2014 IEEE International Conference on Robotics & Automation (ICRA)
Hong Kong Convention and Exhibition Center
May 31 - June 7, 2014. Hong Kong, China

Autonomous Acquisition of Generic Handheld Objects in Unstructured Environments via Sequential Back–Tracking for Object Recognition

Krishneel Chaudhary, Yasushi Mae, Masaru Kojima and Tatsuo Arai

***Abstract*— Robots operating in human environments must have the ability to autonomously acquire object representations in order to perform object search and recognition tasks without human intervention. However, autonomous acquisition of object appearance model in an unstructured and cluttered human environment is a challenging task, since the object boundaries are unknown in prior. In this paper, we present a novel method to solve the problem of unknown object boundaries for handheld objects in an unstructured environment using robotic vision. The objective is to solve the problem of object segmentation without prior knowledge of the objects that human interacts with daily. In particular, we present a method that segments handheld objects by observing human-object interaction process, and performs incremental learning on the acquired models using SVM. The unknown object boundary is estimated using sequential back-tracking via exploitation of affine relationship of consecutive frames. The segmentation is achieved using identified optimal object boundaries, and the extracted models are used to perform future object search and recognition tasks.**

***Index Terms*— Handheld object segmentation, Incremental Learning, Sequential Back–Tracking (SBT), Support Vector Machine (SVM)**

I. INTRODUCTION

Autonomous learning and acquisition of object appearance models from the environment with no human intervention is fundamental for robots operating in human environments. For example, if a robot can learn from observation of human-object interactions, it can keep track of the objects being manipulated and alarm in case of dangerous or hazardous objects. The learning of new skills is achieved using object recognition, which allows robots to interact with objects using visual perceptions. Most research on object recognition focuses on offline learning process, where object appearance models are provided manually in advance to achieve robust performance, or are acquired from the environment by placing markers and tags on the objects [1]. Offline teaching and learning is a time-consuming process with additional training effort for new objects, while the placing of markers and tags on objects is also a time-consuming and unpractical task.

In this work, we consider an alternative to manually providing of object models by performing autonomous acquisition of object representations via observation of human actions. The goal is to equip the robot with an autonomous object acquisition and recognition module, so that the robot can discover new objects, learn their appearance models autonomously and perform object search and recognition tasks without supervision. This incremental learning capabilities would allow more robust performance, since errors can be corrected gradually over time, as more exemplary generic appearance models are accumulated in the dynamic environment. However, autonomous acquisition of object model poses difficulties in performing precise object segmentation, since the object boundaries and attributes are unknown.

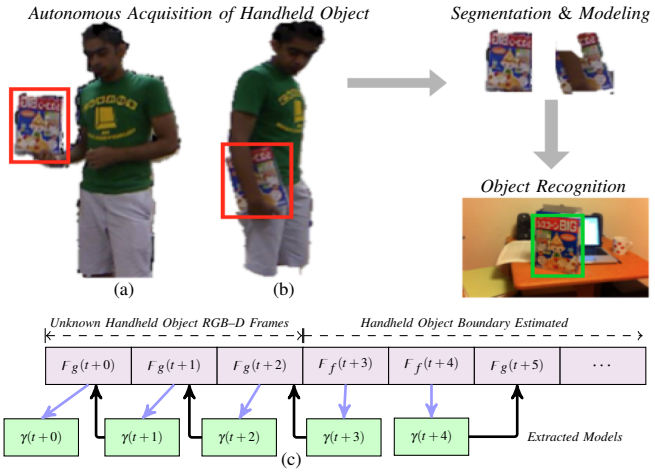


Fig. 1: Autonomous learning by observation of handheld object. a) Handheld object non-contiguous to the human body. b) Handheld object contiguous to the human body. c) SBT sequence diagram

In this paper, we present a novel method to perform autonomous acquisition of handheld object appearance models via observation of human-object interaction in an unstructured environment, as an alternative to manually providing of object models. More precisely, we solve the problem of unknown object boundaries and performing object segmentation without previous knowledge of the object being manipulated. Our approach uses the RGB-D data to segment the potential target object regions in a non-stationary and unstructured environment. The unknown object boundary is estimated when the handheld object is not in connection to the body space, Fig.1(a). The object is segmented and used to extract unknown handheld objects that are connected to the body space, Fig.1(b), by SBT i.e., we build generic handheld object models from future frames in which objects are not in connection to the body space and sequentially back–track to segment handheld object from the former frames, as shown in Fig.1(c), so that object informations are maximized. The segmented objects are accumulated as an appearance model, and are used for future object search and recognition tasks.

This work was supported in part by a Grant-in-Aid for Scientific Research (C) 23500242. K. Chaudhary, Y. Mae, M. Kojima and T. Arai are with Graduate School of Engineering Science, Osaka University. {krish, mae, kojima, arai}@arai-lab.sys.es.osaka-u.ac.jp

The remaining of this paper is organized as follows: Section II gives the overview of the related works. In Section III, we discuss the proposed Sequential Back-Tracking algorithm. The acquisition of handheld object models connected to the human body space and that of non-connected are discussed in Section IV and V respectively. The experimental results and discussions are provided in the Section VI, followed by conclusion in Section VII.

II. RELATED WORKS

Much research has been done in the area of computer vision for learning object models in controlled environments, but limited literature focuses on learning models from human actions in an unstructured environment.

The use of semi-supervised learning [2] reduces the need to provide a complete labeled training data. Welke et al. [3] presented a method for autonomous segmentation and modeling of objects placed in the hand of a ARMAR-III robot operating in a cluttered environment using RGB-D data. In their approach, the robot models the object held in its hands by rotating the object in front of its camera. The limitation is that the object modeling is performed only when the object is placed in robots hand. Roth et al. [4] presented a system for incremental learning through tracking of an unknown hand held object. They used background subtraction for object boundary detection and to initialize the MSER tracker. Background subtraction is suitable for stable objects and thus not robust in a non-stationary environment.

Incremental learning through human demonstration without any prior knowledge of the object has grown in recent years. Arsenio [5] used developmental learning on humanoid robots in natural environment for object learning through tracking and segmentation. The robot performs incremental learning of objects demonstrated by a human teacher and uses geometric hashing representation for storage. [6] and [7] presents an incremental learning of objects in a cluttered environment using RGB-D data focusing on object representation and its applications in object recognition on real time systems. The object's to be modeled is presented within the peripersonal space by a human teacher with the closest visual region selected as candidate region of the object. However, the limitation is the intentional demonstration of the objects to the robot, which requires human effort.

In contrast to the aforementioned research, we attempt to eliminate the need for intentional demonstration of objects, developing a module that learns and acquires object representations autonomously from observation of human actions. This enables a better adaption to the environment with gradual reduction of errors over time and also eliminates the need for a human intervention.

III. SEQUENTIAL BACK–TRACKING ALGORITHM

The segmentation of an unknown handheld object, denoted $\gamma(t)$ at any time $t \in \{1,...,T\}$, is a challenging task, since the object boundaries are unknown in prior, with added complexity of induced human motion and integration of environmental and human body cues. On a 2D image space $\gamma(t)$ is either connected to the body space, denoted $\digamma_g(t)$ or unconnected, denoted $\digamma_f(t)$ as shown in Fig.1(b) and (a) respectively. The estimation of unknown object boundaries are efficiently achieved in $\digamma_f(t)$, while in $\digamma_g(t)$ the separation of $\gamma(t)$ from the human body cues are difficult, because the depth information of the human body space and the object space are approximately same and therefore can not be separated as in $\digamma_f(t)$. We solve the problem of segmenting $\gamma(t+r)$ from $\digamma_g(t+r)$, where $r = \{0,1,...,k\}$[1] by first segmenting $\gamma(t+s)$ from $\digamma_f(t+s)$, where $s = \{1+k, 2+k, ..., p+k\}$[2], since the $\gamma(t)$ boundaries are efficiently estimated in $\digamma_f(t)$. In other words, $\digamma_g(t+r)$ and $\digamma_f(t+s)$ are decomposed into two separate temporal problems and segmentation is first performed for $\digamma_f(t)$ followed by $\digamma_g(t)$ over different time frame via back–tracking as shown in Fig.1(c).

The acquisition of object appearance model from $\digamma_f(t+s)$ *(see Sec.IV)* is achieved using the RGB-D data and the 3D skeleton hand joint position, $\xi_{(x,y,z)}$. The human-object region denoted $\lambda_f(t+s)$, are extracted, skin visual cues subtracted (Fig.2(b)), the boundaries of $\gamma(t+s)$ are estimated using kNN and $\gamma(t+s)$ is accumulated as the object appearance models of object Ψ_l (Fig.2(c)). The autonomous acquisition of $\gamma(t+r)$ from $\digamma_g(t+r)$ *(see Sec.V)* is realized using sequential back-tracking, which is defined as: all acquired frames $\digamma_g(t+r)$ with unknown object boundaries are solved for $\gamma(t+r)$ using $\gamma(t+s)$ acquired from $\digamma_f(t+s)$ via back–projection for estimation of unknown object boundaries in $\digamma_g(t+r)$ of object Ψ_l. The back–tracking is achieved via exploitation of affine relationship between consecutive frames and adequate handheld object segmentations are performed. The sequence diagram of SBT is shown in Fig.2.

In other words, $\gamma(t+s)$, are extracted from $\lambda_f(t+s)$ and accumulated as an appearance model of object Ψ_l at $t+\forall s$. The feature descriptors are computed from the acquired object appearance models and clustered to Bag of visual Word (BOW) model [8], which is trained using SVM (Fig.2(d)). Next, $\lambda_g(t+r)$ are extracted from $\digamma_g(t+\forall r)$ and categorized. In this work, we use an assumption, $\tilde{\mathbf{A}}$ that consecutive frames, i.e φ_{t+1} and φ_t are closely related by affine transformation for same object Ψ_l. Therefore the distinction between $\digamma_f(t+(1+k))$ and $\digamma_g(t+k)$ are relatively small. Based on $\tilde{\mathbf{A}}$ we used $\gamma(t+s)$ for estimating the optimal boundary of $\gamma(t+r)$ in $\lambda_g(t+r)$ as shown in Fig.2(f). The estimated boundary is used for extracting $\gamma(t+r)$ using graph-cut segmentation, and the accumulated models are used to perform incremental learning as shown in (Fig.2(g)).

IV. GENERIC MODEL EXTRACTION FROM $\digamma_f(t+s)$

In this section, we describe the method for boundary estimation and segmentation of unknown handheld object, $\gamma(t+s)$, when the object is not in connection to the human body space. The acquired object appearance models $\gamma(t+s)$ are then used in estimating the boundaries of $\gamma(t+r)$ from

[1] k is the total number of frame of $\digamma_g(t)$, with unknown object boundaries

[2] p is the total frame of $\digamma_f(t)$, and $p + k = n$ is the total number of frame in the human action observation sequence.

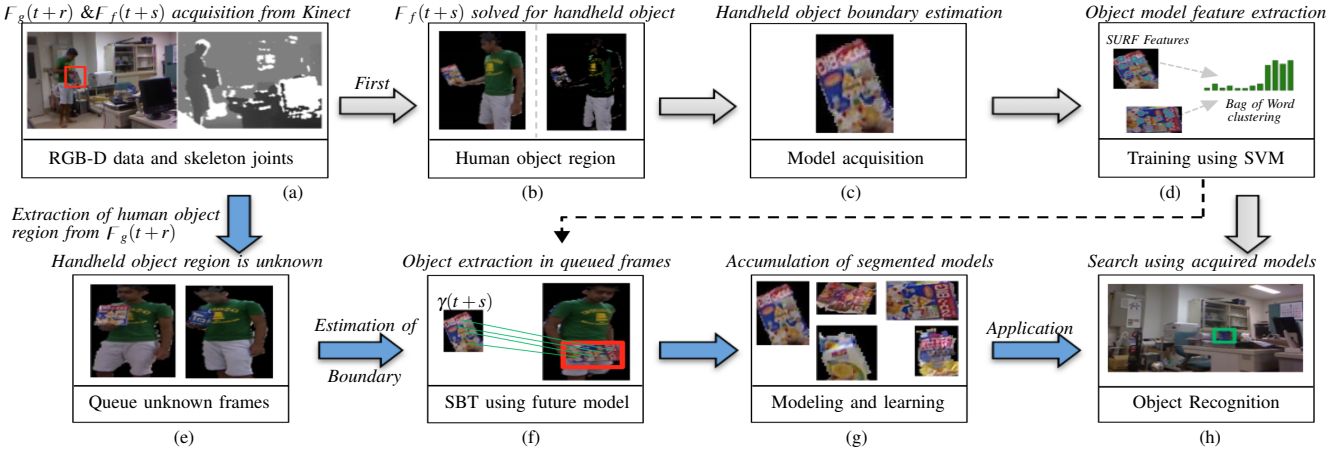


Fig. 2: Flow diagram of the Sequential Back-Tracking Method. a) Acquisition of RGB–D frame from Kinect with unknown objects. b) Extraction of human-object regions, from sequentially acquired frames. Objects connected to the body are queued (e) and solved for object boundaries after models are segmented from objects not connected to body space. c) Unknown boundary estimation and object appearance models segmentation. d) Feature extraction and training of all acquired models. f) Estimation of object boundaries in the unknown frames in (e) using SBT. g) Accumulation and learning of segmented object appearance models. h) Object recognition using the acquired models.

$\Gamma_g(t+r)$ for a complete sequence of Ψ_i via back-tracking, such that no preceding knowledge of the object is lost.

A. Detection and Extraction of Human–Object Region

The skeleton tracking capabilities of Kinect sensor are used to detect the presence of a human in the environment. Skeleton information in Kinect SDK is described by the length of the links and the joint angles. Specifically, it provides the three dimensional Euclidean coordinates, *(x, y, z dimensions)* and the orientation matrix of each joint with respect to the sensor.

The depth data refers to the distance of discrete pixels from the sensor, and is used for extracting the human-object, $\varsigma_f(t+s)$, region from the depth space, which is an efficient method of separating the cluttered environment, compared to RGB data where the color distribution has to be known. The extracted $\varsigma_f(t+s)$ is equivalently mapped to RGB space, and the corresponding human–object region, $\lambda_f(t+s)$ in RGB space is extracted. The skin visual cues θ_{ij} are subtracted from $\lambda_f(t+s)$ region and equivalently from $\varsigma_f(t+s)$, as shown in Fig.2(b).

$$\rho_f(t+s)_{depth} = \varsigma_f(t+s) - \theta_{ij} \tag{1}$$

In this work, the skin color subtraction was performed using the Skin Probability Maps described in [9].

B. Estimation of Handheld Object Boundary using kNN

The nearest neighbor classifier is one of the simplest methods of performing general, non-parametric classification [10]. We use *k*NN classification algorithm to estimate the optimal boundary on the depth data, $\rho_f(t+s)$, belonging to the $\gamma(t+s)$. The *k*NN classifier simply looks at the k points in the training set, Π that are nearest to the test point $\xi_{(x,y)}$, *(only known variable in RGB-D space)* and counts the number of each class in the set and returns a mathematical estimate of $\xi_{(x,y)}$ as measured by the distance metric.

$$p(\varpi = c|\xi_{(x,y)},\Pi,k) = k^{-1} \sum_{i\,\varepsilon N_{(\xi_{(x,y)},\Pi)}} \Gamma(\varpi_i = c) \tag{2}$$

where ϖ identifies the class c, $c \in \{1,...,G\}$, to which $\xi_{(x,y)}$ belongs. N are k nearest points to $\xi_{(x,y)}$ in Π. Γ is a boolean indicator function.

$\gamma(t+s)$ is identified to the correct contour in $\rho_f(t+s)$ belonging to the hand-object region using *k*NN. The training set vector is represented as:

$$\Pi_i \in < (x_1,y_1),(x_2,y_2),...(x_n,y_n) >$$

where (x_r,y_r) denotes the rth contour coordinate in the image domain. k is experimentally evaluated by choosing different values of k on randomly selected frames based on $\xi_{(z)}$, and angular human variation for possible human poses. k = 3 with lowest error rate was used in the experiment.

C. Feature Extraction and Learning

The features are extracted using the Speed Up Robust Feature (SURF) [11] local feature descriptors. The SURF descriptor is preferred over other descriptors due to its concise descriptor length, scale invariance and robustness to rotation. The SURF descriptor defines a 128-dimensional feature vector for each interest point in the image. The feature vectors of $\gamma(t+s)$ and $\lambda_g(t+r)$ are computed by SURF and used for estimating the unknown hand–object boundaries in $\lambda_g(t+r)$, as shown in Fig.2(f).

The extracted local features, $\hbar(\vec{i},j)$, from the images are vector quantized into a BOW model [8], so that each image is represented using a fixed length feature vector. In this work, we built a dictionary size of 150 clusters, such that each image is represented by a feature vector, $\Phi(\vec{i},j)$ of length [150 x 1]. The extracted feature vectors, $\Phi(\vec{i},j)$ are trained using SVM [12], [13]. The Radial Basis Function (RBF), in

(3) was chosen as a kernel function of SVM based on the good optimized results compared to other kernel functions *(only results of RBF are shown due to space limitation)*.

$$K(x,x^{'}) = \exp(-\frac{\|x-x^{'}\|^2}{2\sigma^2}) \quad (3)$$

where σ is the degree of generalization applied to the training set. As the incremental learning of the object models increases, the σ is reduced. Next, all $\gamma(t+s)$ are categorized and labeled. Labeling in this work is done numerically and is incremented for every new object not in the database.

V. OBJECT MODEL SEGMENTATION FROM $F_g(t+r)$

The human–object region, $\lambda_g(t+r)$ are extracted from $F_g(t+r)$ and used for estimating the unknown boundary of $\gamma(t+r)$ connected to the human body space.

A. Estimation of Unknown Object Boundary in $\lambda_g(t+r)$

After classification, the corresponding boundary, ψ_i of $\gamma(t+r)$ in $\lambda_g(t+r)$ is determined using Random Sample Consensus (RANSAC) algorithm [14]. RANSAC algorithm is used to exclude outliers from a large data set and randomly select the best inliers of model $\gamma(t+s)$ to $\lambda_g(t+r)$. Note, we define a square search space of length $\bar{\chi}$ (4) on $\lambda_g(t+r)$ centered at $\xi_{(x,y)}$ circumscribing hand–object neighborhood using knowledge of $\gamma(t+s)$ for better boundary estimation.

$$\bar{\chi} = 2 \times max(l,w) \quad (4)$$

where l is the length of $\gamma(t+s)$ and w is width of $\gamma(t+s)$. Based on assumption $\bar{\mathbf{A}}$, the region of $\gamma(t+r)$ in $\lambda_g(t+r)$ corresponding to the group of good match inliers is derived using perspective transformation, and the respective object model, $\gamma(t+r)$ is segmented.

However, the extracted model may be integrated with visual cues of human body region or the estimated ψ_i maybe larger than the optimal object boundary, as shown in Fig.3(a). Such integrations increases the misclassification rates when used for recognition, therefore are eliminated. The difficulty lies in the identification of the true object and the true integrated cue region for multi-colored objects that are to be separated. To solve this, we treat the segmented model as an energy minimization problem [15], which is efficiently solved using graphcut segmentation approach.

B. Learning Object Color Model and Integrated Cues

The back–tracking method exploits the affine relationships between the frames based on assumption $\bar{\mathbf{A}}$, therefore the color distribution of $\gamma(t+s)$ consisting of pixels $a^f_{(i,j)}$ can be characterized by a Gaussian Mixture Model (GMM) (5).

$$p(x|\lambda) = \sum_{i=1}^{M} w_i g(x|\mu_i,\Sigma_i) \quad (5)$$

where x is the feature vector, M is the number of Gaussian components, w_i is the weight component, μ_i and Σ_i are mean and covariance matrix respectively. The foreground $\Gamma_i = 1$, and the background $\Lambda_i = 0$, color models are built in $\gamma(t+r)$

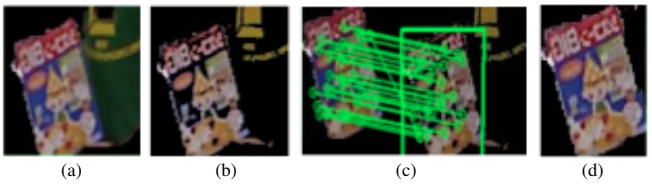


Fig. 3: a) Segmentation using initial estimated boundary. b) Object color model construction using GMM. c) Optimal boundary estimation. d) Segmentation using optimal boundary

by learning the RGB color distribution of $a^f_{(i,j)}$ in $\gamma(t+s)$ using the GMM, shown in Fig.3(b). The learning process is performed once for every new object and achieved using an Expectation Maximization algorithm [16]. Since the majority of pixels, $a^f_{(i,j)}$ in $\gamma(t+r)$ belongs to the handheld object, the classification of Γ_i and Λ_i is based on the assumption that the larger vector contains the handheld object model distribution. However, sudden changes in illumination in the uncontrolled environment may affect the RGB color distribution in $\gamma(t+s)$ and $\lambda_g(t+r)$, which may result in false clustering of $a^f_{(i,j)}$ to Γ_i or Λ_i.

Consequently, when the segmented boundary ψ_i is larger than the optimal object boundary ν_i with the integrated cues, incorrect segmentation occurs due to erroneous minimization. To solve the aforementioned problem, we used SURF to abetment graphcut in eliminating additional cues while maintaining $\gamma(t+r)$ boundaries. Basically, SURF is used to redefine the ψ_i when the initial ψ_i is larger than the ν_i, such that erroneous cues are minimized as shown in Fig.3(c).

C. Segmentated Object Model Evaluation

Features $\hbar_f(\vec{i},j)$ of $\gamma(t+s)$ are computed and matched first with features $\hbar_\Gamma(\vec{i},j)$ of Γ_i. This is because our assumption is that Γ_i contains the object model. The evaluation of the feature match between the 2 images is defined as a ratio, $\vec{\Omega}$:

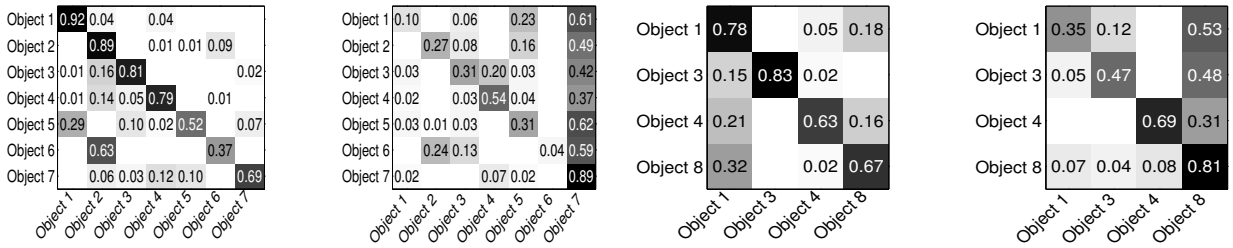
$$\vec{\Omega_i} = \frac{\kappa^i_{miss}}{\kappa^i_{match}} \quad (6)$$

where κ_{miss} is the features with different labels in the 2 images and κ_{match} is the matched features in the 2 images. If $\vec{\Omega}$ is greater than the threshold, τ than the object model is in Λ_i, hence the features $\hbar_f(\vec{i},j)$ are matched to features of Λ_i, $\hbar_\Lambda(\vec{i},j)$ and the new boundary is estimated, the object segmented and the method is repeated until the optimal object boundary is achieved, i.e when $\vec{\Omega_{i-1}} = \vec{\Omega_i}$. The object model is segmented using the computed ν_i, see Fig.3(d).

VI. RESULTS

A. Setup and Initialization

The experiment was performed using Microsoft Kinect sensor for data acquisition; it outputs RGB-D frame of 640 x 480 pixels at 30Hz with a workable range of 1-3m. The experimental data was acquired in 2 rooms with different structure and illumination intensities as shown in Fig.6. Note that *none of the objects were known to the system in prior.*



(a) Room 1 light on (b) Room 1 light off (c) Room 2 light on (d) Room 2 light off

Fig. 4: Leave-one-out cross-validation confusion matrix of the handheld objects acquired using sequential back-tracking from: (a) Lab room 1 with lights on. (b) Lab room 2 with the illumination intensity of approx. 40%. (c) Lab room 2 with lights on and (d) Lab room 2 with illumination of approx. 60%. All acquired objects representations shown in Fig.5 are labeled numerically.

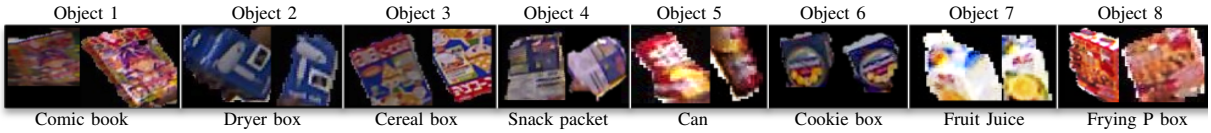


Fig. 5: Object appearance model samples acquired using sequential back-tracking. These are the 8 objects used in the experimentation

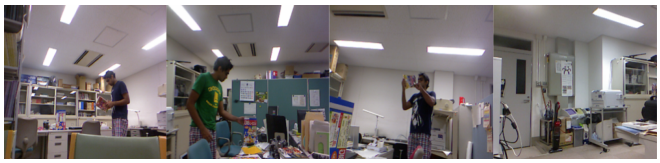


Fig. 6: Sample frames showing the environmental structure where the experiment was performed.

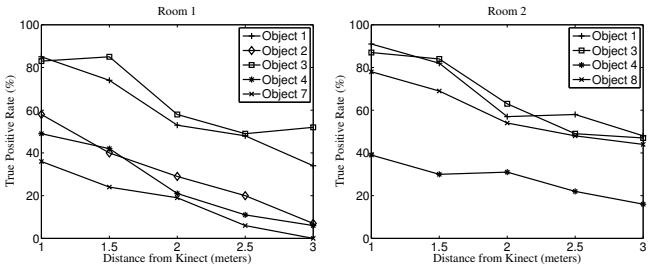


Fig. 7: Object recognition performance using the acquired handheld object models with lights-on in: left) Room 1 where 5 objects were used, right) Room 2 where 4 objects were used.

B. Evaluation of Acquired Objects Models

First the experiment was performed with all lights turned on, and then we turned some lights off so that the generality of the implemented system could be shown. During experimentation, the segmented models of size greater then 30 x 30 pixels *(condition α)* were acquired, since it contains distinctive features for good modeling performance. Samples of selected models accumulated via SBT are shown in Fig.5.

The acquired models were used for validation. However, during autonomous acquisition of the handheld object, some acquired models satisfied condition α, but were results of partial occlusion or incorrectly calculated object boundary therefore acquiring truncated object appearance models. These models create ambiguity among other object models, significantly degrading performance. For the purpose of this work, model selection was done basically by selecting models with features greater than threshold θ_i for each object.

$$\theta_m = \frac{max(\forall_p(\hbar_p(\vec{i},j)))}{2} \tag{7}$$

where m is the number of different objects and p is the number of acquired models of the same label m.

The validation using a confusion matrix constructed from leave-one-out cross validation of the acquired generic handheld object appearance models is shown in Fig.4. Note in this section for validation of acquired object models, the features were computed using SURF and RGB color histogram descriptors. Matrix (a) in room 1 with the light on, the large misclassification of object 6 is evident due to few distinguishable features. Other models sustains distinctive features, making exemplary object representations. Matrix (b) with reduced illumination in room 1 the misclassification among objects increased significantly. The partial reflective material of object 4 reflects light making it distinctive from other objects while, object 7 in reduced illumination is more distinctive from other objects due to its color composition. Matrix (c) with light on in room 2, the ambiguity between object 1 and 8 is evident as both objects have similar color and texture characteristics. Matrix (d) with reduced illumination in room 2, object 8 was more distinctive while misclassification of objects 1, 2 increased similar to room 1

C. Object Search and Recognition

The acquired models were used to perform recognition of the objects in the cluttered environment. We evaluated the success rate at different distance of the object from the sensor as shown in Fig.7. Objects 5 and 6 was not used in recognition, as the success rate was very low due to small object size and very few distinctive features as distance varied. As shown in Fig.7, the success rate in both rooms degrades

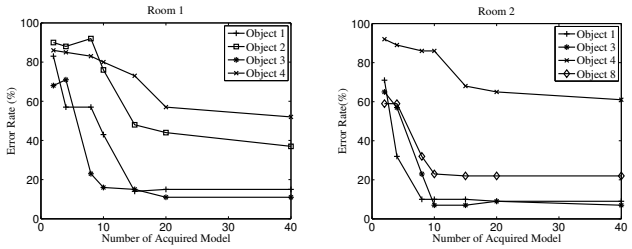


Fig. 8: Incremental learning error rate of first 40 models of each objects acquired using Sequential Back–Tracking in 2 rooms.

significantly as the distance from the sensor increases. This results as increase in distance increases ambiguity between the true object region and the environmental clutters, since the object size decreases, hence increase in misclassification rate. Note that at distance of 2–3m, experiment was performed on a less cluttered environment. Fig.8 shows the result of errors as model acquisition increases. Objects with high feature attain significant drop in error rate while smaller objects had gradual but slow decrease in error.

D. Discussion

Overall, the performance of SBT was better when the lights were turned on with average success of 73% in correctly estimating the unknown object boundaries without truncation for each handheld object.

In this work, the human action based object modeling was performed for a one individual in the environment with only one handheld object. We also made an assumption that the object and the human clothing color distribution are not same. In case of similar color distribution, the object cannot be separated from the background and therefore the optimal object boundary cannot be determined. This is because the addition cues are considered as part of the object and hence the energy minimization function fails to correctly identify the object boundary from the integrated cue's.

Moreover, in the current implementation we also placed a constrain of condition α. As part of the future prospect, we plan to reduce the parameter of condition α, so that smaller household objects such as cups, could be accumulated as from human actions at variable distances. Furthermore, we plan to extend autonomous acquisition module of SBT to acquire appearance models from multiple persons as well as accumulation of multiple objects in human hands.

VII. CONCLUSION

In this paper, we presented a novel method for autonomous acquisition of object appearance models by observation of handheld objects in an unstructured environment for object recognition. We considered the problem of offline training for object recognition that is not only time consuming but also models have to be provided manually. In our approach, we focus on performing incremental learning of objects autonomously over time. The problem of unknown object boundaries was solved using *k*NN for objects not connect to the human body space, while objects connected to the human body space was solved using proposed sequential back–tracking. We performed experiments using unknown objects in 2 different rooms with different structure and illumination conditions. The acquisition rates were higher with the lights turned on and for objects with rich textures. On average SBT successful estimated 73% of unknown boundaries without truncation. The acquired models were used for object recognition in cluttered environments; at closer distances, the recognition rates were higher, while at increased distances, the environment clutters affected the recognition rate significantly. Finally we believe that the proposed system is a powerful tool for autonomous learning of object models by robots for search and recognition purposes, since no previous knowledge about the object is required.

REFERENCES

[1] J. Wu, A. Osuntogun, T. Choudhury, M. Philipose, and J. Rehg, "A scalable approach to activity recognition based on object use," in *Computer Vision, 2007. ICCV 2007. IEEE 11th International Conference on*, 2007, pp. 1–8.

[2] C. Penaloza, Y. Mae, K. Ohara, T. Takubo, and T. Arai, "Generic object classifiers based on real image selection from the web," in *Pattern Recognition (ACPR), 2011 First Asian Conference on*, 2011, pp. 239–243.

[3] K. Welke, J. Issac, D. Schiebener, T. Asfour, and R. Dillmann, "Autonomous acquisition of visual multi-view object representations for object recognition on a humanoid robot," in *Robotics and Automation (ICRA), 2010 IEEE International Conference on*, 2010, pp. 2012–2019.

[4] P. M. Roth, M. Donoser, and H. Bischof, "On-line learning of unknown hand held objects via tracking," in *Int. Conf. on Computer Vision Systems*, 2006.

[5] A. Arsenic, "Developmental learning on a humanoid robot," in *Neural Networks, 2004. Proceedings. 2004 IEEE International Joint Conference on*, vol. 4, July 2004, pp. 3167–3172 vol.4.

[6] H. Wersing, S. Kirstein, M. Gtting, H. Brandl, M. Dunn, I. Mikhailova, C. Goerick, J. J. Steil, H. Ritter, and E. Krner, "Online learning of objects in a biologically motivated visual architecture." *Int. J. Neural Syst.*, vol. 17, no. 4, pp. 219–230, 2007.

[7] C. Goerick, I. Mikhailova, H. Wersing, and S. Kirstein, "Biologically motivated visual behaviors for humanoids: Learning to interact and learning in interaction," in *Proceedings of the IEEE/RSJ International Conference on Humanoid Robots (Humanoids 2006), Genoa, Italy*. IEEE/RSJ, 2006.

[8] G. Csurka, C. R. Dance, L. Fan, J. Willamowski, and C. Bray, "Visual categorization with bags of keypoints," in *In Workshop on Statistical Learning in Computer Vision, ECCV*, 2004, pp. 1–22.

[9] G. Gomez and E. F. Morales, "Automatic feature construction and a simple rule induction algorithm for skin detection," in *In Proc. of the ICML Workshop on Machine Learning in Computer Vision*, 2002, pp. 31–38.

[10] K. P. Murphy, *Machine Learning: A Probabilistic Perspective*. Mit Press, 2012.

[11] H. Bay, T. Tuytelaars, and L. V. Gool, "Surf: Speeded up robust features," in *In ECCV*, 2006, pp. 404–417.

[12] O. Chapelle, P. Haffner, and V. Vapnik, "Support vector machines for histogram-based image classification," *Neural Networks, IEEE Transactions on*, vol. 10, no. 5, pp. 1055–1064, 1999.

[13] C.-C. Chang and C.-J. Lin, "LIBSVM: A library for support vector machines," *ACM Transactions on Intelligent Systems and Technology*, vol. 2, pp. 27:1–27:27, 2011.

[14] M. A. Fischler and R. C. Bolles, "Random sample consensus: a paradigm for model fitting with applications to image analysis and automated cartography," *Commun. ACM*, vol. 24, no. 6, pp. 381–395, June 1981.

[15] Y. Boykov and M.-P. Jolly, "Interactive graph cuts for optimal boundary amp; region segmentation of objects in n-d images," in *Computer Vision, 2001. ICCV 2001. Proceedings. Eighth IEEE International Conference on*, vol. 1, 2001, pp. 105–112 vol.1.

[16] T. Moon, "The expectation-maximization algorithm," *Signal Processing Magazine, IEEE*, vol. 13, no. 6, pp. 47–60, 1996.