

M. S. Thesis Pre-Proposal  
Enhancing the Implicit Shape Model via Object  
Tracking

Michael J. Mertsock

[mjm4837@rit.edu](mailto:mjm4837@rit.edu)

[www.mertsock.com/blog/category/rit/thesis/](http://www.mertsock.com/blog/category/rit/thesis/)

Rochester Institute of Technology

Department of Computer Science

102 Lomb Memorial Drive

Rochester, New York 14623

July 3, 2006

---

Michael J. Mertsock

---

Dr. Roger S. Gaboriski, PhD  
Committee Chair

## 1 Motivation

Object recognition—the identification of an object as a member of some known category—is one of the fundamental problems of computer vision. An ideal recognition system would automatically and accurately recognize the object (if it exists in the image), segment (locate) the object, and learn a model of the object category. Such an object category model would ideally provide the knowledge needed to recognize an object at any configuration or scale, in a cluttered scene with lighting and occlusion variations. Finally, the system should be able to perform these tasks upon arbitrary images or video with little information beyond the pixel values.

This is a lofty goal, but a system that efficiently satisfies most or all of these requirements would be beneficial in many ways. It could be used in monitoring applications to notify users of new subjects after long periods of empty scenes. Robots and other interactive vision systems could have more intelligent interfaces with their environments. True content-based image retrieval in databases would be more feasible, and innovative image editing tools could open new possibilities for media applications.

## 2 Background

*Object recognition* can simply be thought of as a decision problem: given an image  $\mathbf{D}$ , is an object belonging to some category  $C$  visible in the image? Most object category recognition systems, including the human visual system, naturally extend the solution to this problem to include detection, an estimate of the object's location. *Segmentation*, in this context, is the task of separating  $\mathbf{D}$  into two primary regions: figure (the identified object) and ground (the background), either of which might not be contiguous. *Object tracking* is the identification of an object's path through the spatiotemporal volume formed by a video. These activities all share many common aspects, and seem to be interrelated and interdependent in the human visual system. Therefore, it is feasible to search for a computer system that performs all of these tasks in an integrated manner.

Leibe and Schiele observed that “results from human vision indicate . . . that object recognition processes can operate before or intertwined with figure-ground organization and can in fact be used to drive the process,” [1] and this observation motivated the authors to develop the *interleaved* object recognition/segmentation method, also known as the Implicit Shape Model. This method is one of the now-popular parts-based approaches [2, 3, 4, 5] to object modeling and recognition.

The basic interleaved model [1] represents objects of some category (e.g. cars) as a collection of parts. Each part has an appearance—an exemplary image patch—and a vector relating the part to the object centroid. To detect (locate) an object in an image, candidate parts are located in the image using an interest point detector, and the pixels in each point's neighborhood are matched to the model's parts. When a match occurs, the location vector of the matched part points to the possible center of the object; this is recorded as a *vote*. A successful detection will find

that many of the object’s parts and the location vectors will point to roughly the same location, creating maxima in voting space. The parts that contribute to the maxima are recorded as belonging to the object and provide the information needed to determine the location and bounds of the found object. Matched parts that did not vote for this object’s centroid are discarded. Segmentation is carried out probabilistically, by considering the locations of the parts that voted and the weight of their votes. The interleaved model is created by extracting patches around interest points in segmented training images. Typically thousands of these patches are extracted from a set of training images; they are clustered into the final set of parts by (mutual) similarity of appearance using Normalized Grayscale Correlation [1], and the appearance of that part is represented by the average of the pixel intensities throughout the cluster. Each part retains all of the location vectors of its source patches, so during recognition a matched part may actually vote for multiple locations.

The interleaved model has been extended in various ways. Lebo [6] introduced various improvements to enhance the accuracy of recognition. He implemented *co-activation networks*, which adds a holistic perspective to the matched parts during recognition. This addresses the issue of strong false hypotheses by asserting that a true object will have matching parts whose relative locations are consistent with the training data; this favors hypotheses that incorporate the whole shape of the object as represented by the location vectors of the model. A concept followed by most authors is that a matching part belongs to no more than one actual object. Clark [7] implemented this using hypothesis starving, and Lebo developed a one-shot rule for voting.

Clark also applies the interleaved model to object tracking in video, using the Kalman filter to estimate an object’s future position and a probabilistic scheme for linking found objects in a frame of video to tracked objects in previous frames. Object tracking using a parts-based model has been implemented by other authors, including Ramanan. One approach taken by Ramanan [5] is to perform static image recognition to find a strong match in one frame of video, using a model representing a typical human pose (strong shape constraints) with no appearance constraints. An appearance model is then built for the found object based on the information in the video frame, and this enhanced object instance model is used in the remainder of the video to perform robust tracking.

### 3 Goals, Proposed Approach, and Evaluation

The proposed work will focus on a more robust and flexible application of the Implicit Shape Model (ISM) to recognition and tracking in video. While retaining the core of the ISM and interleaved recognition approach, an online learning method in the manner of Ramanan [5] will be investigated for tracking object category instances. The information learned while tracking object instances can be retained to incrementally build a richer object category model that can recognize both general instances of the category and also specific instances or subcategories. Instead of

the Kalman filter, CONDENSATION [8] will be investigated to achieve continuous tracking of objects. The graph cut algorithm by Boykov and Jolly [9] will be considered for final segmentation of the recognized/tracked objects. Whenever possible, the design and implementation of the model and tracking system will follow a biologically-inspired perspective, as has been done to various degrees in previous interleaved recognition efforts [1, 6].

The core of the object model will retain essentially the same structure and learning strategy as described by Lebo. An alternative representation for individual “parts” will be investigated: instead of raw image patches, curves extracted from the edge transform of the object’s image will be clustered and learned. This is based on the *boundary fragment model* described by Opelt et. al. [10]. It is assumed that a curve-based model will reduce the influence of the color and texture of training data and produce a more compact model that focuses on accurate location and shape.

Recognition, tracking, and segmentation in video will be performed in an integrated fashion, following the philosophy of interleaved object recognition. The system will process one entire video at a time. Given a video, the first step of recognition and tracking will be to perform Lebo’s interleaved recognition algorithm against individual frames of video. Co-activation networks and other advanced approaches of Lebo and other authors will be used to search only for high-confidence object detections. Using the new model, which should be less dependent upon appearance, and enforcing a high-confidence agreement to the structure of the model in the detected object, should result in higher-accuracy recognition at the cost of many frames of missed objects.

As explained by Ramanan [5], the low recall rate is inconsequential to the next step of the tracking algorithm, while the high precision is important. Inspired by his tracking method, the highest-confidence object detections will be used to build a high-quality instance model for the object being tracked. Appearance information (in the form of image patches or other color/texture features) will be extracted and appended to each part from the original shape model. Co-activation and other higher-level information can also be strengthened based on the statistics of the object detections being considered. This new instance model is then used to perform high-quality recognition and tracking of the object throughout the remainder of the video, as Ramanan does with the related pictorial structures model.

Leibe and Lebo already perform segmentation based on foreground/background likelihood information generated per pixel, based on nearby matched parts (or the absence of such parts). The segmentations have consistent quality in all reported results [1, 6] but are not “crisp”; that is, the procedure results in a fuzzy segmentation boundary based on the change in likelihood values rather than a strict binary segmentation. However, it should be straightforward to use the ISM recognition data to perform a graph cut segmentation [9]. It is proposed that the use of graph cuts will result in crisp and more accurate segmentations. Also, since graph cuts generalize to  $N$  dimensions, the method is directly applicable to segmentation in video with no additional work.

A common problem seen in the results of frame-based tracking algorithms is

discontinuity: the segmentation or bounding box may split, merge, flicker, change size, or move in arbitrary ways from frame to frame, even if there is an attempt to enforce some temporal coherence (e.g. through the use of Kalman filters). ISM has the additional difficulty of many hypotheses (voting space maxima) generated in each frame, which may cause the tracking algorithm to *drift* (erroneously track a string of false detections that lie tangent to the true path of the object). Instead of simply applying the recognition and segmentation procedures independently to each frame of video, it is proposed (and suggested in Clark’s conclusions [7]) that the CONDENSATION filter [8] can use the per-frame recognition information to track the object continuously and smoothly. This filter has the ability to handle multiple objects per frame and sort multiple detections and false positives into continuous tracks.

This approach to recognition and tracking opens some possibilities for incremental learning and improvement of the main object category model. Inspired by the work of McEuen [11], the instance models created during tracking could be saved as “exemplars”, and may be recalled when the same object instance or a similar object appears in the scene at a later time. For some object categories, such exemplars may accurately model entire subcategories instead of just single instances.

Another possibility to investigate is the use of interleaved tracking, recognition, and segmentation to learn novel poses (viewing angles such as front, side, and rear) of known object categories. For example, a model of side views of cars may be applied to a video in which the car rotates such that it directly faces the camera for a portion of the video. The recognition and tracking procedure described above will probably not perform well the forward-facing portion of the video, but (since the motion of an object is continuous) there should be continuous periods of tracking and segmentation on each end of the forward-facing time period. Also, the graph cut segmentation may extend from the tracked portions of the video a few frames into the untracked portion. Making the assumption that these extra segmented frames represent a novel view of the tracked object, a new pose model can be initialized. This model can be applied to all of the untracked frames, and a segmentation-training-matching cycle could be repeated until the number of recognized and tracked frames fails to increase. The result of this incremental learning strategy would be a model of a novel view of a known object category.

For verification of the foundations of this work, the same database <sup>1</sup> of training images and segmentations is available that was used by both Lebo and Leibe. The full system, of course, will focus on video, so video segments of cars and other objects taken at RIT will compose the bulk of the final data for recognition and tracking. The method will be tested quantitatively in a variety of ways for speed and accuracy of detection and segmentation.

As described above, the proposed work involves a number of enhancements to the ISM by combining methods from various works in the recognition, tracking,

---

<sup>1</sup> The TU Darmstadt Database,  
<http://www.pascal-network.org/challenges/VOC/databases.html#TUD>

and segmentation literature. These efforts will be guided by four common themes: a focus on biologically-inspired methods, improved tracking of objects in video, enhanced performance and accuracy through multiple levels of model specificity, and extensive quantitative testing.

## References

- [1] B. Leibe and B. Schiele, “Interleaved object categorization and segmentation,” *British Machine Vision Conference*, pp. 759–768, 2003.
- [2] S. Agarwal and D. Roth, “Learning a sparse representation for object detection,” in *ECCV ’02: Proceedings of the 7th European Conference on Computer Vision-Part IV*. London, UK: Springer-Verlag, 2002, pp. 113–130.
- [3] R. Fergus, P. Perona, and A. Zisserman, “Object class recognition by unsupervised scale-invariant learning,” in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, vol. 2. IEEE, June 2003, pp. 264–271.
- [4] M. P. Kumar, P. H. S. Torr, and A. Zisserman, “Extending pictorial structures for object recognition,” in *Proceedings of the British Machine Vision Conference*, 2004.
- [5] D. Ramanan, D. A. Forsyth, and A. Zisserman, “Strike a pose: Tracking people by finding stylized poses,” in *CVPR ’05: Proceedings of the 2005 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR’05)*, vol. 1. Washington, DC, USA: IEEE Computer Society, 2005, pp. 271–278.
- [6] T. Lebo, “Guiding object recognition: A shape model with co-activation networks,” Master’s thesis, Rochester Institute of Technology, July 2005.
- [7] D. S. Clark, “Object detection and tracking using a parts-based approach,” Master’s thesis, Rochester Institute of Technology, September 2005.
- [8] M. Isard and A. Blake, “Condensation—conditional density propagation for visual tracking,” *International Journal of Computer Vision*, vol. 29, no. 1, pp. 5–28, 1998.
- [9] Y. Boykov and M.-P. Jolly, “Interactive graph cuts for optimal boundary & region segmentation of objects in n-d images,” in *Proceedings of the International Conference on Computer Vision*, vol. 1. IEEE, July 2001, pp. 105–112.
- [10] A. Opelt, A. Pinz, and A. Zisserman, “A boundary-fragment-model for object detection,” in *Proceedings of the European Conference on Computer Vision*, 2006, to appear.
- [11] M. McEuen, “Expert object recognition in video,” Master’s thesis, Rochester Institute of Technology, October 2005.