Spatio-Temporal Semantic Object Segmentation using Probabilistic Sub-Object Regions

D.J. Thirde[1],G. A. Jones[1] and J. Flack[2]
[1]Digital Imaging Research Centre, Kingston University, UK
{d.thirde,g.jones}@kingston.ac.uk
[2]DDD Group Plc, Bentley, Australia
jflack@ddd.com

Abstract

The MPEG-4 standard generated a need for the extraction of Video Object Planes for usage in video retrieval and description. Two later standards MPEG-7 and MPEG-21 further the need for systems requiring minimal user interaction for the accurate extraction of semantic video objects. Many previous approaches have either relied too heavily on user interaction or made compromises in end accuracy to achieve a faster segmentation process. With the advancement of computer processing power we propose a higher quality segmentation process with end application requiring a hardware implementation. The focus of our approach is to achieve a reliable and high quality segmentation mask per frame using sophisticated offline techniques to minimise the user interaction within the process.

1 Introduction

The segmentation of semantic objects within a video sequence is becoming increasingly important in many multimedia applications. The concept of a semantic object is dependent on the application — an object may have different significance for a human visual system or for video compression performance. We propose a flexible system to segment semantic video objects in generic scenes based on the perception of objects in the scene with regards to the human visual system, allowing meaningful objects to be located, segmented and tracked over a series of video frames. Since our work is developed in the context of video post production we introduce a user input stage. Unsupervised techniques for the segmentation of semantic objects from generic scenes do not exist [12] therefore human interaction is required to define the semantic video objects in one or more frames within the video sequence. Beyond the operator input we also take advantage of sophisticated offline algorithms requiring more processing power to improve the segmentation quality and reliability.

Segmentation methods for 2-D images may be divided primarily into region-based and boundary-based methods. The approach is either to locate objects based on homogeneity of localised feature vectors or by measuring gradient information in the feature space to locate object boundaries. Many previous attempts at segmenting using region-based

BMVC 2003 doi:10.5244/C.17.55

methods have applied chromatic, spatial, temporal or motion information to distinguish the homogeneous regions belonging to scene objects and have been used for applications such as visual surveillance [6], MPEG-4 encoding [10, 9] or more generic object segmentation schemes [2]. Under the framework of the European ACTS-MoMuSys project, the Video Object Generator with User Environment [9] provided a set of algorithms for spatial and temporal segmentation and object tracking, although the requirement of balancing efficiency and robustness provides scope for improving upon the segmentation process taking advantage of the increasing computational power of desktop workstations.

In a supervised region-based approach many researchers have turned the problem into that of classification. We take the image data to be an array of feature vectors and use a classifier to assign the set of user-provided object labels to this data set. There are many different types of classifier that can be used for this assignment, such as neural networks, clustering, decision trees, histograms or support vector machines. The feature vector extracted at a pixel often includes chromatic information alone [6, 11, 10]. The problem with this approach is that in scenes with several objects of similar colour distribution the objects can be assigned the same label due to a lack of spatial distinction. Gaussian mixture models have been proven to be an efficient and valid way to classify feature vector distributions in chromatic feature space. Additionally, spatial information can be appended to the feature vector to define the shape of an object [4, 8]. Other information used may include texture [1], motion [2] or depth information if available [7]. Texture and motion measurements use neighbourhood based operations. We have found they tend to degrade segmentation accuracy at the object boundaries in generic video scenes and are more suited to specific applications. For example, texture descriptors have been shown to improve the segmentation of animals exhibiting natural textured camouflage where colour alone cannot be used to separate the scene objects [4].

Gaussian based distributions are very useful for modelling data since they can be updated efficiently in contrast to other types of classifier. Clustering using Gaussian mixture models for segmentation and tracking of semantic scene objects has been explored in many papers. McKenna *et al* [11] and Marlow and Connor [10] showed how a mixture of Gaussians could be used to segment scene objects using a Bayesian segmentation rule. On a per frame basis, Gaussian mixture model's can be updated using the *expectation-maximisation*(EM) algorithm [3], this can be adapted to changes in the spatio-chromatic distribution over a video sequence. Figueiredo and Jain [5] propose an unsupervised algorithm for learning a finite mixture model that avoids several drawbacks of the standard EM algorithm — it is capable of selecting the number of components, is robust to initialisation and avoids the possibility of convergence to singular values within the feature space.

The addition of spatial data to Gaussian mixture models has proved to be a difficult integration. Whilst a Gaussian parametric model can be used to coarsely model the spatial extent of an object [1, 2], the updating and adaption of the model has proven a difficult task. Mixture models can propagate small segmentation errors over the course of an image sequence and the parametric form of the model is not well suited to modelling the spatial extent of an object over time. Everingham and Thomas [4] showed a non-parametric Gaussian kernel shape model to be a superior technique for fine modelling of the spatial extent of an object in a joint distribution with a coarse spatio-chromatic mixture of histograms. The spatial model in this approach did not take into account what we term articulated motion — significant differences in the motion of individual sub-object regions, knowledge of which can improve the adaption of the shape model for objects exhibiting regions of varying motion.

In this paper we present an algorithm combining Gaussian modelling of chromatic data with Gaussian kernel shape modelling of internal sub-object regions to overcome many shortcomings of previous approaches. The sub-object regions model the spatial areas with colour homogeneity. These regions are tracked spatially using an $\alpha - \beta$ filter and are updated on a per frame basis using a Bayesian segmentation scheme combining spatial and chromatic information.

2 Modelling the Probability Density Function

For a semantic video object, Λ', we can determine the joint probability of the spatial and chromatic information for all feature vectors belonging to that object in the current frame, I_t. For a feature vector, **a**, the joint distribution is given by the form

$$p(\mathbf{a}) = p(\mathbf{f})p(\mathbf{x}) = p(r,g,b)p(x,y) \tag{1}$$

The form of this expression assumes an independence between the chromatic information of an object and its spatial distribution. We model this *probability density function* (pdf) for each object using a combination of parametric and non-parametric models. We model the observed chromatic distribution for each object using a Gaussian mixture model comprising K multivariate densities of the form [3]

$$p(\mathbf{f}_i|\phi_k) = \frac{1}{(2\pi)^{\frac{d}{2}} \left|\blacksquare_k\right|^{\frac{1}{2}}} \exp\left[-\frac{1}{2}\left(\mathbf{f}_i - \mu_k\right)^T \blacksquare_k^{-1} \left(\mathbf{f}_i - \mu_k\right)\right] \tag{2}$$

where $\blacksquare_k$ is the covariance matrix and μ_k is the chromatic mean of the $k'th$ Gaussian cluster, ϕ_k. $\mathbf{f}_i$ is the chromatic observation at a pixel i. The *a posteriori* probability that this observation, $\mathbf{f}_i$, belongs to cluster ϕ_k is given by Bayes theorem

$$P(\phi_k|\mathbf{f}_i) = \frac{p(\mathbf{f}_i|\phi_k)P(\phi_k)}{p(\mathbf{f}_i)} \tag{3}$$

where

$$p(\mathbf{f}_i) = \sum_{k=0}^{K} p(\mathbf{f}_i|\phi_k)P(\phi_k) \tag{4}$$

Using this model for chromatic data we can determine the probability that an observation, $\mathbf{f}_i$, belongs to the object Λ'. We model the observed spatial distribution for each object using a mixture of sub-object Gaussian kernel shape models [3]. An object, Λ', contains S sub-object regions. The spatial pdf of each sub-object region, λ_s (where $s \in \{0 \ldots S\}$), is modelled using a Gaussian kernel density of the form

$$p(\mathbf{x}_i|\lambda_s) = \frac{1}{N}\sum_{n=0}^{N} g\left(\mathbf{x}_i, \mathbf{x}_n, \sigma\right) \tag{5}$$

where $\mathbf{x}_i$ is the spatial observation at pixel i, $g\left(\mathbf{x}_i, \mathbf{x}_n, \sigma\right)$ represents a bivariate Gaussian density function and $\mathbf{x}_n$ represent the N spatial observations at pixels within the sub-object region, λ_s. Evaluating this function will be expensive if the value of N is large.

Since the spatial distribution is 2-D we achieve a faster evaluation by forming a binary image for each sub-object region and convolve with a 2-D Gaussian kernel. This convolution forms a discrete map of the density over the image:

$$p[\mathbf{x}_i|\lambda_s] = \left[\frac{1}{N}\sum_{n=0}^{N}\delta\left(\mathbf{x}_i - \mathbf{x}_n\right)\right] \otimes G\left(\mathbf{x}_i, \sigma\right) \tag{6}$$

where δ is the Dirac delta function. Changing the value of σ in the model allows the uncertainty of the object shape to be changed, in scenes with large motion of sub-object regions setting σ higher allows the spatial model to adapt to such motions. The *a posteriori* probability that a spatial observation, $\mathbf{x}_i$, belongs to the sub-object region λ_s is given by Bayes theorem:

$$P(\lambda_s|\mathbf{x}_i) = \frac{p(\mathbf{x}_i|\lambda_s)P(\lambda_s)}{p(\mathbf{x}_i)} \tag{7}$$

where

$$p(\mathbf{x}_i) = \sum_{s=0}^{S} p(\mathbf{x}_i|\lambda_s)P(\lambda_s) \tag{8}$$

2.1 Initialisation

User input at a key frame is given in the form of a segmentation mask, as shown in Figure 1.

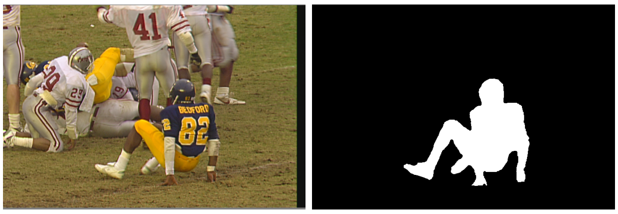

Figure 1: Showing the key frame(right) for the foreground object for frame 70 of the cif *football* sequence(left)

The key frame mask is not a ground truth segmentation mask for the frame, the areas of uncertainty around the object boundary are not marked. Using this key frame, we generate a Gaussian mixture model for each object in chromatic feature space using the Figuiredo and Jain [5] algorithm. This initialises the mixture and determines the number of clusters in an unsupervised process. On a per pixel basis for an object, Λ', spatial regions are generated from the Gaussian mixture model. This is achieved by choosing the cluster index k^* for pixel i that maximises the *a posteriori* probability, thus creating a disjoint partition of the image into K regions for each object.

$$k_i^* = \arg\max_k P(\phi_k|\mathbf{f}_i) = \arg\max_k \frac{p(\mathbf{f}_i|\phi_k)P(\phi_k)}{p(\mathbf{f}_i)} \tag{9}$$

To generate the initial sub-object regions we run a connected components algorithm on the disjoint partitions. The sub-object regions for the first frame are shown in Figure 2 for the foreground object. A minimum component size is used to remove semantically insignificant regions, the regions removed by this process are assigned to valid sub-object regions in the per frame update scheme.

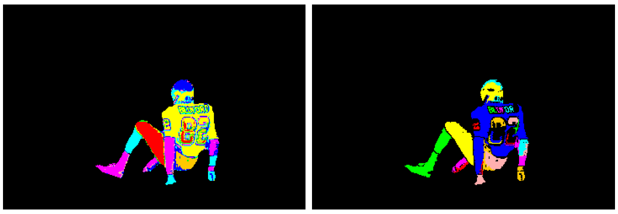

Figure 2: Showing false colour representations of the disjoint cluster partition(left) and the result of a connected components algorithm(right) with minimum region size of 25 pixels. Frame 70 for the cif *football* sequence.

For a sub-object region, λ_s, we parametrically model the assumed homogeneous colour distribution using a single multivariate Gaussian in chromatic feature space. This chromatic information is used in the Bayesian update rule for the sub-object spatial models shown in Section 2.2.

2.2 Update

On a per frame basis the sub-object spatial models are tracked and updated to adapt to the changing spatial distribution of the parent object. The chromatic models are not adapted, if the scene objects change appearance such that the chromatic information is no longer fitted by the model we assume the operator will use a new key frame from which the process is reinitialised.

For a new frame, the mean of the modelled sub-object spatial pdf is tracked and updated from previous observations using α-β filter tracking of the form

$$R_t = \alpha M_t + (1-\alpha) R_{t-1} \tag{10}$$

where R_t represents the state for the current frame, R_{t-1} is the state in the previous frame and M_t is the current frame observation data. When used to predict the translation velocity of the modelled sub-object spatial pdf, $\mathbf{v}$, the α-β filter becomes:

$$\mathbf{v_t} = \alpha \left(\mu_\mathbf{t} - \mu_\mathbf{t-1} \right) + (1-\alpha) \mathbf{v_{t-1}} \tag{11}$$

with μ denoting the mean of the modelled sub-object spatial pdf and the new observed velocity at time t, $\mathbf{v}'_\mathbf{t} = \mu_\mathbf{t} - \mu_\mathbf{t-1}$. The choice of the regularising factor α affects the damping characteristics of the update scheme, a value of 1 makes the tracking update independent of the state history, while making $\alpha = 0$ represents a constant velocity model where the current velocity is unchanged by data observations.

The sub-object spatial models are translated by the tracking process, dependent on the velocity predicted by this process. From the predicted sub-object spatial models we can formulate the new spatial distribution of the sub-object regions using the MAP update rule at each pixel observation $\mathbf{a}_i \in \Lambda'$, choosing the sub-object index s^* such that

$$s_i^* = \arg\max_s P(\lambda_s|\mathbf{a}_i) = \arg\max_s \frac{p(\mathbf{a}_i|\lambda_s)P(\lambda_s)}{p(\mathbf{a}_i)} \tag{12}$$

where, from Equation 1,

$$p(\mathbf{a}_i|\lambda_s) = p(\mathbf{f}_i|\lambda_s)p(\mathbf{x}_i|\lambda_s) \tag{13}$$

$p(\mathbf{x}_i|\lambda_s)$ is obtained from the sub-object spatial model and $p(\mathbf{f}_i|\lambda_s)$ is obtained using the Gaussian density model of the chromatic distribution in the sub-object region.

2.3 Video Object Segmentation

Bayes theorem is applied to classifying pixels in the scene with the most probable object label by choosing the label associated with the object classification Λ^* that maximises the *a posteriori* probability for a given observation at pixel i:

$$\Lambda_i^* = \arg\max_\Lambda P(\Lambda|\mathbf{a}_i) = \arg\max_\Lambda \frac{p(\mathbf{a}_i|\Lambda)P(\Lambda)}{p(\mathbf{a}_i)} \tag{14}$$

where, from Equation 1

$$p(\mathbf{a}_i|\Lambda) = p(\mathbf{f}_i|\Lambda)p(\mathbf{x}_i|\Lambda) \tag{15}$$

The prior probability for each object is obtained using the previous observed segmentation mask.

3 Results

To evaluate the proposed approach we use subjective analysis of the segmentation results of our process and that of an adaptive Gaussian mixture model [11] in spatio-chromatic image space. Objective measures of segmentation quality cannot yet describe the accuracy of the segmentation in respect to the semantics we define as being important. We analyse our results for two cif sequences — *football* (26 frames) and *claire* (50 frames), shown in figure 3, as well as presenting image compression/articulated motion analysis for the MPEG-1 *ice skater* sequence (81 frames) and key frame analysis for the qcif *foreman* sequence (200 frames).

The segmentation results for the proposed process are shown in figure 3. For the *claire* sequence the connected components algorithm partitioned the foreground object into 22 sub-object regions and the background object into 70 regions. Likewise, for the *football* sequence the sub-object region count was 48 and 226 respectively. For both sequences the minimum initial region size was empirically set to 16 pixels, equivalent to a 4×4 neighbourhood. In the update procedure we set the sub-object spatial model smoothing parameter empirically to $\sigma = 3$. We found in general scenes values between $\sigma = 1$ (for low motion) and $\sigma = 5$ (for high) to be sufficient for constrained spatial model adaption. In the sub-object tracking module, α was set to 0.5.

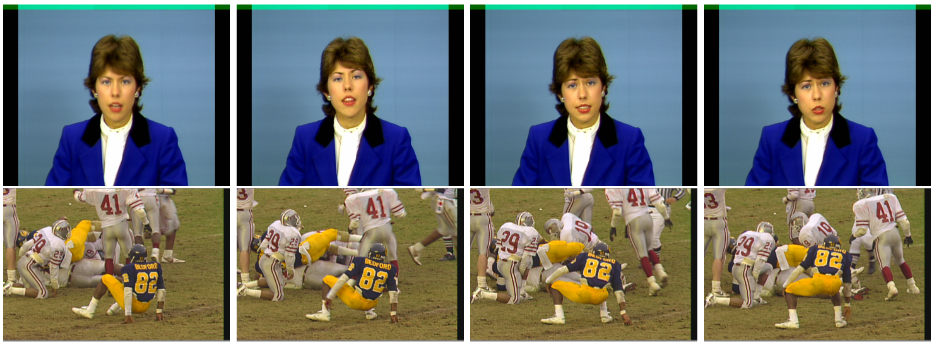

Figure 3: The sequences used for evaluation — the cif *claire* sequence (50 frames) and the more difficult cif *football* sequence (26 frames).

The results for the *claire* sequence show the proposed approach can produce a high quality segmentation in sequences with well defined objects. The more difficult *football* sequence highlights the problem of spatial adaption between video objects sharing identical chromatic space. The proposed approach demonstrates a well constrained spatial model in this sequence. By the final frame the segmentation mask, although degraded beyond usefulness, still contains a significantly greater proportion of foreground object pixels to mislabelled pixels. Our segmentation process has a failure mode when foreground regions occlude background regions of similar chromatic signal allowing the sub-object regions to migrate from foreground to background objects and vice versa, causing fragmentation of the segmentation mask.

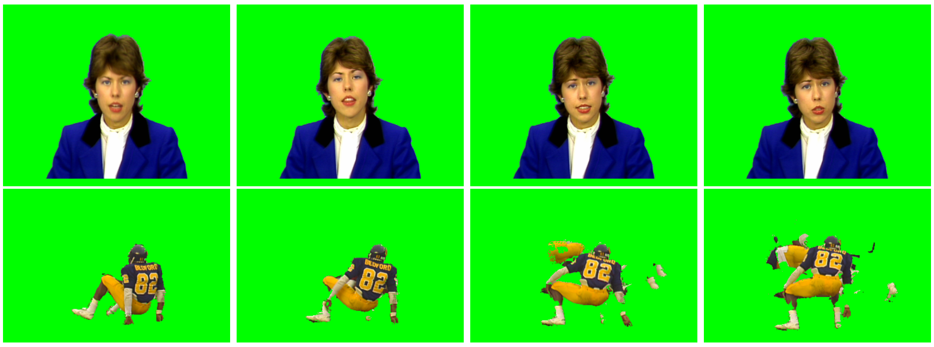

Figure 4: Segmentation Results using the proposed approach for the cif *claire* sequence (50 frames) and the more difficult cif *football* sequence (26 frames).

When comparing the results of our approach to an adaptive Gaussian mixture model, it can be seen that the unconstrained adaption of the mixture model in the spatio-chromatic feature space causes problems at object boundaries, where motion blur or translucent areas (e.g. hair) combine the chromatic signal of neighbouring objects. While the mixture model can model the initial few frames of the sequence sufficiently, the adaption is unconstrained and the parametric form of the Gaussian mixture is too generalised to model the subtle spatial changes of the semantic objects over a video sequence. In the *claire* sequence the mixture model performs adequately, the segmentation mask degrades when the model adapts to the background spatio-chromatic feature space due to the translucent hair silhouette. The *football* sequence shows how the parametric form for the Gaussians inadequately model the spatial distribution of the foreground object and subsequently the mask quality degrades significantly over the initial few frames. The number of clusters used for modelling the feature space were 9 for the background object and 19 for the foreground object in the *claire* sequence and 22 clusters for both foreground and background objects in the *football* sequence.

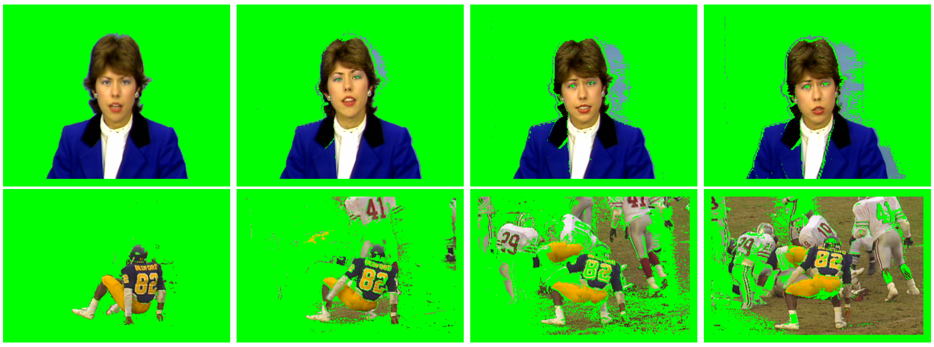

Figure 5: Segmentation Results using an adaptive Gaussian mixture model for the cif *claire* sequence (50 frames) and the more difficult cif *football* sequence (26 frames).

3.1 Image Compression, Articulated Motion and Sensitivity to Key Frame Choice

Compression artifacts in the video sequence can require the user to reduce the spatial model smoothing factor, σ, in an attempt to prevent the mis-adaption of the model over poorly defined object boundaries. In this scenario, the presence of articulated motion in sub-object regions can cause an object-level kernel shape model [4] to fail, as seen in figure 6. Using a mixture of sub-object kernel shape models, the image regions representing the hands of the skater can be tracked and updated successfully over the sequence. Notice how the MPEG-1 artifacts degrade the overall segmentation quality, especially when using an object level kernel shape model where the adaption of the model does not take into account the localised sub-object motion and chromatic distribution.

Figure 7 shows how a poorly chosen key frame leads to segmentation problems that can be avoided. The key frame at frame 0 contains little or no colour information for the eyes, hair or grey clothing of the subject. At frame 114 errors can be seen in the segmentation mask around the subjects clothing, hair and eyes. These areas have not been modelled sufficiently by the sub-object regions. By choosing a later key frame these

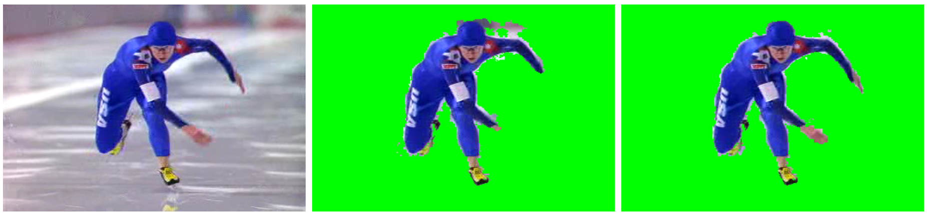

Figure 6: Showing the original frame (left), segmentation using an object level kernel shape model (centre) and segmentation using a mixture of sub-object kernel shape models (right) for frame 19 of the MPEG-1 *ice skater* sequence

problems are rectified and the clothing, hair and eyes are modelled more sufficiently.

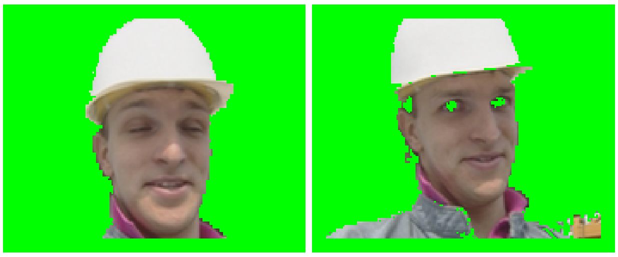

Figure 7: Showing the key frame segmentation 0 (left) and the segmented object at frame 114 (right) for the qcif *foreman* sequence

4 Conclusions and Future Work

We have shown in this paper a sophisticated offline algorithm that provides flexible and stable segmentation of semantic video objects within a video sequence. The user input is minimal, with only one key frame and three parameters required to start the process. We have demonstrated that modelling a joint probability using a Gaussian mixture model in chromatic feature space and a mixture of sub-object Gaussian kernel shape models as a valid approach for propagating a semantic segmentation mask over a video sequence.

Future work is anticipated in many areas of the proposed approach. These include improving the handling of occlusion with multiple scene objects, providing a frame work for unsupervised detection of new scene objects and allowing constrained adaption of the colour model. We aim to look at unsupervised techniques for choosing the algorithm parameters — the spatial model smoothing parameter, σ, the minimum region size in the connected components algorithm and the $\alpha - \beta$ filter α value.

References

[1] A.Rares and M.J.T.Reinders. “Object Tracking by Adaptive Modelling”. In *IEEE International Conference on Image Processing (ICIP2000)*, pages 74–77, Vancouver, Canada, September 2000.

[2] E. Chalom and V.M.Bove Jr. “Segmentation Of An Image Sequence Using Multi-Dimensional Image Attributes”. In *IEEE International Conference on Image Processing (ICIP96)*, pages 525–528, Lausanne, Switzerland, September 1996.

[3] Richard O. Duda, Peter E. Hart, and David G.Stork. *“Pattern Classification Second Edition”*. John Wiley and Sons, 2000.

[4] Mark Everingham and Barry Thomas. “Supervised Segmentation and Tracking of Non-Rigid Objects using a Mixture of Histograms Model ”. In *8th IEEE International Conference on Image Processing (ICIP2001)*, pages 62–65, October 2001.

[5] Mario A.T. Figueiredo and Anil K. Jain. “Unsupervised Learning of Finite Mixture Models”. *IEEE Pattern Analysis and Machine Intelligence*, 24(3):381–396, March 2002.

[6] Nir Friedman and Stuart Russell. Image segmentation in video sequences: A probabilistic approach. In *Uncertainty in Artificial Intelligence: Proceedings of the Thirteenth Conference (UAI-1997)*, pages 175–181, San Francisco, CA, 1997. Morgan Kaufmann Publishers.

[7] M. Harville, G. Gordon, and J. Woodfill. “Foreground Segmentation Using Adaptive Mixture Models in Color and Depth”. In *Proceedings of the IEEE Workshop on Detection and Recognition of Events in Video*, pages 3–11, Vancouver, Canada, July 2001.

[8] I. Kompatsiaris and M. G. Strintzis. “Spatiotemporal Segmentation and Tracking of Objects in Image Sequences”. In *IEEE International Conference on Image Processing (ICIP’99)*, pages 155–158, Kobe,Japan, October 1999.

[9] B. Marcotegui, P. Correia, F. Marques, R. Mech, R. Rosa, M. Wollborn, and F. Zanoguera. “VOGUE: The MoMuSys Video Object Generator with User Environment”. In *Proc. of Workshop on Image Analysis for Multimedia Interactive Services 1999 (WIAMIS’99)*, pages 25–28, Berlin, Germany, May/June 1999.

[10] Sean Marlow and Noel 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.

[11] Yogesh Raja, Stephen J. McKenna, and Shaogang Gong. “Segmentation and Tracking Using Color Mixture Models”. In *Asian Conference on Computer Vision*, volume 1, pages 607–614, 1998.

[12] D. S. Zhang and G. Lu. “Segmentation of Moving Objects in Image Sequence: A Review”. *Circuits, Systems and Signal Processing (Special Issue on Multimedia Communication Services)*, 20(2):143–183, 2001.