3D Shape Reconstruction from a Humanoid Generated Video Sequence

P. A. Martínez[1] D. Varas[2] M. Castelán[1] M. Camacho[2] F. Marques[2] G. Arechavaleta[1]

***Abstract*— This paper presents a strategy for estimating the geometry of an interest object from a monocular video sequence acquired by a walking humanoid robot. The problem is solved using a space carving algorithm, which relies on both the accurate extraction of the occluding boundaries of the object as well as the precise estimation of the camera pose for each video frame. For data acquisition, a monocular visual-based control has been developed that drives the trajectory of the robot around an object placed on a small table. Due to the stepping of the humanoid, the recorded sequence is contaminated with artefacts that affect the correct extraction of contours along the video frames. To overcome this issue, a method that assigns a fitness score for each frame is proposed, delivering a subset of camera poses and video frames that produce consistent 3D shape estimations of the objects used for experimental evaluation.**

I. INTRODUCTION

Most humanoid robots rely on vision systems in order to perceive the environment and resemble human capabilities. In particular, monocular vision is preferred for small-sized humanoids that are certainly constrained to be equipped with lightweight, low-cost and low-energy consumption devices. For these robots, there is a tradeoff between a suitable camera and the quality of the images acquired during the biped march, as the stepping impacts cause jerky camera movements which generate continuos blurring along the related video sequence. Localizing the robot is an additional complex problem due to discrepancies in time among sensor readings, i.e., the orders of magnitude from the acquired frequency signals differ for each sensor and the rate of divergence from the walking reference trajectory is high for small distances.

In the context of 3D object reconstruction, analyzing a monocular video sequence acquired by a humanoid robot represents a difficult task which involves solving for camera localization as well as extracting meaningful image features under challenging motion conditions. This paper investigates the feasibility to estimate, in a multi-view fashion, the 3D geometry of an interest object from the video frames generated along the march of a humanoid. In order to capture multiple views, the robot performs a circular trajectory generated through a locomotion control that corrects the positions and orientations of the robot in accordance with vectors lying on a virtual circle of known radius.

[1]P. A. Martínez, M. Castelán and G. Arechavaleta are with the Robotics and Advanced Manufacturing Group, Centro de Investigación y Estudios Avanzados del Instituto Politécnico Nacional, Ramos Arizpe, Coah, 25900, México. pablo.martinez@cinvestav.edu.mx

[2]David Varas, Margarita Camacho and Ferran Marques are with the Department of Signal Theory and Communications, Universitat Politècnica de Catalunya, Barcelona, 08034, Spain. david.varas@upc.edu

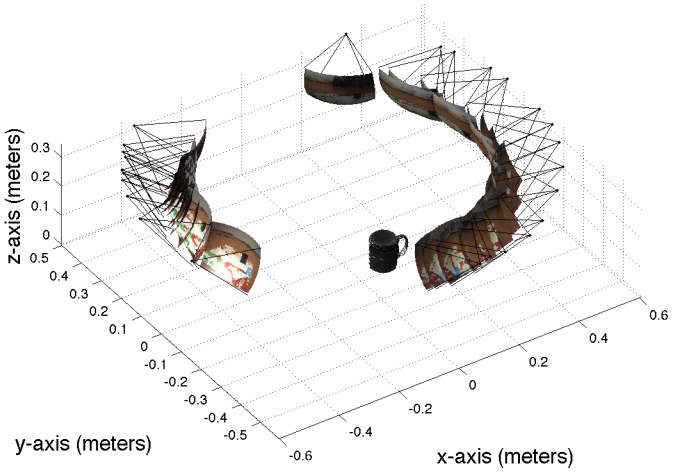


Fig. 1: **The proposed strategy.** A humanoid robot records a video sequence that samples multiple views of the shape of an interest object. A visual-based locomotion control uses the monocular localization of the robot to correct its stepping and perform the required trajectory. An analysis of the recorded sequence is applied in order to determine the suitability of each frame for the purposes of contour extraction. Finally, from the selected frames, a particle filter-based object segmentation process is coupled with a space carving algorithm for estimating the geometry of the object. The figure shows the 25 camera poses of the selected video frames and the estimated 3D shape of the object.

For object segmentation, a strategy has been developed that aims at selecting a suitable set of video frames for robustly reconstructing the 3D shape of the object. In a first stage, blurred images are eliminated from the sequence as well as those frames where parts of the object appear outside the image limits. From this subset, object segmentation is performed using a region-based particle filter approach, from which a consistency score is assigned for each frame. The video frames with the highest scores that also observe a uniform distribution of the sampled object views are finally selected for 3D shape recovery. The process is illustrated in Figure 1, where the final selected video frames are shown as camera poses surrounding an object of interest. In this sense, the main contribution of this paper is a method that is capable of analyzing a video sequence generated by a humanoid robot for the purposes of 3D object reconstruction from multiple views.

The rest of the paper is organized as follows: Section II presents the relevant work that approaches both 3D shape estimation and object segmentation; Section III describes the proposed strategy to record the video sequence of the object; Section IV depicts the method developed to analyse

the recorded video and select the image frames with the most suitable contours for 3D reconstruction; Section V shows the 3D shape recovery results for three different objects and finally Section VI provides concluding remarks and future lines of research.

II. Related work

The classical computer vision problem of monocular 3D object reconstruction has been adapted to robotic platforms in order to provide them with a way to perceive the environment and interact with it, i.e. make a decision, develop a navigation task or grasp an object. For humanoid robots provided with a stereo vision system, solutions have been proposed using next-best-view (NBV) techniques. The problem, commonly approached in a multi-objective optimization manner, consists of estimating the next camera pose that maximizes the unknown volume of the object given a current voxel grid. When coupling NBV with humanoid platforms, the new camera poses are also required to agree with the set of admissible head and body configurations of the robot.

The first attempts in reconstructing the shape of an object by a humanoid robot have probably been described in [1]. Here, an optimal set of camera and body poses were calculated in order to acquire a set of views of the object for achieving a partial reconstruction. For isolating object from background, a red table was used to place the object. The 3D model was estimated from the registration of five disparity maps obtained from stereo images of the object. While this work did not focus on estimating the complete geometry of the object, it showed how a partially reconstructed model would suffice for recognizing the object in cluttered environments. A method for achieving a complete reconstruction that considered obstacle avoidance was later proposed in [2], with successful simulation results. In this approach, a 3D occupancy grid covered the object and an updating operation of the occupancy values of the grid was performed for each captured stereo frame.

More recently in [3] a strategy was presented for acquiring monocular views of an object by a small size humanoid. From these image views, contours of the object were obtained to reconstruct 3D shape using the space carving approach of [4]. The task of extracting contours was simplified using a color based thresholding technique and, as a consequence, the objects to reconstruct were painted in red while the acquisition scenario was covered in blue. For controlling the robot, an extended Kalman filter approach was developed to estimate the position and width of the object with respect to the robot. Camera localization was achieved by solving for extrinsic camera parameters from a set of eight colored landmarks of known distance.

A common aspect to note from the above approaches is that image features and therefore the 3D shape of the object are inferred from still camera poses. In other words, in order to calculate its next state, the robot has to make a pause and interrupt its motion. Rather than being related with mechanical limitations, this pause is a consequence of the optimization process implied in the tasks of calculating the next camera pose or updating information about the object's shape. Although these approaches profit on the versatility of head and body pose configurations of the humanoid robot, exploiting the full range of data contained in the frames of a video sequence has been neglected.

Indeed, when video information is considered, some problems arise. One of the conventional image quality degradation is the blurring of moving camera sequences [5]. In [6], a system that detects blurred images and classify them using the magnitude and the direction of its gradients is proposed. Althought experiments show satisfactory results, an annotated database is needed to train an SVM classifier. Besides fast camera motion, other difficulties such as changes of the object position and shape are problems that should be handled by the system [7].

In contrast with other tracking methods, particle filters [8] can robustly track objects from a sequence of different views as they neither are limited to linear systems nor require the noise involved in the process to be Gaussian. In [9], a particle filter with edge-based features is proposed. This method has been widely used since it provides a robust framework for tracking curves in clutter. However, the space of possible deformations is limited and some transformations of the object shape may not be correctly estimated. This restriction could be critic especially in a multiview scenario. We adapt this idea using shape descriptors without any restriction in the space of possible deformation.

Image-based features for particle filters were introduced by [10]. In it, color histogram is used to robustly track objects in the scene. This feature has the advantages of being scale invariant and robust to partial occlusions and rotations. Moreover, it can be efficiently computed. In our work, we use the Diffusion distance [11] instead of the Bhattacharyya distance [12] for histogram comparison since it leads to better perceptual performance. As the color of an object can vary through time, the target model is adapted during temporally stable image observations in [13]. Note that [10], [13] do not provide shape estimation.

In our work, we propose to use the region-based particle filter presented in [14] to allow tracking and segmenting objects in sequences of different views. This is a suitable algorithm for this task as a geometrical shape is not considered to represent the object. Instead, its contours are propagated between image pairs of consecutive views.

As far as the separation of object from background is concerned, efforts have been developed for coupling 3D object segmentation with successful results in grasping tasks. In [15], a model-free algorithm was proposed to partition the surface normals of depth images acquired with an RGB-D sensor, identifying the connecting regions that belong to several objects in a cluttered scene. Other approaches have also been introduced for stereo images, as in [16], where the principle of fixation by an active observer was used to emulate foveated vision, resulting in an improved selection of the object's contours. While both approaches are capable of performing segmentation from background as new objects are included in the observed scene, the sensors remain fixed

and this assumption does not fit into biped robotic platforms.

It is worth commenting on the growing popularity of RGB-D sensors, which has allowed the recent development of the now called RGB-D SLAM (Simultaneous Localization and Mapping) systems [17] [18]. These methods have proved successful in SLAM tasks which include a dense 3D reconstruction of the observed scene. However, they have been tested over databases that consider smooth transitions between video frames such as hand-held camera and wheeled robot motion [19]. Unfortunately, as the march of a humanoid robot implies constant swinging, the risk of sudden changes in the motion of the camera may compromise the applicability of these approaches.

III. Monocular vision-based locomotion control

Acquiring multiple views of the object of interest is the first step in the reconstruction of its geometry. For each video frame, it is also necessary to register the position and orientation of the camera, which is estimated under a monocular vision-based framework. We have chosen PTAM (the Parallel Tracking and Mapping software of the University of Oxford [20]) to solve this problem as it is able to track hundreds of features, perform both local (incremental) and global bundle adjustments and grow the 3D map when new keyframes appear. These tasks are computed in parallel resulting in real-time applications. For the monocular case, an initialization that simulates a stereo pair to approximate the depth of the initial 3D points is crucial to obtain coherent results.

For generating the robot trajectory we propose a monocular vision-based locomotion control that drives the camera of the robot to face towards the center of the table where the object is located. Additionally, the position of the robot is constrained to keep a constant distance from the center of the table in order to emulate a surrounding trajectory.

A. Robot localization

The output of the camera localization process is a rotation matrix $\mathbf{R}_w$ and a translation column vector $\mathbf{t}_w = [x_w, y_w, z_w]^T$ that relate the world and the camera, in other words, the rigid body transformation from the axis of the world to the axis of the camera. For controlling the march of the robot we are only concerned with the position of the robot on the xy- plane and its orientation angle. Note that the head (camera) of the robot has been locked to be fully aligned with its body, thus, the homogeneous matrix that maps the robot body frame to the world frame can be approximated as

$$\mathbf{T}_b^w \approx \mathbf{T}_c^w = \begin{bmatrix} \mathbf{R}_w & \mathbf{t}_w \\ 0 & 1 \end{bmatrix}, \tag{1}$$

with w, c and b respectively standing for world, camera and body. The position of the robot in world coordinates can be directly taken from the translation vector $\mathbf{t}_w$. The orientation angle of the robot can be found from the first two elements of the translation vector transformed into camera coordinates as

$$\theta_c = \tan^{-1}(y_c/x_c), \tag{2}$$

Fig. 2: **Three objects to reconstruct.** From left to right, example images of recorded sequences Mug, Duck and Action Man.

where $[x_c, y_c, z_c] = -\mathbf{R}_w^T \mathbf{t}_w$.

B. Locomotion control

In order to solve the problem of multi-view 3D reconstruction from object segmentations the robot needs to surround the object of interest. For this task, we propose a locomotion control that directs the next position of the robot to lie along the circumference of a known radius circle while its orientation is directed towards the center of the circle. For an effective translation to occur along the circumference of the circle, an angular displacement s from the current to the following robot's state has to be considered.

Let $\mathbf{x}_{CoM}^{ref} = [x_w, y_w]^T$ be the reference position of the center of mass (CoM) on the xy-plane of the world taken from the translation vector $\mathbf{t}_w$. The orientation angle θ_c^{ref} can be calculated as shown in Eq. 2. The current state of the robot is defined by the pair $(\mathbf{x}_{CoM}^{ref}, \theta_c^{ref})$ and its projected position lying on the radius r circumference may be defined as the vector $\mathbf{x}_p = [x_p, y_p]^T = r(\mathbf{x}_{CoM}^{ref}/||\mathbf{x}_{CoM}^{ref}||)$. The target state of the robot at time $k+1$ is defined as $(\mathbf{x}_t, \theta_t)$ and can be obtained by adding the angular displacement s to the projected vector $\mathbf{x}_p$. The target orientation θ_t is directly estimated from $\mathbf{x}_t$ and its direction is inverted as the robot is facing towards the center of the circle.

The reference linear velocity of the CoM, $\dot{\mathbf{x}}_{CoM}^{ref}$ is computed considering a proportional control based on the distance between the current estimate of the robot's CoM position and the computed target position. Likewise, for the reference angular velocity, $\dot{\theta}_c^{ref}$, the difference between the current and target orientation is used. Therefore, the errors

$$\mathbf{e}_x = \mathbf{x}_{CoM}^{ref} - \mathbf{x}_t \quad \text{and} \quad e_\theta = \theta_c^{ref} - \theta_t$$

are regulated by imposing the exponential convergences

$$\dot{\mathbf{e}}_x = -\lambda_x \mathbf{e}_x \quad \text{and} \quad \dot{e}_\theta = -\lambda_\theta e_\theta,$$

where λ_x and λ_θ are experimentally tuned constant proportional gains. This procedure is performed while the robot does not reach the end of the surrounding trajectory and is formally described in Algorithm 1.

The input of the walking pattern generator (WPG) is given by $\dot{\mathbf{x}}_{CoM}^{ref}$ while the output considers a dynamically stable trajectory of the CoM, the position of the foot in contact and the next footstep placement. The WPG solves quadratic programs with a predefined time horizon as it is proposed in [21]. In this case, the reference orientation θ_c^{ref} is used to express the inequality constraints that define the admissible region to place the next footstep. The computation of the

Data: Localization $\mathbf{x}_{CoM}^{ref}$ at current time k,
orientation θ_c^{ref} at current time k,
radius r,
angular step s.
Result: Reference linear velocity $\dot{\mathbf{x}}_{CoM}^{ref}$ at time $k+1$,
reference angular velocity $\dot{\theta}_c^{ref}$ at time $k+1$.
while $\mathbf{x}_{CoM}^{ref}$ *outside stopping region* **do**
$[x_p, y_p]^T = r(\mathbf{x}_{CoM}^{ref}/||\mathbf{x}_{CoM}^{ref}||)$
$\mathbf{x}_t = [x_t, y_t]^T = r \begin{bmatrix} \cos(\tan^{-1}(y_p/x_p)+s) \\ \sin(\tan^{-1}(y_p/x_p)+s) \end{bmatrix}$
$\theta_t = \tan^{-1} -(y_t/x_t)$
$\dot{\mathbf{x}}_{CoM}^{ref} = -\lambda_x(\mathbf{x}_{CoM}^{ref} - \mathbf{x}_t)$
$\dot{\theta}_c^{ref} = -\lambda_\theta(\theta_c^{ref} - \theta_t)$
Apply a WPG given $(\dot{\mathbf{x}}_{CoM}^{ref}, \dot{\theta}_c^{ref})$
Generate locomotion with inverse kinematics
end

Algorithm 1: The robot performs a surrounding trajectory in accordance with a circle of radius r and an angular displacement s.

TABLE I: Camera position and orientation error

Sequence	Mean position error	Mean orientation error
Mug	7.4 mm ± 4.3 mm	3.05° ± 1.82°
Duck	8.9 mm ± 5.4 mm	3.35° ± 1.95°
Action Man	7.5 mm ± 4.1 mm	2.85° ± 1.90°

joint trajectories of the robot from the WPG outcome is based on the real-time inverse kinematics method suggested in [22].

For experimental evaluation, the video sequences of three different objects, which will be referred to as Mug, Duck and Action Man, were recorded using the proposed locomotion control. The radius of the circular trajectory was set to 0.6 m and the angular step s was set to 3°. For all sequences, the video acquisition rate was of nine frames per second with a resolution of 640×480 pixels. The number of frames recorded for the Mug, Duck and Action Man sequences was 2148, 2370 and 2123, respectively. Example images taken from these sequences are shown in Figure 2. For all experiments, rich textured papers were placed on the surface of the table in order to guarantee an appropriate environment for monocular SLAM.

Results of applying the control for acquiring the Mug database are shown in Figure 3, where the bird-eye view of the performed trajectory is shown in (a) with a close-up view in (b). The red circle appearing in the figure depicts the target positions to be reached by the robot during the video acquisition. The actual trajectory of the robot is depicted in blue. As a consequence of performing SLAM, a sparse 3D map of the world is incrementally built during the acquisition of the video sequence. The final generated map appears in Figure 3 (a) as grey dots inside the circular trajectory and the position of the mug is marked with a circle approximately centered at $(0.1, -0.1)$. Additionally, a subset of four images from the recorded video is shown in Figure 3 (c) in order to provide a visual idea of the observed scenario along the performed trajectory. From the figure, it is noticeable how the footstep generation of the robot successfully achieves the expected circular trajectory, with camera orientations pointing towards the center of the table.

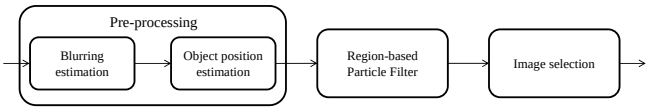


Fig. 4: **Proposed framework.** The aim is to robustly extract a set of object segmentations from different views imposing a minimum quality for its reconstruction. A pre-processing step discards images in which the object may not be correctly segmented. Then, a region-based particle filter obtains object segmentations. Finally, best segmentations are selected to generate the 3D reconstruction.

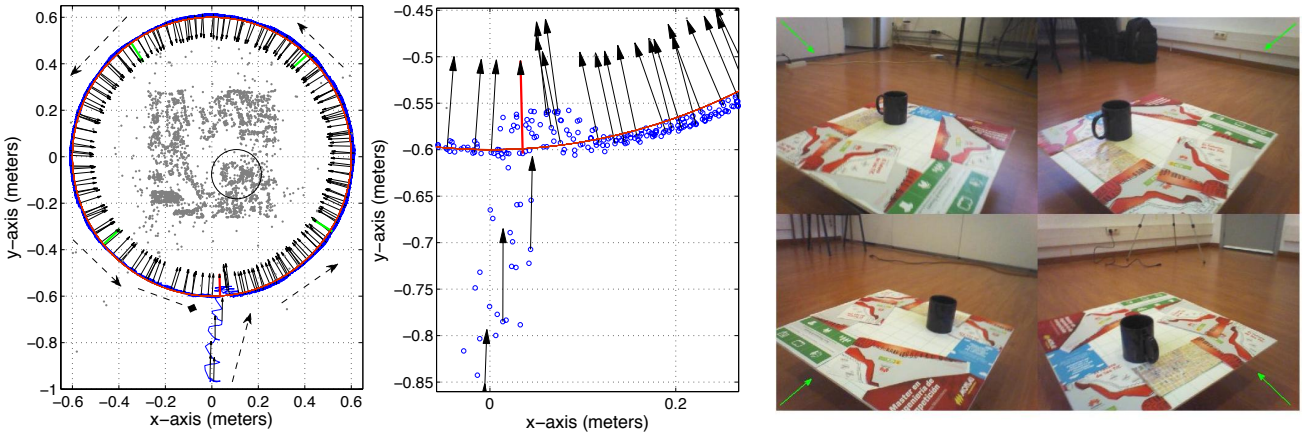
In order to provide a quantitative performance of our control, for each video sequence we measured the distance from every estimated camera position to the nearest position (projection vector of length 0.6 m) on the reference circumference, as the aim of the control is to achieve a trajectory resembling a circle. These distances were considered once the camera poses first entered the circular area, i.e., from the red arrow in Figure 3 (b). The average error in position is shown in Table I, where the proposed scheme delivers a dismissible error of at most 1 cm for the three sequences. The orientation error was also measured considering the angular distance between the current orientation camera angle and the angle of the nearest vector on the reference circumference. The average difference reveals departures of 3° from the expected behavior, confirming the applicability of this strategy for recording video sequences to sample multiple views of the interest object.

IV. Object segmentation from a video sequence

Once the video has been recorded, the object must be segmented in order to create a 3D model. In this work, we adapt the region-based particle filter presented in [14] to extract the 2D shape of the object from partitions (See Figure 5(b)) associated with a set of multiple views of the object. Only images containing the entire object without blur are processed and a subset of these images is finally selected to reconstruct the object in accordance with their final segmentation quality. A diagram of the proposed framework is presented in Figure 4.

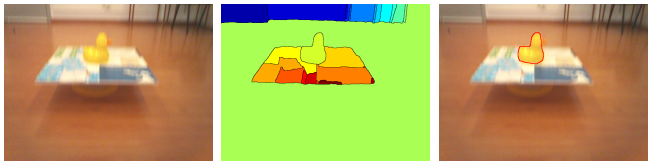
A. Pre-Processing

As the shape of the object is extracted from a partition, the final quality of the model is highly dependent on the image partitions generated along the sequence. These object segmentations at different views are used by a space carving algorithm to reconstruct the 3D model. Although the smoothness of this model increases with the number of segmentations extracted from different views, the larger the number of images considered in this process the higher the probability of an erroneous object shape estimation at least in one view. Thus, a subset of images from the sequence is selected to robustly create a 3D reconstruction of the object.



(a) Bird-eye view of trajectory. (b) Detail of trajectory. (c) Example video frames along the generated trajectory.

Fig. 3: **Applying the proposed control to record a video sequence.** The blue path depicts the estimated robot positions along the circular trajectory. The virtual circle that helps controlling the actual trajectory is shown in red. The counterclok-wise direction of the march is highlighted with dashed arrows. Orientations of the robot separated every 20 frames are shown with arrows facing towards the center of the circle. The projection of the sparse 3D map recovered, as a consequence of SLAM, during the robot's march (mostly features over the table where the object was placed) is shown as grey dots. The location of the mug over the table is marked with a small circle approximately centered at $(0.1, -0.1)$. A close up view of (a) is presented in (b), while example video frames corresponding to green arrows in (a) appear in (c).



(a) Original image (b) Partition (c) Best estimation

Fig. 5: **Discarding blur.** In (a) a blurred image is presented. Images (b) and (c) show its associated partition and the best estimation of the object given this partition, respectively. The blurring effect creates erroneous contours which are not capable to represent a correct segmentation of the object. In this example, the beak of the duck is not included and, as a result, this part of the duck would not be reconstructed.

Two main situations can be found in which a region-based particle filter may not correctly recover the shape of the object. First, when a part of the object is not present in the image. And second, when the blurring effect degradates the quality of the object contours. In order to avoid erroneous estimations, two pre-processing steps select those images in which the object can be correctly segmented. These steps estimate the position of the object in the scene and the blurring of the image respectively.

1) Blurring estimation: Blur is one of the conventional image quality degradations and it can be caused by various factors. In our application, this effect arises due to the rapid camera movement of the robot. The quality of partitions decreases drastically when the blurring effect appears, producing corrupted contours and mixing object and background pixels in the same regions (Figure 5).

Since the image gradient is highly related to image blurring [5], our blur detector computes the magnitude of this gradient to estimate the blur present in an image. Then, a histogram of the gradient is built (in this work, 20 bins have been used). As the contours of a clear image are more precisely defined than the contours of a blurred image, its histogram is expected to contain some contours with large values. On the contrary, contour magnitudes of blurred images should be small.

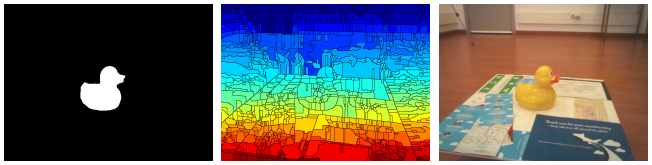
To this extend, the accumulation of the last 10 bins of the histogram is used to classify the image. If this summation represents more than 0.5% of contour pixels, the image is classified as clear. Otherwise, the blurring effect is said to be present.

2) Position estimation: The position of the object in the scene is computed using its relative position with respect to the camera. Due to the camera movement, the object may not be completely observed, and some of its parts can be projected out of the image.

When this situation arises and the image is selected to generate the 3D model, the part which is not included in the scene will not appear in the final reconstruction even if it is correctly segmented in other views. To avoid this problem, a color-based particle filter [13] is used to estimate the position and the bounding box of the object along the sequence. Following a conservative policy, images where the detected bounding box is closer than 25 pixels to an image border are not taken into account to extract the object contours.

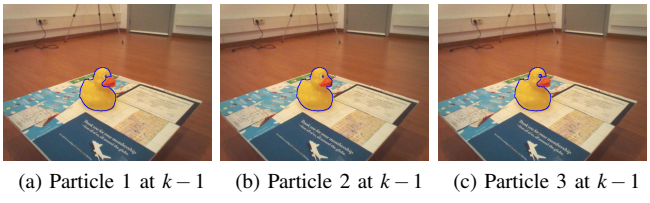
B. Region-Based Particle Filter

This method segments the object along a sequence propagating its shape through time. To this end, similarities between regions are analyzed. Then, parts associated with both the object and background are put in correspondance for each pair of views.

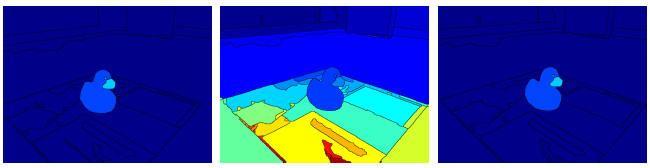


(a) Object mask (b) Partition (c) Original image

Fig. 6: **Region-based particle filter initialization.** In (a) the object mask of the first frame can be observed. Image (b) shows an oversegmentation of the original image, which is presented in (c).



(a) Particle 1 at $k-1$ (b) Particle 2 at $k-1$ (c) Particle 3 at $k-1$



(d) PSP at $k-1$ (e) Partition at k (f) LP at k

Fig. 7: **The region-based particle filter.** In (a), (b) and (c) three particles at $k-1$ are presented. The PSP created by the intersection of these particles is depicted in (d). The new partition at k (e) and the result of the label propagation process (f) are also shown.

In [14], a region-based particle filter is presented in which Monte Carlo methods and a representation of the image in terms of regions are combined. This algorithm, does not only provide the position of the object as in the color-based approach. Instead, it also estimates the shape of the object along the sequence, given an object mask of the first frame. This mask should be provided to the algorithm and in our experiments it is set by hand (See Figure 6). The object mask is used to create a color model of the object in the image that serves as a reference to weight particle estimations. In this work, a histogram has been used as object model. In the region-based approach, the measurement at time k, z_k, is composed by the input image and its partition, whereas the estate, x_k, is formed by a union of regions from the partition associated with the input image. In this work, partitions are obtained using [23]. An example of image partition created with this technique can be observed in Figure 6.

Each particle stands for a state represented by a union of regions that define an estimation of the object. As particles represent different estates of the tracked object, they are also represented by unions of regions. Thus, in order to form the new set of particles, not any propagation is allowed. In other words, the measurement (partition) is used in the propagation process. Then, the weight update equation can be written as:

$$w_k^{(i)} \propto w_{k-1}^{(i)} \sum_c p(z_k|x_k^c) p(x_k^c|x_{k-1}^{(i)}) \tag{3}$$

where $w_k^{(i)}$ is the weight of the *i-th* particle at time k and c swaps all the possible states.

This summation becomes intractable using a brute force approach. For each particle, its probability of being represented by all the possible combinations of regions of the next partition should be computed ($p(x_k^c|x_{k-1}^{(i)})$) and evaluated ($p(z_k|x_k^c)$). To solve this problem, the algorithm takes advantage of the two steps of a usual tracker: *prediction* (movement prediction) and *perturbation* (particle randomness).

a) Prediction: In this step, the shape of the object in the next frame is estimated to ensure a minimum quality of the new set of particles. In order to perform this process in a robust and efficient manner, a *particle support partition* (PSP) is created taking into account the intersections between particles. Using this partition, all particles can be propagated with a single optimization process: *label propagation*. This process labels regions from the new measurement with labels from the particle support partition optimizing similarities between regions which are adjacent as it can be observed in Figure 7. These similarities are computed over the contour elements using color, texture and distance information. As it is expected that the object shape does not change abruptly between consecutive views, only regions in a neighborhood of 50 pixels are considered adjacent. This adapts the concept of adjacency presented in [24] (multiple static partitions) and in [14] (rapid changes) to a multiple view scenario. The result of the process is a *labeled partition* (LP) in which regions from the new partition have been labeled with labels from the PSP.

b) Perturbation: For each particle, N changes are randomly proposed separately. These changes consist on adding/removing regions that belong to the particle or its neighborhood. Then, a greedy algorithm is proposed in which those changes that improve a similarity measure (Diffussion distance) between the particle and the model, are stored and combined to form the final particle. Details of this algorithm can be found in [14].

Finally, the estimation of the object is obtained as the combination of the state of the particles. Note that in the region-based case each particle has its own associated object shape obtained through the two previous steps. Thus, the object shape is estimated combining the masks of all the particles. As a result of this combination, a certain probability of belonging to the object is assigned to each region. The final object shape is estimated considering those regions with a probability higher than a given threshold (In this work, 50% has been used). The capacity of estimating the 2D shape of the object in an image view given its shape in a similar view makes this algorithm suitable for reconstruction applications.

C. Image selection

As it has been previously commented, errors in the object shape estimation rapidly degradate the quality of the final reconstruction. In order to avoid this degradation, only a subset of the views analyzed by the region-based particle filter are used to create the 3D model.

(a) Camera positions of the selected video frames for Duck and Action Man.

(b) Rendered views of the estimated 3D geometry.

(c) Incorrect reconstructions.

Fig. 8: **3D shape estimation results.** The selected video frames for reconstruction are shown, for two of the objects, in (a). Random views of the recovered 3D models are shown in (b). Incorrect reconstructions of the object as a consequence of adding a low quality segmentation appear in (c). For rendering the different views, we applied a voxel coloring method that assigns, for each surface voxel, its corresponding pixel color taken from the camera that is closest to the voxel.

Images are selected according to the Diffusion distance between the segmented object and the model. Moreover, the circular distribution of the cameras is taken into account to correctly represent the entire 3D object. The image with the highest coefficient is chosen first. Then, from the rest of images, the view with the highest coefficient which is not included in a temporal window of 7 frames centered in any chosen image is selected. This process is iterated until 25 frames are chosen or the coefficient falls below a given threshold. The resulting set of views is used to robustly reconstruct the object.

V. 3D RECONSTRUCTION RESULTS

As far as the 3D reconstruction of the object is concerned, we used a method based on shape from occluding boundaries known as space carving [4]. Roughly, this method uses the camera matrix in order to reproject, towards the world, the area bounded by the silhouette of the observed object in the image. The camera matrix is calculated as

$$\mathbf{P} = \mathbf{K}[\mathbf{R}_w \mathbf{t}_w], \tag{4}$$

where $\mathbf{K}$ is the matrix of intrinsic parameters of the camera. From a set of multiple silhouettes with known camera matrices, a 3D model is finally recovered from the intersection of all reprojected silhouettes into the voxel map. This process, which resembles sculpting (carving), is usually posed using a turntable and a fixed camera, which greatly simplifies the tasks of object segmentation and estimation of the camera matrices. On the contrary, our method is capable of dealing with a challenging video sequence recorded from a humanoid robot in motion.

In this section, we show how by coupling a robust object segmentation method with a trajectory that guarantees exhaustive sampling of the image views of the object, it is possible to estimate a visually coherent 3D geometry of the interest object.

The final results of our complete framework are illustrated in Figure 8, where we present the chosen video frames after applying the image selection process described in the previous section. In column (a), we have only included scenarios for Duck and Action Man sequences as the corresponding scenario for Mug appears in Figure 1. It is worth commenting on the differences between the set of selected images from the different scenarios. Particularly, in the Mug experiment a large region of camera poses was left out of the quality set, which can be explained as a consequence of the robot being too far from the object along certain regions of the performed path. In this case, the object was not placed in the center of the table and, as a consequence, it appeared too small for an accurate segmentation to become feasible. Nonetheless, the rendered views of the 3D reconstruction provided in Figure 8 (b) reveal that the shape of the mug was reasonably recovered. Likewise, the rendered views obtained from Duck and Action Man video sequences provide a range of object views that qualitatively agree with an expected reconstructed shape of the object.

Table II shows how the recovered 3D models are also consistent with the objects' physical dimensions in the world, as we measured the width, depth and height of each object

TABLE II: Departures from physical dimensions

Object	Width	Depth	Height
Mug	5.6 mm	6.9 mm	1.0 mm
Duck	5.2 mm	5.0 mm	6.0 mm
Action Man	5.0 mm	4.5 mm	5.2 mm

and each reconstruction in order to obtain a degree of similarity between their corresponding bounding boxes. As shown in the table, the average error was 5.3 mm in width, 5.5 mm in depth and 4.1 mm in height. Considering the accuracy in the recovered geometry, this could be potentially used in further tasks such as grasping and identification.

Finally, rendered views corresponding to incorrect 3D reconstructions are additionally depicted in Figure 8 (c). For this experiment, we included a single video frame observing a low score in the image selection process. Note how the recovered 3D models exhibit important missing parts such as the handle, the beak and the head, respectively for Mug, Duck and Action Man video sequences.

It is worth commenting that our method presents important differences with other 3D reconstruction approaches. For example, while [1] emphasizes the full body posture of the HRP-2 in order to get still images of the object, we favor multiple view exploration by relying on a powerful computer vision segmentation strategy that works over humanoid-locomotion generated video sequences. Also, while they focus on the partial reconstruction of the object for the purposes of later identification, our approach values the full 3D reconstruction of the interest object. For this reason, comparing approaches would not provide enough insight and has not been included in this paper.

VI. Conclusions and future work

We have shown how a state-of-the-art technique in video analysis can be adapted to the challenging video sequences generated by a humanoid robot in motion. Particularly, the complex task of estimating the 3D shape of an interest object has been achieved by relying on a monocular visual-based control of the robot and the robust extraction of silhouettes from a selected subset of highly scored video frames. The reported results are promising and future improvements can be drawn in two directions. As far as the control of the robot is concerned, relaxing the circular supposition might generate camera views of the object that contain important information, as the robot would be able to get closer or farther from the object when needed. Also, incorporating constraints to use a wider range of body postures is desirable for the purposes of generating a richer set of camera views. The capabilities of the video analysis strategy can be as well extended and we are considering incorporating 3D information related to the sparse cloud of points generated during the march of the humanoid robot, which can be useful for a rough object-from-background separation.

Acknowledgment

This work has been developed in the framework of the project BIGGRAPH- TEC2013-43935-R, financed by the Spanish Ministerio de Economía y Competitividad and the European Regional Development Fund (ERDF).

References

[1] O. Stasse, D. Larlus, B. Lagarde, A. Escande, F. Saidi, A. Kheddar, K. Yokoi, and F. Jurie, "Towards autonomous object reconstruction for visual search by the humanoid robot hrp-2," in *IEEE/RAS International Conference on Humanoid Robots*, 2007, pp. 151–158.

[2] T. Foissotte, O. Stasse, P.-B. Wieber, A. Escande, and A. Kheddar, "Autonomous 3d modeling by a humanoid robot using an optimization-driven next-best-view formulation," *Humanoid Robots*, vol. 1, no. 1, pp. 1–23, 2011.

[3] J. Delfin, O. Mar, J. Hayet, M. Castelán, and G. Arechavaleta, "An active strategy for the simultaneous localization and reconstruction of a 3d object from a humanoid platform," in *IEEE/RAS International Conference on Humanoids Robots*, 2012, pp. 384–389.

[4] K. N. Kutulakos and S. M. Seitz, "A theory of shape by space carving," *International Journal of Computer Vision*, vol. 38, pp. 199–218, 2000.

[5] R. Fergus, B. Singh, A. Hertzmann, S. T. Roweis, and W. T. Freeman, "Removing camera shake from a single photograph," *ACM Trans. Graph.*, vol. 25, no. 3, pp. 787–794, Jul. 2006.

[6] in *Advances in Multimedia Modeling*, ser. Lecture Notes in Computer Science, S. Satoh, F. Nack, and M. Etoh, Eds., 2008, vol. 4903.

[7] A. Yilmaz, O. Javed, and M. Shah, "Object tracking: A survey," *ACM Computing Surveys*, vol. 38, p. no 4, 2006.

[8] M. S. Arulampalam, S. Maskell, N. Gordon, and T. Clapp, "A tutorial on particle filters for online nonlinear/non-gaussian bayesian tracking," *IEEE Trans. on Signal Processing*, vol. 50, no. 2, 2002.

[9] M. Isard and A. Blake, "CONDENSATION - conditional density propagation for visual tracking," *International Journal of Computer Vision*, vol. 29, pp. 5–28, 1998.

[10] K. Nummiaro, E. Koller-Meier, and L. V. Gool, "A color-based particle filter," *First International Workshop on Generative-Model-Based Vision*, pp. 53–60, 2002.

[11] H. Ling and K. Okada, "Diffusion distance for histogram comparison," in *CVPR*, 2006.

[12] A. Bhattacharyya, "On a measure of divergence between two statistical populations defined by their probability distributions," *Bulletin of Calcutta Mathematical Society*, 1943.

[13] K. Nummiaro, E. Koller-Meier, and L. Van Gool, "An adaptive color-based particle filter," *Image and Vision Computing*, vol. 21, 1, pp. 99–110, 2003.

[14] D. Varas and F. Marques, "Region-based particle filter for video object segmentation," in *CVPR*, 2014.

[15] A. Ückermann, C. Elbrechter, R. Haschke, and H. Ritter, "3d scene segmentation for autonomous robot grasping," in *IEEE/RSJ International Conference on Intelligent Robots and Systems*, 2012, pp. 1734–1740.

[16] A. Mishra, Y. Aloimonos, and C. Fermuller, "Active segmentation for robots," in *IEEE/RSJ International Conference on Intelligent Robots and Systems*, 2009.

[17] P. Henry, M. Krainin, E. Herbst, X. Ren, and D. Fox, "Rgb-d mapping: Using kinect-style depth cameras for dense 3d modeling of indoor environments," *International Journal of Robotics Research*, vol. 31, no. 5, pp. 647–663, 2012.

[18] F. Endres, J. Hess, J. Sturm, D. Cremers, and W. Burgard, "3-d mapping with an rgb-d camera," *to appear in IEEE Transactions on Robotics*, 2014.

[19] J. Sturm, N. Engelhard, F. Endres, W. Burgard, and D. Cremers, "A benchmark for the evaluation of rgb-d slam systems," in *IEEE International Conference on Intelligent Robots and Systems*, Vilamoura, Portugal, 2012.

[20] G. Klein and D. Murray, "Parallel tracking and mapping for small AR workspaces," in *Proc. Sixth IEEE and ACM International Symposium on Mixed and Augmented Reality (ISMAR'07)*, Nara, Japan, November 2007.

[21] A. Herdt, H. Diedam, P.-B. Wieber, D. Dimitrov, K. Mombaur, and M. Diehl, "Online walking motion generation with automatic footstep placement," *Advanced Robotics*, vol. 24, no. 5-6, pp. 719–737, 2010.

[22] O. Kanoun, "Real-time prioritized kinematic control under inequality constraints for redundant manipulators," in *Robotics: Science and Systems VII*, Los Angeles, CA, USA, June 2011.
[23] P. Arbelaez, M. Maire, C. Fowlkes, and J. Malik, "From contours to regions: An empirical evaluation," in *IEEE Conference on Computer Vision and Pattern Recognition*, June 2009, pp. 2294–2301.
[24] D. Glasner, S. Vitaladevuni, and R. Basri, "Contour-based joint clustering of multiple segmentations," in *CVPR*, 2011.