Object Segmentation through Human-Robot Interactions in the Frequency Domain

ARTUR M. ARSENIO

Massachusetts Institute of Technology - MIT Artificial Intelligence Laboratory
200 Technology Square, Room NE43-936, Cambridge, MA 02139, USA
arsenio@ai.mit.edu

Abstract. This paper presents a new embodied approach for object segmentation by a humanoid robot. It relies on interactions with a human teacher that drives the robot through the process of segmenting objects from arbitrarily complex, non-static images. Objects from a large spectrum of different scenarios were successfully segmented by the proposed algorithms.

1 Introduction

Embodied vision [2] extends far behind the concept of active vision - the human/robot body is used not only to facilitate perception, but also to change world context so that it is easily understood by the robotic creature (Cog, the humanoid robot used throughout this work, is shown in Figure 1).

Figure 1: Cog, the humanoid robot used throughout this work. It has two eyes, each eye equipped with one wide-field of view camera and one foveal, narrow-field of view camera.

Embodied vision methods will be demonstrated for simplifying visual processing by being attentive to the human *Hand*, *Arm* or *Finger* - which will be defined as the human actuator (although a robot actuator is also used). This is similar to cues used by primates, who possess specific brain areas to process the hand visual appearance [8]. Focus will be placed on a fundamental problem in computer vision - *Object Segmentation* - which will be dealt with by detecting and interpreting natural human showing behavior such as finger tapping, arm waving, object shaking or poking. Object segmentation is truly a key ability worth investing effort so that other capabilities, such as object/function recognition can be developed on top of it. The framework here described is currently being developed to estimate object kinematics/dynamics, with the goals of task identification, so that a human teacher can guide the robot in the process of learning and executing new activities.

1.1 Embodied object segmentation

Several algorithms have been presented in the literature for image color segmentation or texture segmentation. However, objects composed of just one color or texture do not occur often, being only a particular instance of a much more diversified set. Among previous object segmentation techniques it should be stressed the min-cut algorithm [7]. Although a very good tool, it suffers from several problems which affect non-embodied techniques.

Indeed, object segmentation on unstructured, non-static, noisy and low resolution images is a hard problem. The techniques this paper describes for object segmentation deal with problems such as (see Figure 2):

- Segmentation of an object with similar colors or textures as the background
- Segmentation of an object among multiple moving objects in a scene
- Segmentation of fixed or heavy objects in a scene, such as a table or a sofa
- Segmentation of objects painted or drawn in a book or in a frame, which cannot be moved relatively to other objects drawn on the same page
- Robustness to luminosity variations
- Fast Segmentation
- Low resolution images (for real-time motion, images of size 128×128 are used)

Figure 2: Simulation of a scene using the *3D DataBecker Home Designer package*. In scenes such as the one shown, the drawer and the table cannot be moved to grab the child attentional focus on it - a strategy that if available could enable object segmentation. On the other hand, multiple moving objects in a scene makes harder to introduce an object to a child, since the other objects are distracters. Furthermore, both the wooden table and the wooden chair have common colors and textures, which turns harder the segmentation problem. In addition, a lamp may be on or off changing luminosity in the room. Children rely heavily on a human instructor for learning new objects and tasks. They also depend heavily on poking and waving objects by themselves to learn about these objects.

Two other embodied object segmentation techniques developed recently [1] include 1) active segmentation from poking objects with a robot actuator, and 2) interactive object segmentation on a wearable computer system. These two segmentation scenarios operate on £rst-person perspectives of the world: the robot watching its own motion, or a wearable watching its wearer's motion. The techniques presented in this paper requires rather either a human teacher or a robot. This strategy is suitable for segmenting objects based on external cues. It exploits shared world perspectives between a cooperative human and a robot through an embedded protocol. Objects are presented to the robot according to a protocol based on periodic motion, e.g. waving an object or tapping it with one's £nger [2].

I propose human engineering the robot's physical environment on-line. Through social interactions of a robot with a caregiver, the latter facilitates robot's perception and learning, in the same way as human caregivers facilitate children perception and learning during child development phases.

This paper is organized as follows. Section 2 describes how objects with a strong and narrow frequency content of the motion £eld are £ltered out from the images. The next three sections describe different protocols a human caregiver may use to boost the robot's object segmentation capabilities. Segmentation by passive demonstration is described in Section 3. This technique is especially well suited for segmentation of £xed or heavy objects in a scene, such as a table or a drawer, or objects drawn or painted in books. Section 4 presents object segmentation though active object actuation. Objects are waved by a human teacher in front of the robot. Objects that are dif£cult to waved but easily poked are segmented as described in Section 5. The overall algorithmic control structure is shown in Figure 3. Section 6 discusses the main algorithmic issues, such as robustness to light variations or effects of shadows, which are supported by experimental object segmentation results. Section 7 draws the conclusions and describes future lines of research.

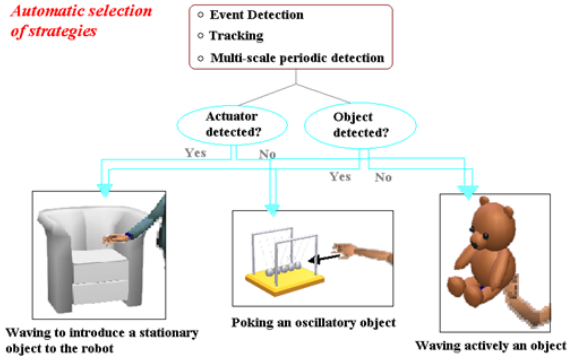

Figure 3: Image objects are segmented differently according to the context. The selection of the appropriate method is done automatically. After detecting an event and determining the trajectory of periodic points, the algorithm determines whether objects or actuators are present, and switches to the appropriate segmentation method.

2 Detection of Events in the Frequency Domain

For events created by human teachers/caregivers, such as tapping an object or waving their hand in front of the robot, the periodic motion can be used to help segment it.

The segmentation processes are activated only after the detection of an event. Image points where the event occurred are initialized and tracked thereafter over a time interval. Their trajectory is evaluated using a Windowed-Fast Fourier transform (WFFT), and tested for a strong periodicity.

2.1 Event Detection

Events are detected through two measurements: a skin-tone mask derived by a simple skin color detector [4]; and a motion mask derived by comparing successive images from the camera and placing a non-convex polygon around any motion found. A real-time, low resolution motion algorithm was developed. It consists of: i) image difference of two

consecutive frames, with a threshold to create a binary image of moving points ii) gaussian £ltering iii) Covering of the resulting image by n overlapping windows. Each image region is covered by 4 of such windows. iv) A convex polygon is used to approximate all the moving points in each window, while windows with less than a minimum number of pixels are eliminated (this removes outliers). The union of all such convex polygons is the desired non-convex polygon.

An event has occurred if both the image's moving region and the skin-tone region are above given thresholds, as illustrated in Figure 4.

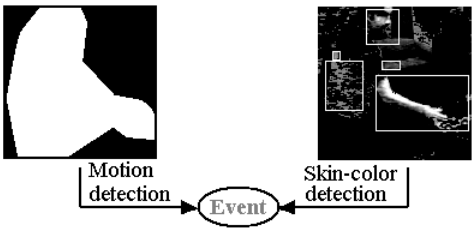

Figure 4: An event detection is triggered by a minimum number of moving points and a minimum number of actuator points in the image at a given time instant.

2.2 Tracking

A grid of points homogeneously sampled from the image are initialized in the moving region, and thereafter tracked over a time interval of approximately 2 seconds (65 frames). At each frame, the velocity V of each point is computed together with the point location in the next frame.

The motion trajectory for each point over this time interval was determined using four different methods. Two were based on the computation of the image optical ¤ow £eld - the apparent motion of image brightness - and consisted of 1) the Horn and Schunk algorithm [6]; and 2) Proesmans's algorithm - essentially a multiscale, anisotropic diffusion variant of Horn and Schunk's algorithm. The other two algorithms rely on discrete point tracking: 1) block matching; and 2) the Lucas-Kanade pyramidal algorithm. We achieved the best results by applying the Lucas-Kanade pyramidal algorithm.

2.3 Multi-scale Periodic Detection

A WFFT is applied to each point motion sequence,

$$I(t, f_t) = \sum_{t=0}^{N-1} i(t')h(t'-t)e^{-j2\pi f_t t'} \qquad (1)$$

where h is usually a Hamming window, and N the number of frames. In this work a rectangular window was used. Although it spreads more the width of the peaks of energy, it does not degrades overall performance while speeding up computation.

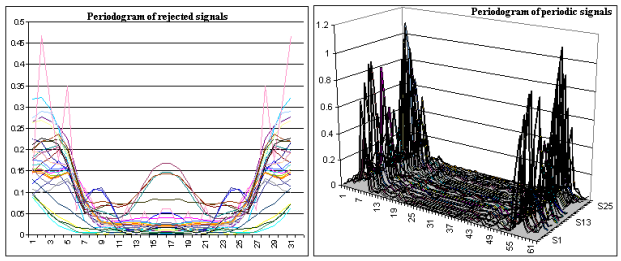

Figure 5: The left graph shows the WFFTs of discarded points, while the right image displays the WFFTs for a set of points contained on an object moving periodically.

Periodicity is estimated as follows. A periodogram is determined for all signals from the energy of the WFFTs over the spectrum of frequencies. These periodograms are then processed to determine whether they are usable for segmentation. A periodogram is rejected if 1) there is more than one energy peak above 50% of the maximum peak; or 2) there are more than three energy peaks above 10% of the maximum peak value; or 3) the DC component corresponds to the maximum energy; or 4) signals have peaks of energy spread in frequency over a threshold. Figure 5 shows either the WFFTs of signals £ltered out and signals passed.

The periodic detection is applied at multiple scales. Indeed, for objects oscillating during a short period of time, the movement might not appear periodic at a coarser scale, but appear as such at a £ner scale. If a strong periodicity is found, the points implicated are used as seeds for color segmentation. Otherwise the window size is halved and the procedure is tried again for each half.

Now that periodic motion can be detected and isolated spatially, the waving behavior will guide the segmentation process.

3 Segmentation by Passive Demonstration

This strategy has the potential to segment objects that cannot be moved independently, such as objects painted in a book (see Figure 6), or heavy, stationary objects such as a table or a sofa. Events of this nature are detected when the majority of the periodic signals arise from points whose color is consistent with skin-tone. The algorithm assumes that skin-tone points moving periodically are probably projected points from the arm, hand and/or £ngers, of a human teacher describing an object to the robot.

An af£ne ¤ow-model (equation 2) is estimated from the optical ¤ow data, and used to determine the trajectory of the arm/hand/£nger position over the temporal sequence, using a least squares minimization criterium for the estimation of A and B, the model parameter matrices.

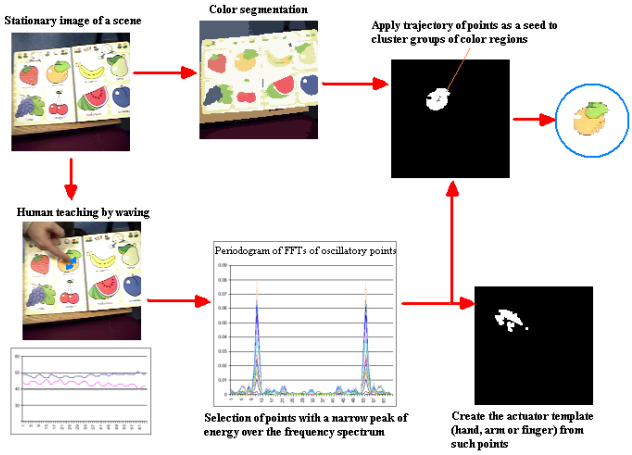

Figure 6: The algorithm works as follows: a human teacher waves on top of the object to be segmented. The motion of the actuator is tracked and the energy per frequency content is determined. A template of the actuator is built from the set of periodic moving points, while the trajectory is used to segment the object from the color segmentation image.

$$\begin{bmatrix} \dot{x} \\ \dot{y} \end{bmatrix} = A \begin{bmatrix} x \\ y \end{bmatrix} + B = \begin{bmatrix} A & B \end{bmatrix} \begin{bmatrix} x \\ y \\ 1 \end{bmatrix} \quad (2)$$

The model uncertainty is given by equation 3, with V being the velocity of an image moving point with image coordinates (x, y), and $[\ \dot{x} \quad \dot{y}\ ]$ the velocity of the same point predicted by the model.

$$unc_V = E\left(\left(V - \begin{bmatrix} \dot{x} \\ \dot{y} \end{bmatrix}\right)(V^T - [\ \dot{x} \quad \dot{y}\ ])\right) \quad (3)$$

Points from these trajectories are collected together, and mapped onto a reference image taken before the waving began (this image is continuously updated until motion is detected). A standard color segmentation [5] algorithm is applied to this reference image. The differentiated clusters of colors hence obtained need to be grouped together into the colors that form an object. This grouping works as follows. Trajectory points are used as seed pixels. The algorithm £lls the regions of the color segmented image whose pixel value are closer to the seed pixel values, using a 8-connectivity strategy.

Therefore, points taken from waving are used to both select and group a set of segmented regions into what will probably constitute the full object (see Figure 7). The clusters grouped by a single trajectory might either form or not form the full object (depending on intersecting or not all the clusters that form the object). But after two or more trajectories this problem vanishes.

Figure 7: Paintings of objects segmented from a book.

4 Segmentation though active actuation

A scene perceived by the robot might contain several moving objects, which may have similar colors or textures as the background. During human-robot interactions, the human often desires to present a speci£c object to the robot for the latter to segment from the image. Multiple moving objects invalidate unambiguous segmentation from motion, while dif£cult £gure/ground separation makes segmentation harder. The strategy described in this section £lters out undesirable moving objects while providing the full object segmentation from motion. Whenever a teacher waves an object in front of the robot, the periodic motion of the object is used to segment it (see Figure 8). This technique is triggered by the following condition: the majority of periodic points are generic in appearance, rather than drawn from the hand or £nger.

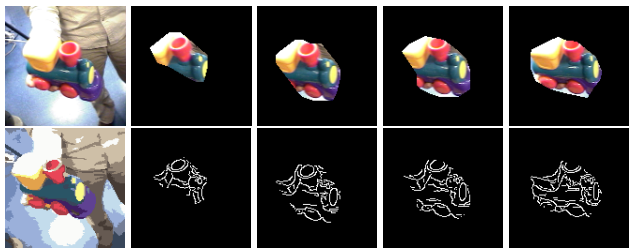

Figure 8: Top row: A image of the scene is shown on the left image, while object templates for the 1^{st} image of the images sequence are shown in the right. Bottom row: Color segmentation of scene on the left, and edge masks of templates obtained by a Canny edge detector.

The set of periodic points tracked over time are sparse, and hence an algorithm is required to group then into a meaningful template of the object of interest. This is done in two steps. First, an af£ne ¤ow-model (equation 2) is applied to the optical ¤ow data at each frame, and used to recruit other points that satisfy equation 2 within the uncertainty given by equation 3. Then, the non-convex polygon approximation algorithm described in Section 2.1 is applied

to all periodic, non skin-colored points, to form the object; it is also applied to all periodic and skin-colored points to form a template of the actuator.

It is worthy to stress that this approach is robust to humans or other objects moving in the background (they are ignored as long as their motion is non-periodic).

5 Segmentation by Poking

The periodic motion induced on an object whenever a robot (or a human instructor) pokes it can be used to segment it (see Figure 9). This technique is basically the same as for the previous case, but triggered by a different condition: while the majority of periodic points are still generic in appearance, the ones drawn from the robot/human actuator do not oscillate, and hence no actuator is detected.

Figure 9: The left image shows a pendular object immediatelly after being poked by the robot. The other images show object segmentations for three different runs.

Similarly to last section strategy, this method is not affected by other scene objects and/or people if their movement is not oscillatory.

6 Experimental Results

This section presents additional experimental results for object segmentation, together with an analysis of such data.

Figure 10 presents the results of segmenting objects from a book. For this particular experiment, the resulting templates might have future use for learning shapes and colors. Figure 11 shows a sample of a number of object segmentations. A large variety of objects with a differentiated geometry/structure were successfully segmented. Segmentations were obtained from a large spectrum of object motions/frequencies.

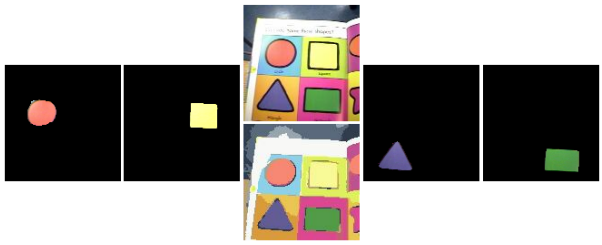

Figure 10: Different shapes/colors segmented from a book by the technique described in section 3.

Some of the objects segmented by the technique described in section 4 may be surrounded by a thin boundary membrane which does not belong to the object. This is due to the size of the search window used for tracking, and occurs whenever there is a lacking of texture on points inside this boundary membrane. These segmentation results are currently being improved by removing this membrane. A Canny edge detector is applied to the object image template (as shown in Figures 8 and 12). Boundary points with signi£cant texture are correctly tracked and hence do not originate false classi£cations (and therefore no undesirable membrane is created). However, points belonging to poor texture surfaces over an object boundary are tracked as moving with the object, being classi£ed as such. However, surfaces lacking texture have low image gradients, and hence are not selected by the Canny edge detector. Initial results of applying a deformable contour to the edge image were very promising on the elimination of such undesirable membrane boundaries.

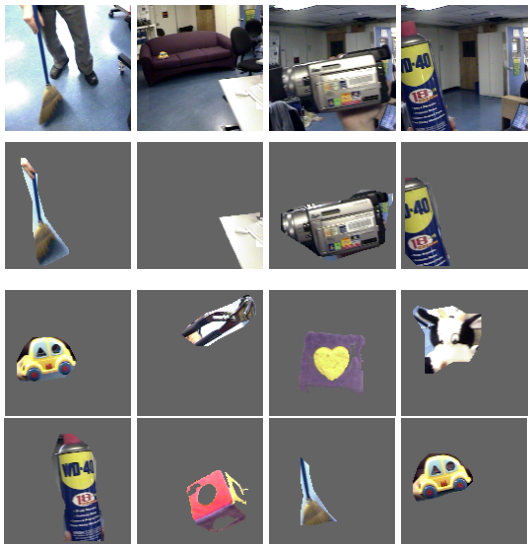

Figure 11: A sample of object segmentations.

6.1 Robustness

A large number of templates is obtained per time interval. Indeed, for objects tracked over n frames, n templates from the object segmentation are available. As shown in Figure 12, reliable segmentations of an hammer are obtained even though some textures and colors of the hammer, the table and the human performing the hammering are similar and thus dif£cult to differentiate.

The algorithms here described are robust to light variations, as can be seen in Figure 13, for results of segmenting objects from a book subject to normal, low and high levels of luminosity.

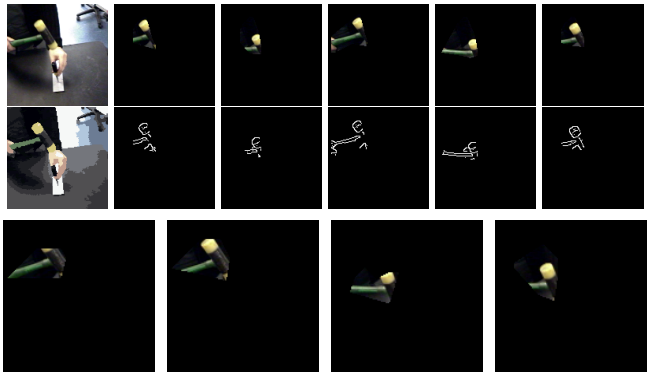

Figure 12: Results for the segmentation of an hammer. Each sequence of images of size n originate n templates for an object. The top row shows 5 templates for £ve frames along the trajectory. The middle row shows the edge images using a Canny edge detector. The bottom row shows four hammer templates obtained from four different image sequences.

7 Conclusions and Future Work

In this paper we introduced the human in the learning loop to facilitate robot perception. By exploiting movements with a strong periodic content, a robot is able to rapidly segment a wide variety of objects from images, with varying conditions of luminosity and a different number of moving artifacts in the scene. The detection is carried out at different time scales for a better compromise between frequency and spatial resolution.

Objects were segmented in several cases from scenes where tasks were being performed in real time, such as hammering. One of the strategies presented also permits the segmentation of objects that are not possible to move and hence to separate from the backgrounds. Such technique is especially powerful to segment £xed or heavy objects in a scene or to teach a robot from books.

It should be emphasized that the techniques presented can be used in a passive vision system (no robot is required), with a human instructor guiding the segmentation process. But a robot may also guide the segmentation process by himself, such as by poking. In addition, learning by scaffolding may result from human/robot social interactions [4].

This work is currently being extended to account with other time events besides periodic events. In addition, the object motion over time is being used to estimate object kinematics and dynamics, with the goals of task identi£cation and of classifying objects by their function. This will enable the robot to learn new tasks from a human teacher and execute them by himself.

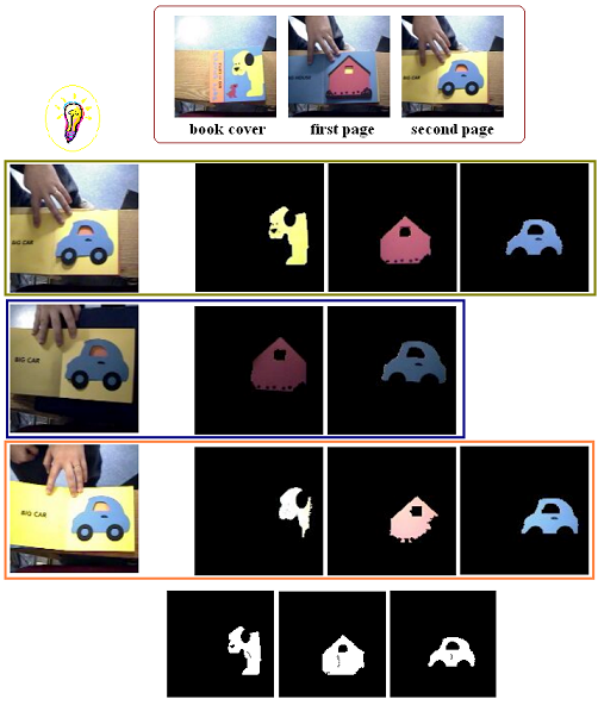

Figure 13: The top images show three different pages of a book. The bottom images show the masks obtained with the actuator trajectory on it. The other rows of images show segmentations for different luminosity conditions.

Acknowledgements

Work funded by DARPA project "Natural Tasking of Robots Based on Human Interaction Clues", contract DABT 63-00-C-10102. Author supported by Portuguese grant PRAXIS XXI BD/15851/98.

References

[1] A. Arsenio, P. Fitzpatrick, C. Kemp and G. Metta, *The Whole World in Your Hand: Active and Interactive Segmentation*. Accepted for publication at the 2^{nd} conference on Epigenetic Robotics, Boston (2003).

[2] A. Arsenio, *Boosting vision through embodiment and situatedness*. MIT AILab Research Abstracts, (2002).

[3] D. Ballard, *Animate vision*. In Arti£cial Intelligence, 48(1), 57. (1986).

[4] C. Breazeal, *Sociable Machines: Expressive Social Exchange Between Humans and Robots*. Doctoral Dissertation, EECS-MIT, (2000).

[5] D. Comaniciu and P. Meer, *Robust analysis of feature spaces: Color image segmentation*. In IEEE Conference on Computer Vision and Pattern Recognition, San Juan, Puerto Rico (1997).

[6] B. Horn, *Robot Vision*. MIT Press, (1986).

[7] J. Shi and J. Malik, *Normalized cuts and image segmentation*. In IEEE. Transactions on Pattern Analysis and Machine Intelligence, 22:888-905, (2000).

[8] D. Perrett, A. Mistlin, M. Harries and A. Chitty, *Understanding the visual appearance and consequence of hand action*. In Vision and action: the control of grasping, pages 163-180. Ablex, NJ. (1990)