

Hierarchical Probabilistic Models for Video Object Segmentation and Tracking

David Thirde & Graeme Jones

Digital Imaging Research Centre
Kingston University
Kingston-Upon-Thames KT1 2EE

<http://www.kingston.ac.uk/dirc>



The Problem: The goal of segmentation and tracking video objects in generic scenes is to segment the objects accurately and consistently depending on a set of semantics defined. Methods previously applied to this problem can be divided primarily into region-based or boundary-based methods. These two distinct approaches attempt to locate an object based on the semantic homogeneity of feature vector regions or by measuring gradient information in the feature space to locate object boundaries. Many techniques for object extraction allow a human operator to locate and define the semantic video objects to be segmented and tracked. Turning this problem into one of classification, the image data can be taken to be an array of feature vectors and a classifier can be used to assign the set of user provided labels to unlabelled feature vectors. The information contained in the feature vector and choice of classifier is dependent on the application and/or semantics defined.

Motivation: Semantic video objects with regard to the human visual system often have a complex underlying probability density function within the image feature space and hence previous approaches to this problem have applied many existing parametric and non-parametric forms of representation to extract such objects on an accurate pixel-wise basis. Parametric models have previously been applied to locate and segment video objects on a per-frame basis (e.g. Noel and O'Connor), although the functional form of the density model may not always provide a good representation of the object PDF within the feature space and the resulting segmentation mask quality is often degraded as a result. To overcome this problem non-parametric methods can be used to model sub-object homogeneous regions in the feature space, where the functional form of the model is determined by the data itself. A major drawback with these models is the lack of temporal stability in cluttered scenes where regions belonging to different objects can exhibit intra-region homogeneity within the feature space, causing severe artifacts in the segmentation masks.

We propose a novel hierarchical technique using parametric models to describe the appearance and location of an object and then use non-parametric methods to model the sub-object regions for accurate pixel-wise segmentation. Our motivation is to use parametric models to locate the object, improving the sensitivity of the non-parametric sub-object region models to background clutter.

Previous Work: There is limited work relating to the hierarchical modelling of video objects. The term 'hierarchy' is often used to describe algorithms where there is a label correspondence between a sub-object region level and an object level region (e.g. Marques and Llach [1]) or where hierarchical clustering algorithms are used. Some approaches to content based video retrieval work (e.g. Fu *et al* [2]) have explored the use of hierarchy to describe scenes using an interactive mapping of low-level motion features into semantic descriptors and also used coordinate system transforms to measure motion model temporal consistency. Our work is developed in the context of region based modelling of video objects and is similar in spirit to algorithms such as Raja *et al* [3], Marlow and O'Connor [4] and Everingham and Thomas [5].

Approach: We present a methodology incorporating a two tier hierarchy — parametric object level models, used for object location, and non-parametric sub-object level models, providing accurate pixel-wise object segmentation of the current scene. Object level regions represent areas of the image that contain semantically homogeneous features. Sub-object regions (which also have membership to a parent object) represent areas that have feature space homogeneity and intra-region inhomogeneity.

The feature vector observed at a pixel can contain chromatic, textural, spatial, motion or other feature measurements. We have found that the use of neighbourhood based features (i.e. texture or motion information) can often degrade the segmentation quality at the boundaries of objects, therefore the feature space PDF is modelled as a joint distribution of the chromatic and spatial signals. The feature vectors observed at the sub-object level are preprocessed using a Hotelling transform to locate them in the co-ordinate system of the moments of the spatial distribution of the parent object. This transformation provides invariance to rotation, scale and translation of the parent object and the resulting co-ordinate system can be seen in Figure 1.

We use a Gaussian spatial model and Gaussian Mixture Model chromatic model to describe the appearance and location of an object at the object level of the hierarchy. At the sub-object level we model both the spatial and chromatic distributions using Gaussian kernel density estimation, which is used to provide accurate pixel-wise segmentation.

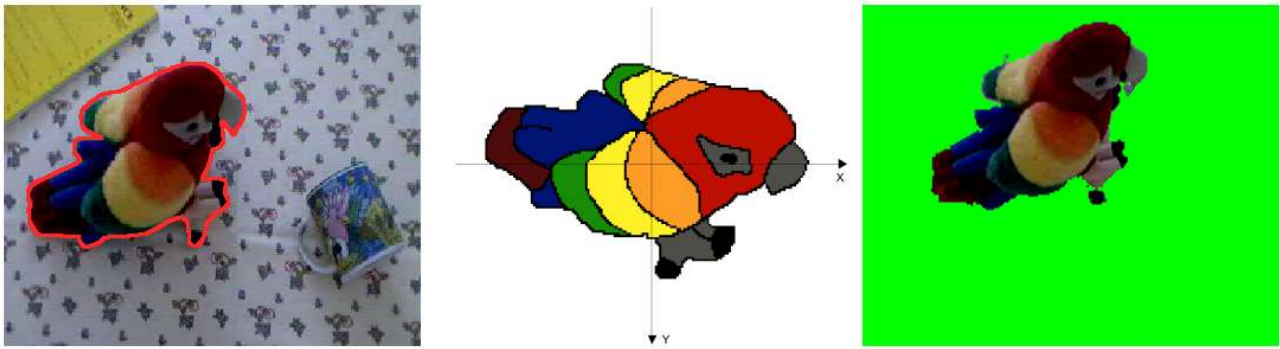


Figure 1 Showing the first frame mask (left), sub-object regions within the co-ordinate system of the parent object (centre) and segmentation result for frame 00018 of the *Parrot* sequence.

The algorithm is initialised with a user provided mask of the video objects (see Figure 1), from which the object level models can be built. The sub-object regions are created using the chromatic mixture model to hypothesise homogeneous regions within the objects boundary. We model the sequence on a per frame basis, that is, we do not attempt to explicitly model the temporal aspect of the data. With the models from frame $t-1$ (we make the assumption that the object at frame t has not moved a significant distance from frame $t-1$) we use an object level maximum a posteriori rule to calculate the object support in frame t , updating the object models. From the newly observed object moments we can perform the reverse Hotelling transform to the sub-object spatial models, from which a pixel-wise segmentation is performed in the image plane using a sub-object level MAP rule. From this we update the sub-object models and propagate both hierarchical level models to the next frame.

Results: A segmentation of the parrot object in the final frame of the Parrot sequence can be seen in Figure 1. The segmentation result contains all the semantically important features of the foreground object, although there are noticeable background artifacts caused by the uncovering of previously occluded background. The algorithm has been shown to outperform object based methods in quantitative comparison studies over a range of generic sequences such as the MPEG-4 standards *Bream* and *Foreman*.

Future Work: Includes making this approach able to handle changes in object pose, the innovation of new regions at both levels of the hierarchy is a difficult and challenging task. The update mechanism used to find the object support on a per frame basis is another interesting area of work – the update can be achieved using object based motion models, camera models, tracking with dynamic models, explicit temporal modelling or key frame interpolation. Beyond video object segmentation the hierarchical form of the model can be applied to other applications such as video surveillance or video compression.

Acknowledgments: The authors would like to thank DDD Group Plc., whom the authors thank for their financial and technical support of this work.

References:

- [1] F. Marqus and J. Llach. “Tracking of Generic Objects for Video Object Generation”. In *International Conference on Image Processing*, volume 3, pages 628–632, 1998.
- [2] Y. Fu, A. Ekin, A. Tekalp, and R. Mehrotra. “Temporal Segmentation of Video Objects for Hierarchical Object-Based Motion Description”. *IEEE Transactions on Image Processing*, 11(2):135–145, Feb 2002.
- [3] Y. Raja, S. J. McKenna, and S. Gong. “Segmentation and Tracking Using Color Mixture Models”. In *Asian Conference on Computer Vision*, volume 1, pages 607–614, 1998.
- [4] S. Marlow and N. E. O’Connor. “Supervised Object Segmentation and Tracking for MPEG-4 VOP Generation”. In *International Conference on Pattern Recognition*, volume 1, pages 1125–1128, 2000.
- [5] M. Everingham and B. Thomas. “Supervised Segmentation and Tracking of Non-Rigid Objects using a Mixture of Histograms Model”. In *8th IEEE International Conference on Image Processing*, pages 62–65, October 2001.