–12.

[8] C. Choi and H. I. Christensen, "3d pose estimation of daily objects using an rgb-d camera," in *In Proc. of the Int. Conf. on Intelligent Robot Systems (IROS)*, 2012.

[9] A. Kanezaki, Z. Marton, D. Pangercic, T. Harada, Y. Kuniyoshi, and M. Beetz, "Voxelized shape and color histograms for rgb-d," in *Int. Conf. on Intelligent Robots and Systems (IROS), Workshop on Active Semantic Perception and Object Search in the Real World*, Sep. 2011.