An Embodied Approach to Perceptual Grouping

The author
The Institute
The laboratory The address
the email

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 opportunities crated by active vision systems [1, 4]. The human (and/or robot) body is used not only to facilitate perception, but also to change the 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 through different segmentation scenarios).

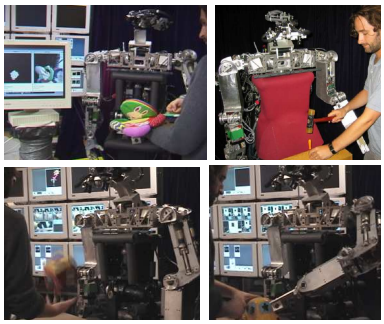

Figure 1. Cog, the humanoid robot used throughout this work, shown through several learning scenarios. These images correspond to real experiments from which objects were separated from the background.

Embodied vision methods will be demonstrated with the goal of simplifying visual processing. This is achieved by selectively attending to the human actuator (*Hand*, *Arm* or *Finger*), or the robot actuator. Indeed, primates have specific brain areas to process the hand visual appearance [11].

Focus will be placed on a fundamental problem in computer vision - *Object Segmentation* - which will be dealt with by detecting and interpreting natural human/robot task behavior such as waving, shaking, poking, grabbing/dropping or throwing objects. Object segmentation is truly a key ability worth investing effort so that other capabilities, such as object/function recognition can be developed.

1.1. Embodied object segmentation

The number of visual segmentation techniques is vast. An active segmentation technique developed recently [7] relies on poking objects with a robot actuator. This strategy operates on first-person perspectives of the world: the robot watching its own motion. However, it is not suitable for segmenting objects based on external cues. Among previous object segmentation techniques it should be stressed the minimum-cut algorithm [12]. Although a good tool, it suffers from several problems which affect non-embodied techniques. Indeed, object segmentation on unstructured, non-static, noisy, low resolution and real-time images is a hard problem (see Figure 2):

- ▷ object may have similar color/texture as background
- ▷ multiple objects might be moving simultaneously in a scene
- ▷ necessary algorithm robustness to luminosity variations
- ▷ requirement of real-time, fast segmentations on low resolution images (128×128 images)

Segmentations must also present robustness to variations in world structure. In addition, mobility constraints (such as segmenting heavy objects) poses additional difficulties, since motion cannot be used to facilitate the problem.

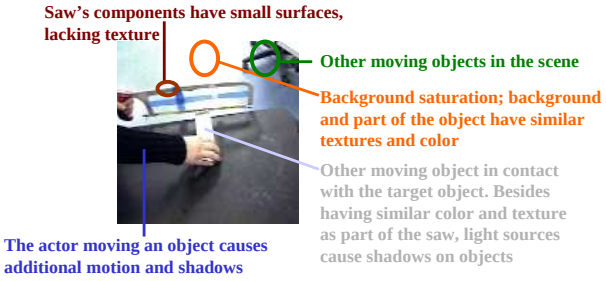

Figure 2. The results presented in this paper were obtained with varying light conditions. The environment was not manipulated to improve natural occurring elements, such as shadows or light saturation from a light source. All the experiments were taken while a human or the robot were performing an activity.

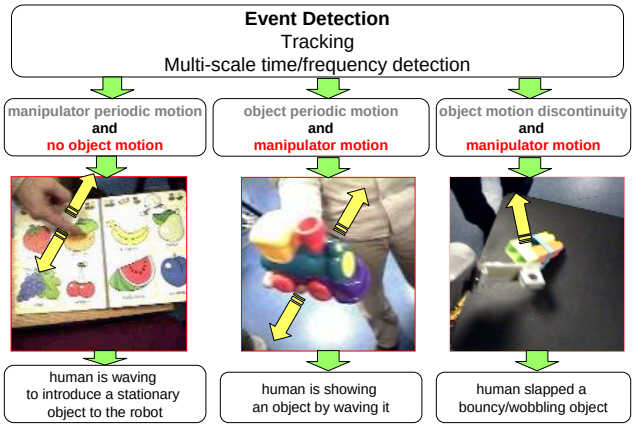

Figure 3. Image objects are segmented differently according to scene context. The selection of the appropriate method is done automatically. After detecting an event and determining the trajectory of periodic points, the algorithm determines whether objects or actuators are present, and switches to the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

Distinguishing an object from its surroundings – the figure/ground segregation problem – will be dealt by exploiting shared world perspectives between a cooperative human and a robot. Indeed, we argue for a visual embodied strategy for object segmentation, which is not limited to active robotic heads. Instead, embodiment of an agent is exploited by probing the world with a human/robot arm. This strategy proves not only useful to segment movable objects, but also to segment object descriptions from books, as well as large, stationary objects (such as a table) from monocular images.

This paper is organized as follows. The next three sections describe different protocols a human instructor might use to boost the robot's object segmentation capabilities (The overall algorithmic control structure is shown in Figure 3). Segmentation by demonstration is described in Section 2. This technique is especially well suited for segmentation of fixed or heavy objects in a scene, such as a table or a drawer, or objects drawn or printed in books. Section 3 presents object segmentation though active object actuation. Objects are waved or shaken by a human actor in front of the robot. Objects that are difficult to wave but easily acted on (for instance, by poking them) are segmented as described in Section 4. Experimental object segmentation results are presented in each section. Finally, Section 5 draws the conclusions.

2. Segmentation by Demonstration

We propose a human aided object segmentation algorithm to tackle the figure-ground segregation problem. Indeed, a significant amount of contextual information may be extracted from a periodically moving actuator. This can be framed as the problem of estimating $p(o_n|v_{B_{\vec{p},\epsilon}}, act^{per}_{\vec{p},S})$, the probability of finding object o_n given a set of local, stationary features v on a neighborhood ball B of radius ϵ centered on location p, and a periodic actuator on such neighborhood with trajectory points in the set $S \subseteq B$.

The following algorithm implements the estimation process to solve this figure-ground separation problem (see Figure 4):

1. A standard color segmentation [6] algorithm is applied to a stationary image (stationary over a sequence of consecutive frames)
2. A human actor waves an arm on top of the object to be segmented
3. The motion of skin-tone pixels is tracked over a time interval (using the Lucas-Kanade Pyramidal algorithm), and the energy per frequency content is determined for each point's trajectory
4. Periodic, skin-tone points are grouped together into the arm mask [3].
5. The trajectory of the arm's endpoint describes an algebraic variety [8] over N^2 (N represents the set of natural numbers). The target object's template is given by the union of all bounded subsets (the color regions of the stationary image) which intersect this variety

An affine flow-model is estimated (using a least squares minimization criterium) from the optical flow data, and used

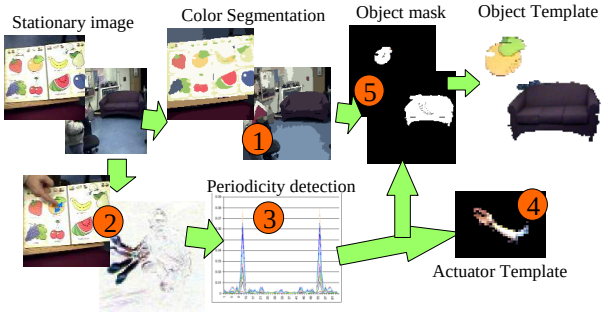

Figure 4. A standard color segmentation algorithm computes a compact cover for the image. The actuator's periodic trajectory is used to extract the object's compact cover – a collection of color cluster sets.

to determine the trajectory of the arm/hand/finger position over the temporal sequence. Periodic detection is then applied at multiple scales. Indeed, for an arm oscillating during a short period of time, the movement might not appear periodic at a coarser scale, but appear as such at a finer scale. If a strong periodicity is not found at a larger scale, the window size is halved and the procedure is repeated again for each half.

The algorithm consists of grouping together the colors that form an object. This grouping works by having periodic trajectory points being used as seed pixels. The algorithm fills the regions of the color segmented image whose pixel values 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 the full object. Clusters grouped by a single trajectory might either form or not form the smallest compact cover which contains the object (depending on intersecting or not all the clusters that form the object). After two or more trajectories this problem vanishes.

2.1. Perceptual Organization

According to Gestalt psychologists, the whole is different than the sum of its parts – the whole is more structured than just a group of separate particles. The technique we suggest segregates objects from the background without processing local features such as *textons* or contours [10]. The proposed grouping paradigm differs from Gestalt grouping rules for perceptual organization. These rules specify how parts are grouped for forming wholes, and some of them are indeed exploited by our grouping method: *Similarity* and *Proximity* rules are embedded on the color segmentation algorithm; moving periodic points in an image sequence are also grouped together. It is worthy to stress that this technique solves very easily the *figure and ground* illusion (usually experienced when gazing at the illustration of a white vase on a black background – the white vase is segregated just by having an human actor tapping on it (or on the black faces, if the human selects them).

2.2. Experimental Results

The strategies which enable the robot to learn from books rely heavily in human-robot interactions. It is essential to have a human in the loop to introduce objects from a book to the robot (as a human caregiver does to a child), by tapping on their book's representations. The segmentation by demonstration method previously presented is then used to segment an object's image from book pages (Figure 5 shows a segmentation sample). This scheme was successfully applied to extract templates for fruits, geometric shapes and other elements from books, under varying light conditions (see Figure 6).

Embodiment of an agent is also exploited by probing the world with a human arm. This strategy proves not only useful to segment object descriptions from books, but also to segment large, stationary objects (such as a table) from monocular images. Figure 7 shows segmentations for a random sample of object segmentations (furniture items), together with statistical results for such objects.

3. Segmentation Driven by Active Actuation

This technique is triggered by the following condition: the majority of periodic points are generic in appearance, rather than drawn from the hand or finger. A visual scene might contain several moving objects, which may have similar colors or textures as the background. Multiple moving objects create ambiguous segmentations from motion, while large similarities between figure and background makes the figure/ground segregation problem harder. However, a human actor can facilitate robot's perception by waving or shaking an object in front of the robot, so that the motion of the object is used to segment it, as follows: Moving image points are initialized and tracked thereafter over a time interval; Their trajectory is then evaluated using a Short Time Fourier transform (STFT), and tested for a strong periodicity. Periodic, non-skin points are then grouped into a unified object.

3.1. 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).

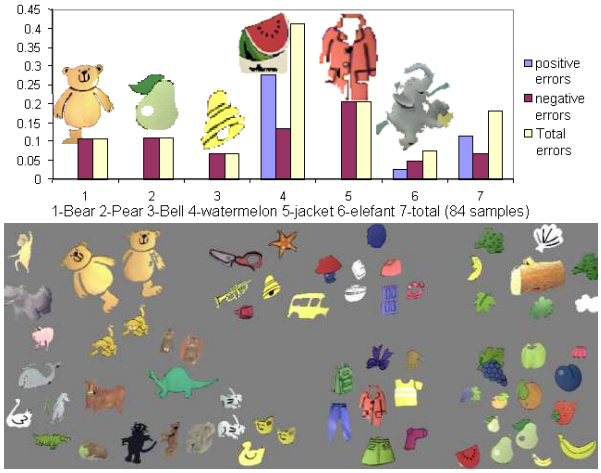

Figure 5. (top) Statistical analysis for object segmentation from books. Errors are given by (template area - object's real visual appearance area)/(real area). Positive errors stand solely for templates with larger area than the real area, while negative errors stand for the inverse. Total errors stand for both errors. The real area values were determined manually. Results shown in graph (7) correspond to data averaged from all these objects. (bottom) Templates for several categories of objects (for which a representative sample is shown), were extracted from dozens of books. Two subjects not acquainted with the algorithm were also briefly instructed on the protocol for interacting with the robot. No noticeable performance degradation was found from such interactions.

At each frame, each point's velocity is computed together with the point's 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 flow field - the apparent motion of image brightness - and consisted of 1) the Horn and Schunk algorithm [9]; 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.

3.2. Multi-scale Periodic Detection

A STFT is applied to each point's motion sequence,

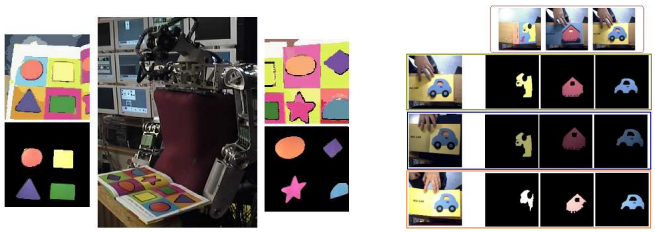

Figure 6. (left) segmentations of geometric shapes from a book (right) top images show pages of a book. The other rows show segmentations for different luminosity conditions.

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

where h is usually a Hamming window, and N the number of frames. In this work a rectangular window was used. Although it spreads the width of the peaks of energy in a larger extent than the Hamming window, it does not degrade overall performance, and decreases computational times.

Periodicity is estimated from a periodogram determined for all signals from the energy of the STFTs over the spectrum of frequencies. These periodograms are processed by a collection of narrow bandwidth band-pass filters. Periodicity is found if, compared to the maximum filter output, all remaining outputs are negligible. The periodic detection is applied at multiple time scales. If a strong periodicity is found, the points implicated are used as seeds for segmentation.

3.3. Perceptual Grouping

Now that periodic motion can be spatially detected and isolated, the waving behavior guides the segmentation process:

1. The set of moving, non-skin [5] points tracked over a time window is sparse. Hence, an affine flow-model is applied to the periodic flow data to recruit other points within uncertainty bounds

2. Clusters of points moving coherently are then covered by a non-convex polygon – approximated by the union of a collection of overlapping, locally convex polygons [3].

This algorithm is much faster than the minimum cut algorithm [12], and provides segmentations of similar quality to the active minimum cut approach presented by [7]. Figure 8 presents statistical results. A random number of segmentation samples are also shown, while Figure 9 shows

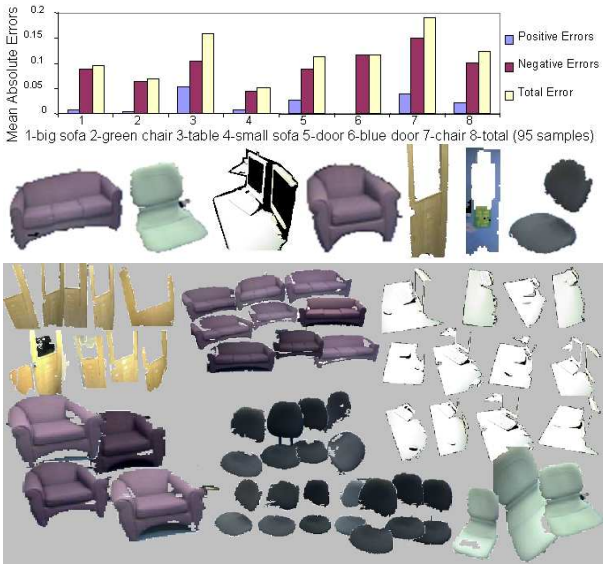

Figure 7. Statistics for the furniture items (a set of segmentation samples is also shown). Results shown in graph (8) correspond to data averaged from all these objects. A black contour was added to the table template for visualization proposes. A chair is grouped from two disconnect regions by merging temporally and spatially close segmentations.

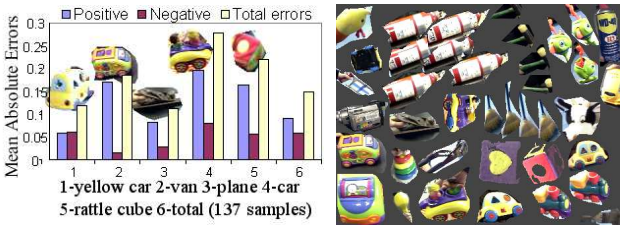

Figure 8. Object segmentations (left) Statistical results for the objects shown. Results shown in graph (6) correspond to data averaged from a larger set of objects on the database (right) sample segmentations from a large corpora consisting of tens of thousands of computed segmentations.

Figure 9. Sample of objects segmented from oscillatory motions, under different views. Several segmentations of the robots arm (and upper arm) are also shown, obtained from rhythmic motions of robot's arm (and upper arm, respectively) in front of a mirror.

results for a few objects under diverse perspective deformations. This approach is robust to other scene objects and/or people moving in the background (they are ignored as long as their motion is non-periodic).

4 Segmentation Through Discontinuous Motion

The discontinuous motion induced on an object whenever a robot (or a human instructor) acts on it can be used for segmenting the object (as shown in Figure 10). In order to detect discontinuous events, an algorithm was developed to identify and track multiple objects in the image:

1. A motion mask is first derived by subtracting gaussian filtered versions of successive images and placing non-convex polygons around any motion found.
2. A region filling algorithm is applied to separate the mask into regions of disjoint polygons (using a 8-connectivity criterion).
3. Each of these regions is used to mask a contour image computed by a Canny edge detector.
4. The contour points are then tracked using the Lucas-Kanade pyramidal algorithm.
5. An affine model is built for each moving region from the position and velocity of the tracked points. Outliers are removed using the covariance estimate for such model.

Spatial events are then defined according to the type of objects' motion discontinuity [3]. This strategy was used to detect events such as grabbing, dropping, poking, assembling, disassembling or throwing objects, and to segment objects from such events by application of the grouping algorithm in Section 3.3. A random sample of segmentations is presented in Figure 11.

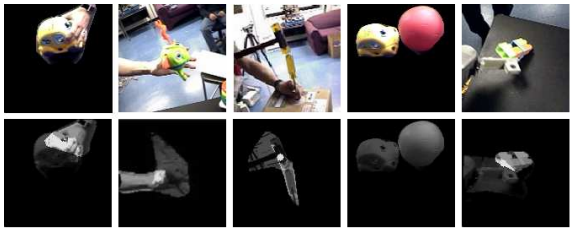

Figure 10. Segmentation of objects (bottom row) during discontinuous events (from left to right) Human grabbing a stationary car toy; Grabbing an oscillating fish toy; Human hammering a nail – segmentations from impact; two objects (a car toy and a ball) crashing; and the robot acting on objects by itself, enabling segmentation.

Figure 11. Sample of object segmentations from object's discontinuous motions actuated by humans and the robot. Not only the object's visual appearance is segmented from images, but also the robot's end-effector appearance.

5 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 or discontinuity content, the robot's visual system segments 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.

We proposed a grouping strategy to segment objects that are not allowed to move and therefore might be difficult to separate from the background. Such human-centered technique is especially powerful to segment fixed or heavy objects in a scene or to teach a robot through the use of books.

Objects were segmented in several cases from scenes where tasks were being performed in real time, such as hammering. 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 [5].

Human teachers facilitate children's perception and learning during child development phases. Similarly, through interactions of a robot with a human instructor, the latter facilitates the robot's segmentation task by providing additional grouping cues.

Acknowledgements

Project funded by DARPA as part of the "Natural Tasking of Robots Based on Human Interaction Cues" under contract number DABT 63-00-C-10102, and by the Nippon Telegraph and Telephone Corporation as part of the NTTand the author's institution Collaboration Agreement. Author supported by Portuguese grant PRAXIS XXI BD/15851/98.

References

[1] J. Aloimonos, I. Weiss, and A. Bandopadhay. Active vision. *Int. Journal on Computer Vision*, 2:333–356, 1987.

[2] Author. *Boosting Vision through Embodiment and Situatedness*. The publisher, 2002.

[3] Author. Embodied vision - perceiving objects from actions. *The publisher*, 2003.

[4] R. Bajcsy. Active perception. *Proceedings of the IEEE*, 76(8):996–1005, August 1988.

[5] C. Breazeal. *Sociable Machines: Expressive Social Exchange Between Humans and Robots*. PhD thesis, MIT, Cambridge, MA, 2000.

[6] 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.

[7] P. Fitzpatrick. *From First Contact to Close Encounters: A Developmentally Deep Perceptual System for a Humanoid Robot*. PhD thesis, MIT, Cambridge, MA, 2003.

[8] J. Harris. *Algebraic Geometry: A First Course (Graduate Texts in Mathematics, 133)*. Springer-Verlag, January 1994.

[9] B. K. P. Horn. *Robot Vision*. MIT Press, 1986.

[10] J. Malik, S. Belongie, J. Shi, , and T. Leung. Textons, contours and regions: Cue integration in image segmentation. In *IEEE International Conference on Computer Vision*, Corfu, Greece, September 1999.

[11] D. I. Perrett, A. J. Mistlin, M. H. Harries, and A. J. Chitty. Understanding the visual appearance and consequence of hand action. In *Vision and action: the control of grasping*, pages 163–180. Ablex, Norwood, NJ, 1990.

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