Video Object Segmentation using Multiple Features

Alvaro Pardo

IIE & IMERL - Faculty of Engineering
Universidad de la República
CC 30, Montevideo, Uruguay
&
Faculty of Engineering and Technologies
Universidad Católica del Uruguay

Abstract. In this paper we present an algorithm for semi-automatic object extraction from video sequences using multiple features. This work is part of an ongoing effort to study video segmentation using multiple features, and the relative contribution of each one of them. For this reason, the algorithm here presented will be very simple and made up from of the shelf algorithms. We will show that even with a simple algorithm, with the right steps, we can successfully segment video objects in moderate complex sequences.

1 Introduction

Video object segmentation is one of the most important and challenging problems in video analysis. Applications range from video surveillance and tracking, to video object-based coding and video databases.

Since the amount of literature about this subject is relatively vast, we will content ourselves with a review of the more relevant references, and the ones that are closely connected with our approach.

We can distinguish two main groups of algorithms: completely automatic algorithms, and the ones that require some interaction with the user. Among the later ones, we have all methods where the user must select the video object to be segmented along the sequence. Although these methods need the interaction with the user to select the object to be segmented, this interaction is indeed minimal. Is important to note that in this step the user introduces semantic information, or high level knowledge of the object. This is one of the reasons why these methods usually perform better than fully automatic ones. Usually, the system can aid the user with an initial coarse segmentation of the first frame. In addition, we can distinguish between region and boundary-based methods.

In existing approaches features may include only chromatic information, for example [1], or a combination of colour, spatial and motion information [2–6]. Usually these methods rely on statistical descriptions of the features. For example, Gaussian Mixture Models (GMM) are used in combination with maximum a posterior (MAP) to classify different regions in the sequence. This approach is

simple and efficient when dealing only with colour information, however, when including spatial and motion information, the results tend to deteriorate at the object boundaries. The reasons for this problem are, on one hand the fact that GMM usually produce small errors that then are propagated to future frames. On the other hand, the difficulty to reliably update the spatial information [7]. To overcome these problems we propose: first to regularize the posterior probabilities of object and background using an isotropic probability diffusion algorithm [8]. Second, to decouple spatial information from motion estimation steps for the update of spatial information. That is, we will estimate the new object shape to feed an Expectation-Maximization step (EM) in order to learn the new GMM parameters.

This work is part of an ongoing project to study video segmentation using multiple features. One of the main goals is to investigate a common framework for all the features used. Although recently several authors presented novel and successful algorithms, we believe that there is still a lack of information in order to judge all the different existing approaches. Specially, usually is difficult to tell which part of the algorithm is responsible of the overall success or failure. For this reason we decided to investigate a simply structured algorithm, made up from of the shelf algorithms without using any fancy and/or complicated methods. In this way, we will be able to evaluate the individual contribution of each feature and each step of the whole process. As we will see in section 2 all the building blocks of our algorithm can be replaced with others.

The structure of the paper is as follows. In Section 2 we present the proposed approach. In Section 3 we summarize some practical issues. In Section 4 we present some examples, and finally in Section 5 we discuss some future work and conclusions.

2 Proposed approach

Our work falls into the category of semi-automatic and region-based object segmentation based on GMM. At the beginning, the user must select the object of interest to be segmented along the sequence. We assume that whenever the object disappears from the scene or is completely occluded, the process must be re-initialised.

With the initial object and background classification, we learn the object and background GMM. This is similar in spirit to [4, 7, 6]. We describe object and background as a set of regions each one modelled with a Gaussian distribution. To learn the optimum parameters of the GMM we apply the well known EM algorithm [9]. The initial mean, and covariance matrices are estimated using the Kmeans algorithm [9]. Before describing the structure of our algorithm, we first describe the features used.

2.1 Features

The whole process relies on three different features: colour, position and motion. Each feature is different in nature and plays a different role in our method. This

contrast with some existing approaches where the full set of features is combined into an unique feature vector [5, 4]. In our case we append into the same feature vector, colour and spatial information, while we leave motion information for the update of object shape.

Colour Colour information is represented using the Lab space that is known to be perceptually meaningful. That means that distances in the Lab space correlate with perceived colour distances.

Position The spatial information is given by the (x, y) object and background pixels coordinates. In our method position plays two roles. Firstly, it is included in a feature vector together with colour information. Second, position constitutes the shape information that will be used to estimate the motion of the object.

The feature vector of colour and position is normalized to $[0, 1]$ before applying Kmeans and EM. The number of components in each mixture is fixed along the whole process. As in [6] the Minimum Description Length can be sued to estimate the optimal number of Gaussians. We will come back to this point in Section 5.

Given a new frame we can update the object and background GMM with the EM algorithm using as initial values the ones of the previous frame. In this way we can cope with variations in object and background along the sequence. This is especially useful in cases where the object or the background changes its model. For example, when the object deforms or moves. If the object moves or zooms, its spatial distribution also moves. Therefore, this step is crucial in order to track the position of each Gaussian in the mixture.

Finally, the posterior probabilities of object and background are regularized with an isotropic vector probability algorithm [8]. This is also very important to avoid problems due to small errors during MAP classification.

Motion Motion information is taken into account to estimate the objects shape. To track the object shape deformation we apply a simple block-matching algorithm between the previous and current frame. We use 3×3 blocks and a search area of ± 5 pixels. With the translation vectors obtained after the block-matching we estimate the objects shape that will be used to update the object and background GMM. Although simple, we observed that this process is quite robust and efficient. We also experimented with optical flow but it turned to be too unstable in complex and noisy scenes. In Section 5 we will come back to this point.

2.2 Algorithm Outline

1. Given the initial video object marked by the user, learn the models of object and background.
2. For all frames in the sequence:

(a) Apply the block-matching motion estimation between frames $t-1$ and t to obtain an estimation of the shape, $\hat{S}(t)$, at current frame t. After block-matching the estimated shape is regularized via mathematical morphology, its holes are filled, and the biggest connected component that matches the video object is selected.
(b) Using the points in $\hat{S}(t)$ the GMM for object and background are updated with EM.
(c) Before applying the MAP step, we regularize the posterior probabilities using the isotropic vector probability algorithm [8]. Given the posterior probabilities $p(o|(x,y))$ and $p(b|(x,y))$ of the pixel (x,y) to be object and background, we define the probability vector:

$$\mathbf{p}(x,y) = (p(o|(x,y)), p(b|(x,y))).$$

The iterative regularization procedure is then:

$$\mathbf{p}^{k+1}(x,y) = \mathbf{p}^{k}(x,y) + 0.25 * \Delta\mathbf{p}^{k}(x,y)$$

where $\Delta\mathbf{p}$ is the Laplacian of $\mathbf{p}$ at (x,y).
(d) Apply a MAP step to obtain the shape of the object at frame t, $S(t)$
(e) Regularize the obtained shape as in step 2a.
(f) Continue to next frame.

3 Practical Issues

In this section we describe several practical issues that we encountered to be crucial for the algorithm presented, and some implementation details.

Initialization Is very important to start the process with a good representation of the object and the background. We found out that if the initial models of the object and background do not correctly represent their content, the segmentation tends to deteriorate along frames. Hence, it is very important to start with good initial guesses for the Kmeans and therefore for EM.

Posterior probabilities regularization We used the isotropic version of the algorithm presented in [8] with four iterations. Although we could use the anisotropic version of the same algorithm that respects borders in a better way, we decided to use this simple one to understand the importance of this step. In fact, this step turned out to be very important to obtain a clear segmentation close to the object border.

Motion We also tested Horn-Schunck, and Lucas-Kanade [10] optical flow methods. However, it turned out that due to the amount of regularization imposed, these methods do not provide a good estimation of the shape. We are aware that there exist methods that allow the extraction of discontinuous optical flows. Nevertheless, since we wanted to use only standard algorithms we did not include them here.

4 Results

We now present some examples to show the performance of our algorithm. We selected two different sequences with different complexity. First, we have Claire sequence. In this case, the background is static and the object moves slowly from frame to frame. As we can see in Figure 1 the algorithm successfully extracts the object along the sequence. Although these results may not be very impressing since this sequence is relatively simple to segment, it is important to note that we are not segmenting the whole body of Claire but her head. This is indeed a harder problem since we need to clearly separate the head and body's features. As already discussed in Section 2 a bad description of object and background leads to misleading results. In this case, the head object tends to include the rest of the body as we increase the number of frames. As we can see in Figure 1, the algorithm proposed successfully segment the head object along the sequence without expanding it to include parts of the image originally marked as background.

The second example is the Foreman sequence (Figure 2). In this case, both, the object and the background move and slowly change. Once again, the results are very stable during the whole sequence. We also show in Figure 3 that the proposed algorithm can cope with occlusions. When the hand occludes the face it is included in the object. Then when it moves away the algorithm successfully selects the head as the main object. This is done mainly using the motion estimation and the position feature in the Gaussian mixture model.

5 Conclusions and Future Work

We showed how using a simple algorithm based on of the shelf algorithms; we can obtain good results in video object segmentation. In addition, we showed how to overcome some problems of GMM via using posterior probability regularization, and decoupled shape and model updating.

Regarding the relative contribution of each feature to the overall performance, we found out, as note in [7], that colour information alone is not very reliable. For this reason colour and spatial information must be appended into the same feature vector. Furthermore, to obtain a good classification a posterior probability regularization is essential.

On the other hand, shape updating must be very accurate in order to be used in the MAP classification step. Otherwise, the results close to object borders tend to deteriorate. Due to this observation, we did not include a shape probability in the MAP classification. In future work we will study other methods for shape motion estimation, for example affine versions of the Lucas-Kanade algorithm [11].

Finally, another important thing is the initialisation of the whole process and the possible change in the number of Gaussians. The initial condition of the EM determines the quality of the results. Therefore, for the future we leave the inclusion of method such as the ones presented in [12]. With respect with a

Fig. 1. Segmentation results for Claire sequence.

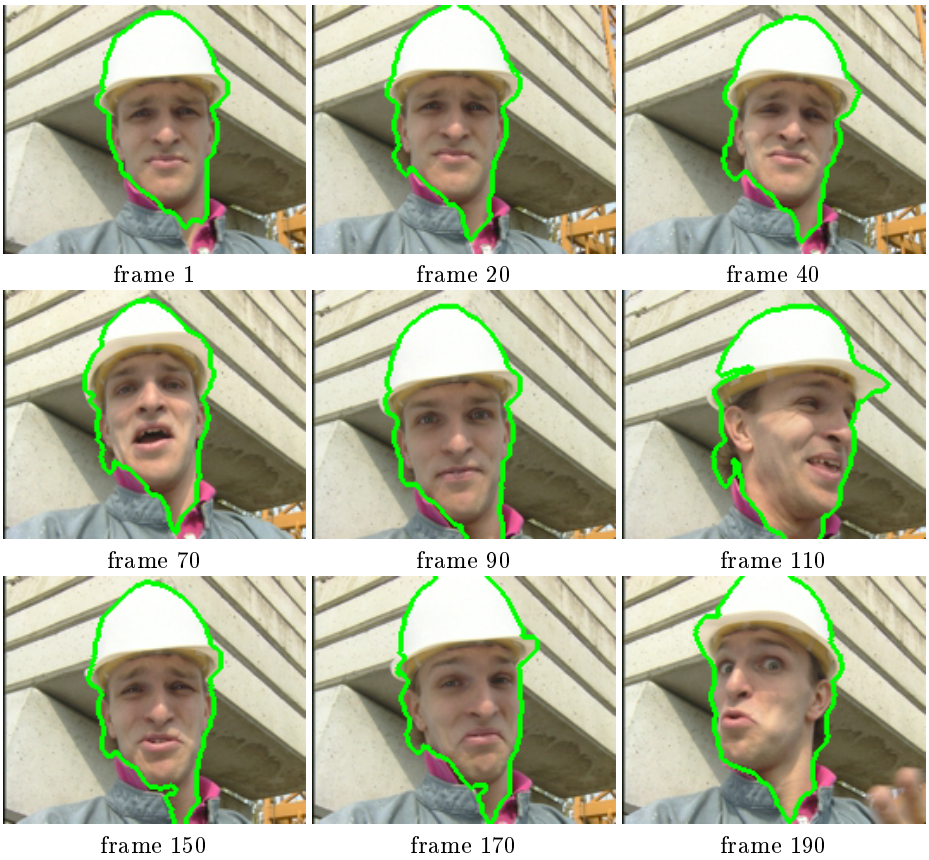

Fig. 2. Segmentation results for Foreman sequence.

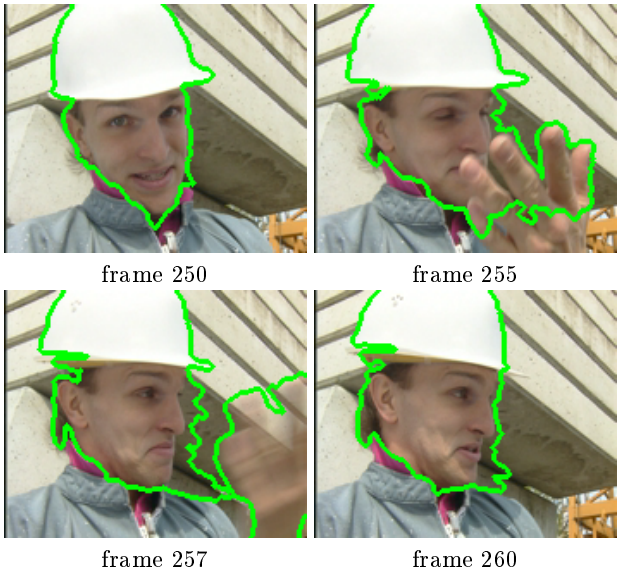

Fig. 3. Segmentation results for Foreman sequence in presence of occlusion.

varying number of Gaussians, we will explore the use of MDL. Although, feasible, this solution seems to be computationally demanding.

References

1. Marlow, S., Oconnor, N.: Supervised Object Segmentation and Tracking for MPEG4 VOP Generation. In: ICPR00 - International Conference on Pattern Recognition. Volume 1. (2000) 1125–1128
2. Castagno, R., Ebrahimi, T., Kunt, M.: Video Segmentation Based on Multiple Features for Interactive Multimedia Applications. IEEE Transactions on Circuits and Systems for Video Technology **8** (1998) 562–571
3. Gu, C., Lee, M.C.: Semiautomatic Segmentation and TRacking of Semantic Video Objects. IEEE Transactions on Image Processing **8** (1998) 572–584
4. Everingham, M., Thomas, B.: Supervised Segmentation adn Tracking of Nonrigid Objects using a Mixture of Histograms Model. In: ICIP01 - International Conference on Image Processing. (2001) 62–65
5. Khan, S., Shah, M.: Object Based Segmentation of Video using Color, Motion and Saptial Information. In: CVPR2001 - Int. Conf. Computer Vision and Pattern Recogbition. Volume 2. (2001) 746–751
6. Greenspan, H., Goldberger, J., Meyer, A.: Probabilistic Space-Time Video Modeling via Piecewise GMM. IEEE Transactions on Pattern Analysis and Machine Intelligence **26** (2004) 384–396
7. Thirde, D., Jones, G., Flack, J.: Spatio-Temporal Semantic Object Segmentation using Probabilistic Sub-Object Regions. In: BMVC2003 - British Machine Vision Conf. (2003)
8. Pardo, A., Sapiro, G.: Vector Probability Diffusion. IEEE Signal Processing Letters **8** (2001) 106–109
9. Duda, R., Hart, P., Stork, D.: Pattern Classification. Second edn. John Wiley and Sons (2000)
10. Barron, J., Fleet, D., Beauchemin, S.: Performance of Optical Flow Techniques. International Journal of Computer Vision **12** (1994) 43–77
11. Baker, S., Matthews, I.: Lucas-Kanade 20 Years on: A Unifying Approach. International Journal of Computer Vision **56** (2004) 221–255
12. Figueiredo, M., Jain, A.: Unsupervised Learning of Finite Mixture Models. IEEE Transaction on Pattern and Machine Intelligence **24** (2002) 381–396