MVA2015 IAPR International Conference on Machine Vision Applications, May 18-22, 2015, Tokyo, JAPAN

A novel multi modal tracking method based on depth and semantic color features for human robot interaction

Aswin Chandarr, Maja Rudinac, Pieter Jonker
Vision Based Robotics, Faculty of Mechanical Engineering
Delft University of Technology

Abstract

In this paper we tackle the challenges of visual tracking for personal robots. We have proposed a novel track-by-detection method that combines a semantic object model with depth properties to obtain target contours. The tracking can be initialized by either 2D or 3D inputs, which are further refined using clustering based background removal to obtain an initial object model. During tracking, we propose to refine the search space by metric constancy and removal of the support plane. Further, the target appearance is modeled using a semantic human centric color descriptor, continuously updated by an online learning algorithm. The spatial compactness of the target is described using a Gaussian model with an initially determined variance. A fusion of the obtained color and depth models based on a target-background dissimilarity measure, is used to perform segmentation based tracking using graph cuts to obtain object contours. The experimental results in a household scenario show a good performance of the algorithm in challenging conditions such as scale, viewpoint change and out of plane rotations.

1 Introduction

Among the major challenges that personal robots face are dynamic learning and exploration of the human environments. To be able to adapt to new environments and users, robots have to learn about objects and human activities and also act in a continuously changing conditions. These require a robust method for object and person tracking, invariant to scale, viewpoint changes and out of plane rotations. Effective manipulation and learning of objects requires acquiring continuous object contours in addition to a target position provided by tracking methods. The challenges posed by these conditions on current state of the art tracking methods are elaborated below.

There are several state of the art 3D tracking methods using pointclouds [1], but they require an a-priori 3D mesh with particle filter which is not suitable for a robot's exploration of unknown objects. In addition, methods based on particle filters do not scale with larger and deformable objects.

The most efficient 2D state of the art tracking algorithms are based on a closed loop track, detect and learn framework [2]. They use local features and optical flow which leads to failure and drift in case of uniformly colored objects and out of plane rotations, commonly encountered by personal robots.

Algorithms like [3] overcome these problems by using an incrementally learning structured output SVM and part based features. Apart from not utilizing the additional depth information available, they only track bounding boxes over the targets and do not provide contours, required for manipulation tasks. Several methods have been proposed that utilize depth in addition to 2D models [4], however, they do not provide contour information and do not tackle the problems of out of plane rotations and large viewpoint change very well.

To obtain object contours several methods have been proposed ([5], [6]). Deformable part based models have been utilized by [5], while [6] uses a graph cut method fusing high level (object detection) and low level (color/motion) features. However, these methods consider only the properties of the target and are not adaptive to varying target background visual characteristics.

In this paper we introduce a robust tracking algorithm that provides contours with multi modal features that can also be used for semantic object recognition [7]. The proposed *track by detection* algorithm shown in Figure 1, uses online learning, semantic color feature description combined with depth to provide target contours while tackling challenging conditions. We propose a background elimination and search space refining step followed by a novel optimal fusion of color and depth information in a graph cut methodology to obtain segmentation. Continuous learning and update of target models decreases drift and results in the generation of a complete feature space of the target.

Our algorithm tackles the common conditions encountered by personal robots such as scale, viewpoint change and out of plane rotation with an algorithm invariant to visual object attributes. The used semantic human centric color description and the generation of target contours can be used for object recognition, manipulation and higher level human robot interaction. The rest of the paper is organized as follows. Section 2 describes target initialization followed by search space refining in section 3. Section 4 explains the color and depth target modeling and segmentation and learning using the optimal fusion of these modalities is detailed in Section 5. The experimental setup and the results are shown in Section 6 followed by conclusions and future work.

2 Initial target selection

Target initialization is the first and critical step as it provides the base model for the entire tracking process.

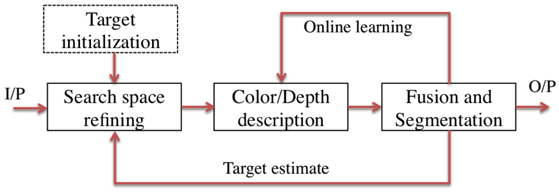

Figure 1: Tracking framework

We use a 2D bounding box for initial target selection as required by human robot interaction applications such as user initiated grasping as well autonomous grasping based on 2D object detection methods. However, the area within the bounding box can include parts of the background, which can cause drift during the tracking process. Given that tracking targets for robotic applications consist of spatially coherent objects, we minimize drift and obtain an accurate initial model. An Euclidean clustering is perfromed on the points in the cloud within the initial bounding box and the dominant cluster is modeled as the initial target.

3 Search space refining

Once the initial model is selected, the object position has to be estimated over the consecutive frames. The knowledge of the target position in the previous frame is used to restrict the search area for detection in the next frame. Additionally, certain geometric properties of the scene enables masking of parts of the scene as definite background. We further explain how these two properties are used to refine the search space of the object for the proposed tracking by detection mechanism.

While the size of the target in pixels can vary based on scale, the metric size of the target (m) is a constant. We exploit the variation of the pixel boundaries based on the metric size, distance and the camera intrinsics as described in the (1). This relates spread of the search space in x, y directions with the focal length (f) with the mean depth of the last frame (μ_{depth}), enlarged by a scale factor s. A scale factor of 2 has been used as object size cannot increase beyond twice between two frames.

$$spread_{new} = \frac{s \times f \times target_{size}}{\mu_{depth}} \tag{1}$$

The next image frame is cropped to $spread_{new}$ to restrict the search space, which is further refined as follows.

It can be also observed that the targets we are interested in are mostly placed on support surfaces such as a table, floor etc. In human environments, these are always planar regions, specifically perpendicular to the target. We model the target orientation with a surface normal (N_{target}) perpendicular to the dominant planar component of the target. Large planar regions in the scene oriented normally to the N_{target} are localized using a RANSAC based plane fitting. They are removed from the source image to obtain a final search space mask. An example result of this process can be seen in Figure 2. Once the search space is obtained, each point here should be assigned a probability of belonging to the target. This is explained in the next section.

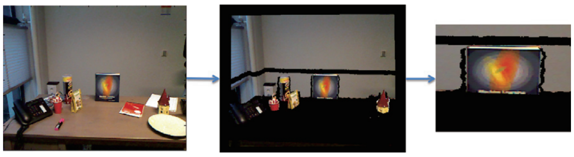

Figure 2: Refining the search space. The planar regions perpendicular to target are removed and search space is restricted based on metric consistency from previous frame

4 Target description

Given the initial target region, dense maps can be obtained for the consecutive frames that describe the probability of each point belonging to the object based on different modalities. In order to track both uniform and textured objects, we have combined semantic color appearance with depth properties.

4.1 Color description

A semantic color appearance model provides robots with perception of colors, similar to human understanding. Color Naming System (CNS) derived from cognitive psychology studies provides a set of 11 basic colors, common to most languages. In a previous research [8], a pixel based probabilistic mapping from the RGB to CNS has been obtained from natural images obtained from online search using a PLSA method. This has been used with reduced dimensions as a feature for tracking in [9] and also for object recognition in [7]. We use this as a color descriptor to obtain an illumination invariant semantic representation of the object. This is significant as it allows for multiple view semantic object recognition and higher level cognitive human robot interaction in the future. We obtain a normalized 11 dimensional feature at each point. A set of positive and negative feature samples are obtained by sampling from the known target and background regions. We learn a discriminative model separating these regions on an 11 dimensional manifold using AROW (Adaptive Regularization of Weights) a fast second order online learning algorithm [10]. This provides a confidence value for each point belonging to the target model. This confidence output is normalized as in (2) to describe the probability of the point belonging to the target (Pr_c).

$$Pr_c = \frac{c - min(C)}{max(C) - min(C)} \tag{2}$$

where, c is the classifier output at a particular pixel. This is further combined with the depth properties of the target.

4.2 Depth description

Since the whole object is considered as a single spatially coherent entity, we use only depth data to represent the spatial compactness of a target. Given the initial region, the target depth is modeled as a single dimensional gaussian parametrized by μ and σ obtained from the initial model. With this, a dense probability map (Pr_d) of the search space is obtained using (3)

$$Pr_d = e^{K\left(\frac{D-\mu}{2\sigma}\right)^2} \tag{3}$$

where D is the depth at a given point K describes the steepness of the target background boundary. Hence Pr_d represents the probability of a point belonging to the target model based on its current depth properties. This kind of depth model assigns higher costs to any other object occluding the target, thereby making the tracking robust to occlusion. Having obtained a probability map based on color and depth, we combine them optimally and use it for segmentation of the search space as explained in the next section.

5 Object detection and learning

The obtained probability maps have been used to obtain a segmentation of the target and the background using a graph-cut formulation [11]. In this approach, the search space is modeled as a markov random field and the segmentation is the configuration of the latent variable (target or background) corresponding to the minimal energy. The solution provides an optimal segmentation considering the data cost (D_c) which represents the cost of pixel (i) labeled to its components (l_i: target/background) based on the observations at the pixel (x_i: color/depth) and a smoothness cost (S_c) which enforces neighboring pixels (Nbr_i) have the same labeling. The solution is the minimization of the total energy $E(x)$ in the Equation (4).

$$E(x) = \sum_{i \in S_s} \left(D_c(l_i|x_i) + \sum_{j \in Nbr_i} S_c(x_i, x_j) \right) \tag{4}$$

The smoothness cost is assigned to a 4-connected neighborhood based on their relative euclidean distances which is proportional to their depths. This cost between neighboring pixels (x, y) defined in (5) enforces spatial coherence in the segmentation process. $Depth_x$ is the distance value of the pixel (x) from the camera.

$$S_c = \frac{1}{|Depth_x - Depth_y| + \varepsilon} \tag{5}$$

The D_c is obtained by an optimal combination of the color and depth probabilities. A target-background dissimilarity measure (Ds) is used for adaptability to different kinds of targets in any environmental conditions,. This is obtained as distances between the normalized histograms representing visual characteristics of target and background. We use the Bhattacharyya distance which provide a measure in the range $[0, 1]$. The spatial distribution(DH_{tg}, DH_{bg}) of the components is represented using histograms of the depth map of target and background quantized into 10cm bins. The chromatic distribution (CH_{tg}, CH_{bg}) is obtained using a 168 dimensional histogram of the image quantized in HSV space as used in [12]. The dissimilarity Ds_c, Ds_d is obtained using a Bhattacharyya distance between target and background histograms. For the compatibility with graph cut, the color and depth probability maps are inverted and scaled to penalties (Pen_c, Pen_d) and fused into a single data cost (D_c) by using normalized weights (w_c, w_d) which are obtained as in (6)

$$\begin{aligned} Pen &= A\,(1 - Prb) \\ D_c &= w_c Pen_c + w_d Pen_d \\ w_c &= \frac{Ds_c}{Ds_c + Ds_d},\; w_d = \frac{Ds_d}{Ds_c + Ds_d} \end{aligned} \tag{6}$$

where, A is a scale to achieve uniformity between penalties for different modalities. Now, having obtained both D_c and S_c, we obtain the segmentation T_{new} of the target by applying the min-cut [11] followed by a few morphological operations. An illustration of the different potentials over the entire frame is shown in Figure 3 where the penalty value increases from blue to red color. To incrementally learn the color model, the regions of color probability map (Pr_c) that have a low value, but still are present in the new segmentation (T_{new}) are updated to the online learner. This ensures that we have a complete model during object exploration applications. To minimize the drift, the μ of the depth model is updated while the σ is kept constant at the initial value. All these components connected together in the framework of Figure 1 result in continuous target tracking. The performance of the algorithm is evaluated as explained in the next section.

6 Experimental setup and Results

The tracking system is targeted for personal/service robots operating in indoor environments using RGB-D sensors. Currently there does not exist a dataset for scenarios encountered by personal robots for object exploration, learning and manipulation. We have implemented the tracking system on a service robot [12] and in order to realistically evaluate the performance of the system, we tested it on a challenging dataset consisting of 7582 frames divided into 15 categories of household objects in various conditions. The data was created using a Microsoft Kinect with a VGA resolution and 15fps. Some of the objects used in the dataset are shown in the Figure 4. We test the tracking performance during conditions of scale change, viewpoint variation and out-of-plane rotations taken in a real household setting with varying illumination conditions. An experiment was performed with a relative scale change between robot and target during the tracking and a viewpoint change from -60^o to +60^o.

6.1 Results

The tracking performance is evaluated by using the overlap criteria defined by (7) as used in [13]

$$precision = \frac{match_{area}(bb_{truth}, bb_{tracked})}{area(bb_{tracked})} \tag{7}$$

The performance measures reported are the percentage of frames that have a precision greater than 0.5 when using different modalities. Table 1 shows the general (overall) tracking performance. Using only color provides a mean precision of 78 % and using only depth has a precision of 72%. When only depth is used, the performance decreases due to the drift of tracked target into nearby regions with similar depth. But when coupled with color information it can be seen that this drift is minimized and we obtain an enhanced tracking performance of 84% which is comparable with the state of the art tracking methods [4]. Table 2 shows the mean precision at different viewpoints which also account for out of plane rotations. It can be seen that the performance when using only color decreases with large viewpoint change. This is due to the large variance of

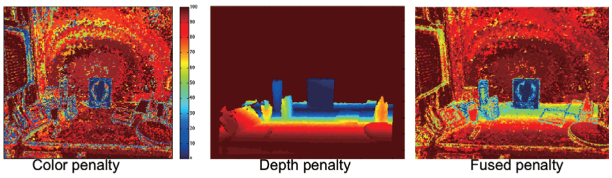

Figure 3: Color, depth and fused potentials

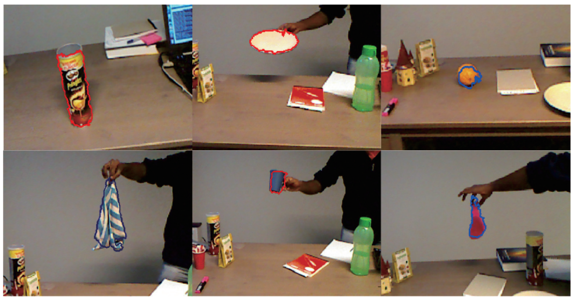

Figure 4: Sample tracked contours

Table 1: Overall tracking performance

	Color	Depth	Color + Depth
Precision (%)	78	72	84

illumination conditions as well different color distributions over the target periphery. Since the depth distribution remains invariant to the change in viewpoints, attributed to the spatial coherence of the targets, combining color and depth provides a good tracking performance inspite of the challenging viewpoint changes.

Table 2: Viewpoint change performance

Viewpoint	-60^o	-30^o	0^o	30^o	60^o
Color (%)	68	72.5	78	74	69
Color + Depth (%)	73	76	85	77.5	72

The influence of scale change on the tracking performance is quantified in the Table 3. When the object is very far from the robot, the target has only very few pixels which are not discriminative enough from the background in the 11 dimensional feature space. This results in lower tracking precision at farther scales while using only color model. It can also be seen that addition of depth properties considerably increases the performance validating the robustness of the algorithm. The algorithm has been successfully implemented and tested on a personal robot [12].

Table 3: Scale change performance

Scale	1(Far)	2	3	4 (Near)
Color (%)	63	67	73	76
Color + Depth (%)	70	73	78	83

7 Conclusions

In this paper a novel algorithm for object and person tracking is introduced for human robot interaction tasks. After the target is initialized using a bounding box over RGBD data the final target model is obtained by using a clustering algorithm. The search space of the consecutive frames is refined using the metric size constancy of the objects and also by removing the planar support surfaces. The object appearance has been modeled using both color and depth modalities. A color feature based on CNS has been used with a classifier to obtain a color probability map while the depth probability map is obtained by using a gaussian model of the object with fixed initial variance. A novel method to perform optimal fusion of different object modalities using a target-background dissimilarity measure has been introduced. This has been used in a graphcut framework to obtain the object contour. Extensive experiments have been performed in a household environment showing good performance under challenging conditions of viewpoint and scale change as well as out of plane rotation. This method, combined with an online color learning mechanism can be applied to robotic object exploration and recognition to obtain a complete appearance model of the object. Additionally the used color features can be simultaneously used for multiple view semantic object recognition, leading to high level human robot interaction in the future.

References

[1] C. Choi and H. Christensen, “Rgb-d object tracking: A particle filter approach on gpu,” in *Intelligent Robots and Systems (IROS)*, Tokyo, Nov 2013, pp. 1084–1091.

[2] Z. Kalal, K. Mikolajczyk, and J. Matas, “Tracking-learning-detection,” *Pattern Analysis and Machine Intelligence*, 2012.

[3] S. Hare, A. Saffari, and P. H. S. Torr, “Struck: Structured output tracking with kernels.” in *ICCV*. IEEE, 2011, pp. 263–270.

[4] S. Song and J. Xiao, “Tracking revisited using rgbd camera: Unified benchmark and baselines,” in *Proceedings of the 2013 IEEE ICCV 2013*.

[5] M. Godec, P. M. Roth, and H. Bischof, “Hough-based tracking of non-rigid objects,” in *Proc. International Conference on Computer Vision (ICCV)*, 2011.

[6] A. Bugeau and P. Pérezz, “Track and cut: Simultaneous tracking and segmentation of multiple objects with graph cuts,” *J. Image Video Process.*, 2008.

[7] A. Chandarr, M. Rudinac, and P. Jonker, “Multi-modal human centric object recognition framework for personal robots,” in *IEEE-RAS International Conference on Humanoid Robots*, 2014.

[8] J. Van De Weijer, C. Schmid, J. Verbeek, and D. Larlus, “Learning color names for real-world applications,” *Image Processing, IEEE Transactions on*, 2009.

[9] M. Danelljan, F. Shahbaz Khan, M. Felsberg, and J. van de Weijer, “Adaptive Color Attributes for Real-Time Visual Tracking,” in *Proceedings of IEEE CVPR 2014*.

[10] S. C. Hoi, J. Wang, and P. Zhao, “Libol: A library for online learning algorithms,” *Journal of Machine Learning Research*, vol. 15, pp. 495–499, 2014.

[11] Y. Y. Boykov and M.-P. Jolly, “Interactive graph cuts for optimal boundary & region segmentation of objects in n-d images,” 2001.

[12] M. Rudinac, “Exploration and learning for cognitive robots,” Ph.D. dissertation, Delft University of Technology, 2013.

[13] X. Wang, M. Rudinac, and P. Jonker, “A robust real-time tracking system based on an adaptive selection mechanism for mobile robots,” in *Control Automation Robotics Vision (ICARCV), 2012 12th International Conference on*, Dec 2012, pp. 1065–1070.