

Unsupervised discovery of objects using temporal coherence

Brandon C. S. Sanders^{1,2}, Rahul Sukthankar^{2,3}

¹Department of Computer Science
University of Rochester
Rochester, NY 14627

²Compaq Research (CRL)
One Cambridge Center
Cambridge, MA 02142

³The Robotics Institute
Carnegie Mellon University
Pittsburgh, PA 15213

sanders@cs.rochester.edu, rahul.sukthankar@compaq.com

Abstract

We present a novel unsupervised method for discovering objects in image sequences. Instead of using spatial homogeneity to partition pixels into regions, we group pixels into large temporally coherent clusters (TCCs) having a unique temporal signature. Each cluster’s temporal signature is explained by hypothesizing the arrival and departure of a small set of objects. Explanations of less ambiguous clusters are used to disambiguate the explanations of their more complicated neighbors. The recovered objects and arrival/departure events explain each frame in the image sequence by specifying (1) which objects are present, (2) the ways in which they occlude each other, and (3) the pixels in the frame that are not part of the background or the objects.

Our framework ignores distracting motion, correctly deals with occlusion (including mutual occlusion), and recovers entire objects even in cases where they are partially occluded in every frame. Because we do not use spatial information in our clustering steps the technique is significantly different from and complements traditional spatially based segmentation algorithms. The recovered 2D object masks are suitable for unsupervised training and initialization of object recognition and tracking systems.

1. Introduction

Computers capable of intelligent interaction with physical objects must first be able to discover and recognize them. “Object Discovery” is the problem of grouping all observations springing from a single object without including any observations generated by other objects. The entire spatial intelligence hierarchy (*i.e.*, understanding objects, relations, actions, and activities) hangs upon the problem of object discovery. Without object discovery, systems addressing any significant portion of spatial intelligence will necessarily be brittle, domain-specific, and require large-scale hand training by humans.

In this paper we describe a novel approach to object discovery that achieves good results for the quasi-stationary case (*i.e.*, the case in which all discovered objects are stationary for some small period of time). Our system performs object discovery by recovering and explaining the temporal structure found in each pixel’s stream of observations. While the results we present superficially resemble those obtained by segmentation algorithms that assume local homogeneity of object characteristics (*e.g.*, color, texture and flow vectors [1, 2, 4]) we use only temporal information to perform pixel clustering. As such, our approach is complementary to the existing body of segmentation work and represents a novel attack on the problem of object discovery. The advantages of our method include: (1) Low frame rate requirements (*i.e.*, 1-5Hz); (2) Entire objects are discovered even in cases where they are always partially occluded; (3) Complex occlusion relationships are recovered (including mutual occlusion); (4) Violation of our assumptions can be robustly detected, allowing us to discard problematic sequences rather than generate false hypotheses to explain them.

2. Algorithm

Our algorithm is divided into low-level, mid-level and high-level phases. The low-level operator considers each pixel independently through time and computes its temporal signature. The mid-level process groups pixels with consistent temporal signatures into large temporally coherent clusters (TCCs). Finally, the high-level reasoning phase explains each TCC at every frame by specifying the set of objects present in the TCC at the frame and the single object that is visible. From the TCC explanations we generate an interpretation of the entire sequence. This interpretation has two parts:

1. A complete description of each stationary object that arrives or departs during the sequence.



Figure 1: Selected frames from the sequence used as an example in this paper. In this sequence a person places and removes a chair, beanbag and wastebin. No frame contains the entire beanbag (*i.e.*, in all frames containing the beanbag it is partially occluded by either the chair or the wastebin). We correctly recover an entire exemplar of each stationary object in the sequence including the beanbag. Furthermore we provide a consistent explanation of every frame in terms of the stationary objects present, how they occlude each other and those pixels that cannot be explained by the stationary objects present (*i.e.*, the true foreground). (The sequence contains 320 frames digitized at 240×320 @5Hz, frames 0, 26, 87, 150, 190, 244, 276, and 299 are shown here.)

2. A complete reconstruction of each frame in the sequence that includes:

- The set of stationary objects present in the frame;
- The ways the objects occlude each other;
- The set of pixels in the frame that cannot be explained by the set of stationary objects present in the frame (*i.e.*, the true foreground pixels).

2.1 Computing Temporal Signatures

Our lowest level operator considers each pixel location independently and constructs its temporal signature. A pixel's temporal signature encodes when the pixel is stable and when it is in the midst of changing from observing one stationary object to observing a different stationary object (see Figure 2).

2.2 Clustering Temporal Signatures

The mid-level process groups pixel locations into temporally coherent clusters (TCCs) according to their temporal signatures. Because our high-level reasoning benefits from large regions of support, the clustering seeks to maximize cluster size while guaranteeing intra-cluster temporal consistency. It is important to note that while the obtained TCCs exhibit a high degree of spatial cohesion, spatial information is not used to construct them. Rather, the spatial cohesion arises from the reliable agreement of temporal structure found in pixels that have observed the same set of objects.

To perform the clustering we begin with a single large cluster containing all the pixels and split it into smaller and smaller clusters. Each time a cluster is split, the constituent clusters are guaranteed to be temporally more consistent than their cluster of origin. Splitting continues until each cluster contains only those pixel locations that agree about

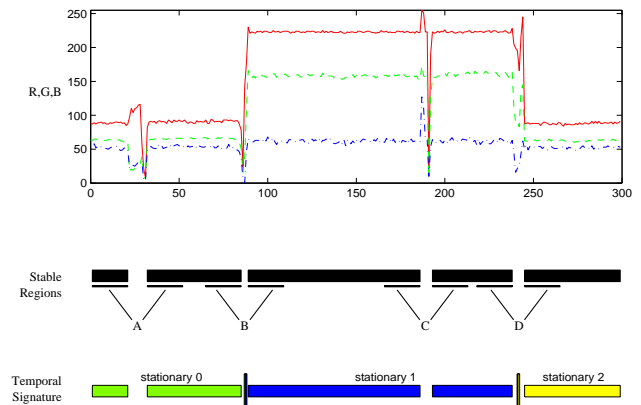


Figure 2: The temporal signature constructed by our system for the pixel at row 195 column 150. Step one examines each *atomic interval* (*i.e.*, a 4 sec long interval) for stability. If every observation on the interval is close to its mean the entire interval is marked as stable. Step two analyzes the temporal signature by comparing the atomic intervals on either side of each unstable region. As noted on the line containing the temporal signature, the comparisons at A and at C both indicate interrupted observation of a single stationary object (*i.e.*, stationary 0 is interrupted at A while stationary 1 is interrupted at C). In contrast, the comparisons at B and D suggest transition from observation of one stationary object to observation of another.

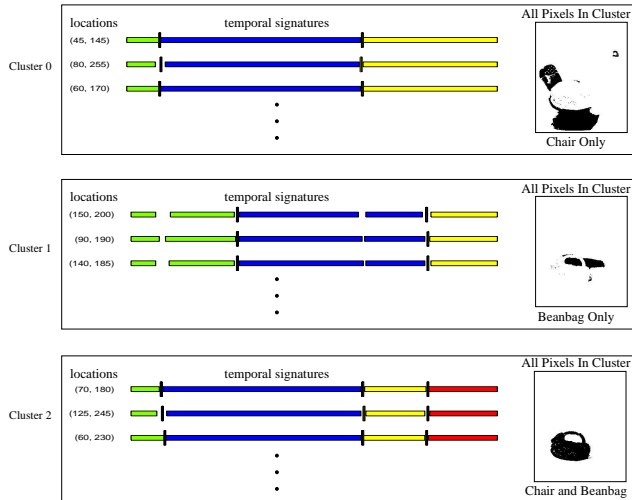


Figure 3: Grouping pixels into temporally coherent clusters (TCCs) according to their temporal signatures (see Figure 2). Starting from a single large cluster, each cluster is split until it contains only pixel locations that agree about when transitions from one stationary interval to the next occurred. The final clusters are temporally coherent, effectively containing only pixels that have observed exactly the same set of objects. For example the top cluster contains those pixels that observe only the chair (for our purposes its shadow and reflection are part of the chair). The middle cluster contains those pixels that observe only the beanbag, while the bottom cluster contains those pixels that observe both the chair and the beanbag. It is important to note that while the obtained TCCs exhibit a high degree of spatial cohesion, spatial information is not used to construct them.

when transitions from one stationary interval to the next occurred. Intuitively, these final temporally coherent clusters contain only pixels having observed exactly the same set of objects.

2.3 Event Labeling and Propagation

The high-level reasoning phase explains each TCC generated by the mid-level temporal grouping on a frame by frame basis. The explanation of a TCC at a particular frame specifies the set of objects present in the TCC at the frame as well as which object is in front of the others (*i.e.*, visible). Thus the reasoning phase proceeds in two distinct steps:

1. The TCC determines the set of objects present at the frame by adopting objects from its less complicated neighbor TCCs;
2. The TCC decides which object is in front of the others at the frame by comparing the appearance of the entire

region at the frame with its appearance in frames for which the front object is already known.

The reasoning in the first step of the event labeling and propagation is tractable because of a single important assumption: *Proportionally few pixels observe the boundaries of two distinct objects.* In other words, while object boundaries often intersect, they are rarely coincident for more than a few pixels at a stretch. Given this assumption, boundaries between TCCs occur at the edges of objects that occlude or are occluded by another object. Accordingly, for two neighboring TCCs the inner TCC has *always* observed exactly the same set of objects its outer neighbor has observed, plus one additional object its outer neighbor has not observed. Thus a TCC with outer neighbors that have each observed $N - 1$ objects has in turn observed N objects. For this reason, the outermost TCCs are the least ambiguous and always contain pixels that have observed a *single* object. Their transitional events can be confidently labeled as the arrival and departure of that object. Because their inner neighbors are guaranteed to have also observed the same object, the outer TCCs' labeled object events are propagated inward. Moving inward, each TCC's set of observed objects is exactly the set of all objects its outer neighbors have witnessed.

Step two of the TCC explanation phase has the same bootstrapping flavor as the first step. Instead of propagating objects and their events between TCCs, appearance models of the objects within the TCC are propagated from simpler frames to more complex frames, in turn generating new appearance models. Knowing which object is in front at a given frame allows us to initialize or update that object's appearance model using the values of the pixels in the TCC at that frame. Because the system knows the set of objects existing in the TCC at every frame, it can use the appearance models initialized in frames containing fewer objects to determine which object is in front for those frames containing comparatively more objects. For details of the reasoning performed in this phase, including special cases not covered in this sketch, please see [3].

3. Results

The event labeling and propagation phase discovers the true background and a small set of objects. The objects are described by 2D object masks along with their arrival and departure events. This set of background, objects, and object events explains the pixels in each frame, except for those pixels observing non-stationary objects in the frame (*i.e.*, the true foreground). An explanation of a particular frame in the sequence thus consists of the background, the set of objects present at that frame, the way in which those objects occlude each other, and the foreground observations of moving objects not explained by the background and stationary objects (see Figure 4).

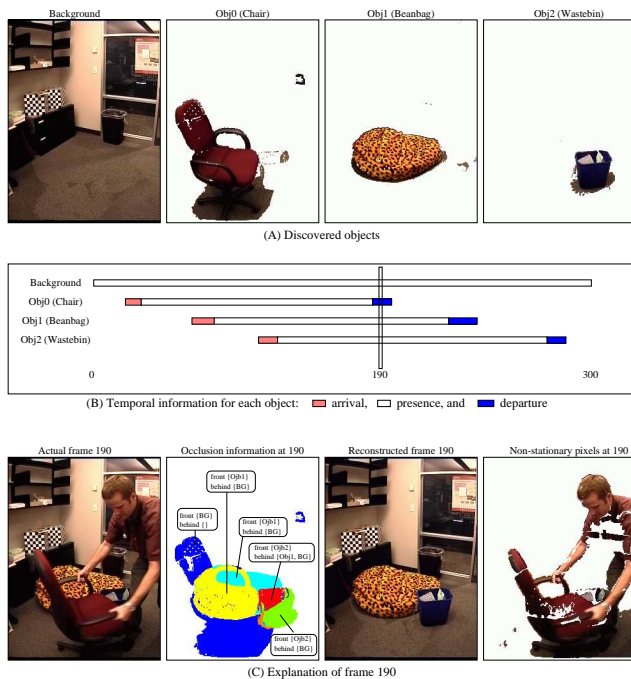


Figure 4: Results for the 320 frame beanbag sequence. (A) The stationary objects discovered include the background, chair, beanbag and wastebin. The entire beanbag is recovered despite being partially occluded whenever present. (B) The arrival, presence and departure intervals for each object. (C) The explanation of each frame in the sequence indicates which stationary objects are present and describes the occlusion relationships among them. Our framework successfully deals with mutual occlusion by determining the front object for temporally coherent regions (TCCs) rather than for entire object masks. Pixels in the actual frame but not in the reconstructed frame are labeled as non-stationary pixels and correspond to the true foreground. Although every frame in the sequence is explained, space constraints limit us to displaying only the explanation of frame 190. The holes in the recovered object appearance models correspond to portions of the object that look so much like the background that they appear stationary throughout the entire sequence. For instance, holes in the top of the chair are caused by the similarity of its appearance to the black £ling cabinet and the dark squares on the checkered cubes.

4. Conclusion

Our framework ignores distracting motion, correctly deals with occlusion (including mutual occlusion), and recovers entire objects even in cases where the objects are partially occluded in every frame. Because we do not use spatial information to perform our clustering, our technique is significantly different from and complements traditional spatially based segmentation algorithms. Additionally, sequences that violate our assumptions may be robustly detected and ignored. This characteristic well suits our approach to train and initialize object recognition and tracking systems without requiring human supervision. As such, our method represents significant progress toward solving the object discovery problem. For more details about the system, please see our companion technical report [3] available at <http://crl.research.compaq.com/>.

5. Acknowledgments

We would like to thank Randal Nelson at the University of Rochester for valuable discussions leading to this research.

References

- [1] Yucel Altunbasak, P. Erhan Eren, and A. Murat Tekalp. Region-based parametric motion segmentation using color information. *Graphical Models and Image Processing*, 60(1):13–23, January 1998.
- [2] Serge Belongie, Chad Carson, Hayit Greenspan, and Jitendra Malik. Color- and texture-based image segmentation using EM and its application to content-based image retrieval. In *Proceedings of the International Conference on Computer Vision ICCV98*, Bombay, India, January 1998.
- [3] Brandon C. S. Sanders and Rahul Sukthankar. Unsupervised discovery of objects using temporal coherence. Technical Report 2001/11, Compaq Research (CRL), One Cambridge Center, Cambridge, MA 02142, Aug. 2001.
- [4] John Y. A. Wang and Edward H. Adelson. Representing moving images with layers. *IEEE Transactions on Image Processing Special Issue: Image Sequence Compression*, 3(5):625–638, 1994.