Segmenting People in Meeting Videos Using Mixture Background and Object Models

Dar-Shyang Lee, Berna Erol, Jonathan J. Hull

Ricoh Innovations, Inc.
2882 Sand Hill Road, Suite 115, Menlo Park, CA 94025, U.S.A.
{dsl,berna,hull}@rii.ricoh.com

Abstract. We have developed a meeting recorder system which captures a panoramic video of a meeting room. Segmentation of people from this video is required for tracking and retrieval applications. However, the application scenario makes it difficult to rely on the usual solution of static background initialization and purely motion-based tracking for segmenting people. In this paper, we describe a novel framework for segmenting people in these videos using adaptive Gaussian mixtures for both background and object modeling. Based on a Bayesian formulation of the problem, results of object segmentation provide feedback to the background segmentation module. Experimental results on real meeting videos are presented.

1 Introduction

Object segmentation is an important component of many multimedia systems with applications ranging from content-based retrieval to object-based coding. Although it has been the subject of intensive study, fully automatic object segmentation is still considered an open problem. In our application, we are interested in tracking humans in videos captured by a single omni-directional camera. The realistic setting of our application makes it infeasible to rely on static background subtraction because some people will always be present when a recording is started. In this paper, we propose a solution using an adaptive mixture background model coupled with a mixture-based object model.

We have developed a meeting recorder prototype which captures a panoramic view of a meeting room. The capture device consists of an omni-directional camera equipped with four microphones. The camera can be positioned anywhere on a table and there are no specified seating locations for the participants. Recording is controlled manually via a VCR-like interface provided on a small touch-screen panel. Captured videos are dewarped to produce a full panorama of the meeting room. Analysis of the four-channel audio input allows us to identify the direction of speakers, and a special playback tool is used to provide a regular perspective view of the meeting which automatically centers on the speaker based on the results of the

Published in IEEE Pacific Rim Conference on Multimedia, Dec.16-18, Taiwan, 2002.
© Springer-Verlag

analysis. To improve the accuracy of view selection and to facilitate person identification, people segmentation is needed.

We propose a two-module system for tracking people using an adaptive background model and a blob-based object model. The background segmentation module identifies regions of interest that the object segmentation module classifies into individual objects. While this is the typical approach to most tracking systems, our system differs in that an adaptive mixture model is used for both the background and the object model. We also provide a Bayesian formulation of the background segmentation problem which allows the results of object segmentation and other domain-specific priors to be incorporated.

There has been a significant amount of work on tracking humans assuming various sensor inputs and environmental constraints [5],[6],[15]. The context of our application is similar to that of [13], where a single Gaussian recursive filter is used for background modeling and three clusters in the color space are used for object modeling. Our work differs in the background and object models used. Given the multimodal nature of the background process, a mixture model is more appropriate than a single Gaussian recursive filter [3],[11], and more efficient than the non-parametric approach of [1].

For background segmentation, we extend the work of [11] using an improved adaptive learning scheme and provide a Bayesian formulation of the segmentation problem. The generalized framework offers a guideline for incorporating domain-specific priors and feedback from the object model. Our object model is based on the blob representation of [15], where a person is composed of several blobs, and each blob is represented as a Gaussian distribution of spatial and color features. This representation was extended to handle multiple people and occlusions in [7]. However, in both applications, foreground is determined after background subtraction with static initialization. The existence of an image of the empty room was also assumed.

The rest of the paper is organized as follows. In Section 2, we provide a short description of our meeting recorder system to set the context for this work. In Section 3, we describe our framework for people tracking. Each of background and object segmentation modules is described in a subsection. Experimental results are shown in Section 4 followed by conclusions.

2 Meeting Recorder System

Recently, several multimedia systems have been proposed to facilitate the recording of meetings [2],[14]. Our meeting recorder system is designed with portability and compatibility with commercial hardware in mind. The hardware configuration consists of a special capture device, a touch screen monitor, as shown in Figure 1, and a PC. The capture device is composed of an omni-directional camera in the center and 4 microphones positioned at corners of the base.

The camera has a parabolic mirror that captures a panoramic view of the meeting in a single *doughnut* video stream. The video, along with digitally mixed stereo

audio, is sent to a video capture card and recorded as an MPEG-2 file. Encoding is done at 640x480 at 30fps. Although several other panoramic video capture systems based on multiple cameras have been proposed [2],[10], we chose a simpler design to avoid dealing with multiple video streams. This makes it feasible to replace the capture PC with a commercial digital video recorder for portability.

Before the audio signals are digitally mixed and sent to the video capture card, they are processed in real-time to determine the direction of speakers. A software filter receives four-channel audio and calculates the sound source direction between 0 and 360 degrees whenever human speech is detected. The results are post-processed to produce a sequence of virtual camera parameters used by our meeting viewer to automatically center on the speaker during playback. However, users can manually control the view using pan, tilt, and zoom operations.

The data on speaker directions is also used in combination with skin detection to extract face images of meeting participants. Background images are extracted from the video to identify the meeting location. This information is displayed in a meeting description document in HTML format along with user-added annotations. The audio is further analyzed to detect significant events. Based on motion analysis performed on the compressed data stream, events involving large spatial activities are identified. All the information associated with the meeting is written to a meta file. The video and meta file are archived and made available on a database server. The system has been used regularly since the beginning of 2002. As a result, we have collected more than 60 meetings and have over 50 hours of video. Additional details of the system are described in [8].

3 Algorithm Description

Our proposed solution consists of a background segmentation and an object segmentation module, as shown in Figure 2. The background segmentation module identifies regions of interest which can be further segmented and tracked by the object module.

In the background segmentation module, a Gaussian mixture is constructed for each pixel in the image to model the distribution of values observed at that pixel over time, effectively color quantizing the observations to discrete processes. A constraint for this quantization process is that it learns in an online mode and is adaptable over time. Each process corresponds to a distinct cluster and can be classified as foreground or background based on its color or other domain-specific attributes. Segmentation is then achieved by classifying a pixel value according to the class of the underlying process to which it belongs.

In the object segmentation module, regions identified as foreground are matched against a list of objects represented as blobs. New objects are created for regions that do not match any existing object. Object parameters are updated using the results of this classification. Information on the updated objects is used to improve the background model. Since objects are represented by spatially and chromatically

coherent blobs, processes similar to the object are less likely to be background. We describe each module in a subsection below.

3.1 Background Segmentation

For background segmentation, the decision at every pixel location over time is to separate observations resulting from a foreground process from those belonging to a background process. From a statistical pattern recognition perspective, this decision should be based on $P(B|x)$, where B represents the background class and x is the pixel value. Considering the multimodal nature of pixel values observed at a location over a time window, a Gaussian mixture is appropriate for modeling the distribution [3],[11].

$$P(x) = \sum_{k=1}^{K} P(x \mid G_k) P(G_k) = \sum_{k=1}^{K} w_k \cdot g\left(x, \mu_k, \sigma_k^2\right) \quad (1)$$

where G_k denotes the k-th Gaussian constituent, and $g(x, \mu_k, \sigma_k^2) \equiv g_k(x)$ is the normal density function computed at x.

Assuming all observations coming from a single Gaussian process belong to either foreground or background, the original posterior probability can be reformulated as

$$P(B \mid x) = \frac{\sum_{k=1}^{K} P(x \mid G_k) P(G_k) P(B \mid G_k)}{\sum_{k=1}^{K} P(x \mid G_k) P(G_k)} \quad (2)$$

The segmentation problem is decomposed as two independent problems of estimating the distribution of all observations at a single pixel, $P(x)$, as a Gaussian mixture, and estimating the probability of each Gaussian in the mixture being background, $P(B|G_k)$. The first problem corresponds to the quantization in Figure 2, and the second problem corresponds to the classification of each process as foreground or background.

The problem of density estimation using Gaussian mixtures is well studied. The constraint for our application is that an online, instead of batch, learning algorithm is needed and the model must adapt to distribution changes over time. We follow an adaptive filtering algorithm in [11] with a few modifications. The most significant change is that an adaptive learning rate schedule is introduced for each Gaussian to improve convergence. In addition to the normal weight, mean and variance parameters, a new parameter c_i is added to count the number of data points that have directly contributed to the update of the i-th Gaussian. We have found that using a separate learning rate for each Gaussian based on this number instead of using a fixed learning rate for all Gaussians in the mixture dramatically improves the convergence speed and approximation results [9].

Unlike the previous problem of density estimation where the objective and desired algorithm behaviors are well defined, estimating $P(B|G_k)$ is largely heuristic and application dependent. Since background is typically observed more often and displays less variation in value, we approximate $P(B|G_k)$ with $b_k^{r,c} = w_k^{r,c} \cdot (E_\sigma / \sigma_k^{r,c})$ where E_σ is the expected value of the variance for a background Gaussian. We estimate this by averaging the variance of the top 25% of Gaussians in the entire image that have the largest $P(B|G_k)$. The decision could also depend on other sensor inputs such as depth [4] if it were available. We also apply domain-specific priors

and neighborhood constraints. To enhance the detection of humans, a strong bias is placed against skin color being background. Skin tone is detected using a single Gaussian distribution in the normalized red-green color space 12. Finally, background estimates at neighboring locations reinforce each other. If the mean of a Gaussian at position (r,c), $\mu_k^{r,c}$, evaluates to be background with a high probability at its neighboring positions (r',c'), then its estimate $b_k^{r,c}$ is also incremented. This local reinforcement helps propagate established background models to regions where background is not well recognized as background. The object models described in the next section also provide feedback to this module.

3.2 Object Segmentation

At the end of the background segmentation process, every pixel position is classified as foreground or background. After smoothing out small noise and performing connected component analysis, regions of interest are identified. The goal of the object segmentation module is to match up these regions with a list of existing objects, or create new objects if necessary. Since the human body is highly deformable, we use a blob-based representation [7],[15].

A blob is simply a set of points sharing certain spatial and visual characteristics. Each object is composed of a set of blobs. This is naturally modeled as a Gaussian mixture of an augmented feature vector X consisting of the spatial coordinate (r,c) and the color components (R,G,B). The probability that a point belongs to an object O_j is

$$P(O_j \mid X) = \sum_{k=1}^{K} w_{j,k} \cdot g_{j,k}(X) \quad (3)$$

Every foreground pixel in a region of interest is matched against the list of existing objects. If most pixels in a region belong to the same object, then all foreground pixels in that region are assigned to that object, and the object model is updated. If the majority of pixels do not match any object well enough, then a new object is created. At the end of object classification, we obtain an object segmentation map where each pixel is labeled with an object ID or as background.

Object parameters are updated using the batch EM algorithm on all the samples in its support map. The parameters learned at time t are used as initial values for the update at time $(t+1)$. When an object is first created, one iteration of an online learning algorithm is used to initialize the Gaussians before applying batch learning.

Objects identified by this module are used to enhance the background segmentation module. If a blob corresponding to the torso has been identified, then Gaussian constituents in nearby background models that look similar to that blob are unlikely to be background. The background model is periodically updated by evaluating its Gaussians against the current list of objects. If $\mu_k^{r,c}$ matches an object well, computed by Eq.(3), its probability of being background is decremented. This feedback improves detection of stationary parts of the body that otherwise would be treated as background and in turn improves the object model.

4 Experiments

The proposed algorithm was tested on meetings recorded at our lab. For the background models, we used K=3 with a temporal retention α=0.99. Since $b_k^{r,c}$'s are expensive to compute, they are updated on every tenth frame. For the object segmentation module, K=5 and α=0.999. To obtain reasonable initial segmentation, object models are not created until 30 frames into the video. Both modules are implemented based on Microsoft's DirectShow. The system runs at 10 Hz with a 360x50 panoramic video on a 2GHz Pentium4 PC.

Segmentation results of a frame 10 seconds into a meeting video are shown in Figure 3. Fig.3(a) shows the video frame. From the motion history in Fig.3(b) it can be seen that all except the person on the right had relatively little motion since frame 0. Fig.3(c) shows a representation of the background model. The person on the right was easily detected as foreground and has become invisible in the background model. For the other three people, regions behind their heads that have not been completely occluded are showing through. Fig.3(d) shows regions of interest extracted by the background segmentation module. The rightmost person is well segmented since a good background estimation was achieved at that position. The bodies of the two people on the left were not segmented correctly because there was little motion. After image processing and object classification, the segmented object map is shown in Fig.3(e). Both people on the right were segmented correctly. However, there were holes in the two people on the left, which indicates that image processing can be improved. Fig.3(f) shows the object models that were constructed. The head and torso were well represented in position and color for the middle two persons. Although the models were not as good for the other two people, they were segmented correctly mainly by spatial proximity.

In summary, the preliminary results were very encouraging. We found that background segmentation worked well. A significant improvement was gained from our proposed adaptive rate algorithm for online mixture learning. We experimented with different selection criterion for Gaussian reassignment, various retention factors, and different color space representations. However, we did not notice any significant difference. One area for future investigation is the development of a more systematic approach to estimate $P(B|G_k)$. It is difficult to determine the relative importance of various priors and heuristics used in the background classifier. A strategy for training is needed.

The object segmentation algorithm can also be a subject of future investigation. Results obtained by simultaneously training background and object models depend on the object motion. While it handles moving objects quite well, it tends to break up hands from bodies if the person stays stationary for a while. This can be improved by incorporating priors into the object models and better object grouping. In addition, morphological processing can help fill up small holes. We are currently investigating these issues and preparing a more thorough evaluation using ground truth data.

5 Conclusions

We proposed an approach for automatic segmentation of people in meeting videos captured in a realistic setting. The application scenario calls for an adaptive background and object model. Our system simultaneously learns a mixture background model and blob-based object models for background and object segmentation. We also presented a Bayesian formulation of background segmentation based on Gaussian mixture modeling with an improved learning algorithm. Experimental results are encouraging.

References

1. Elgammal, A., Harwood, D. and Davis, L., "Non-parametric model for background subtraction," *Proc. 6th European Conference on Computer Vision*, (2000).
2. Foote, J. and Kimber, D., "FlyCam: Practical panoramic video and automatic camera control," *Proc. of Int. Conf. on Multimedia & Expo*, vol.3, (2000), 1419-1422.
3. Friedman, N. and Russell, S., "Image segmentation in video sequences: a probabilistic approach," *Proc. 13th Conf. Uncertainty in Artificial Intelligence*, (1997).
4. Harville, M., Gordon, G. and Woodfill, J., "Foreground segmentation using adaptive mixture models in color and depth," *ICCV Workshop on Detection and Recognition of Events in Video*, (2001), 3-11.
5. Gavrila, D. M., "The Visual Analysis of Human Movement: A Survey," *Computer Vision and Image Understanding*, vol. 73(1), (1999), 82-98.
6. Haritaoglu, I., Harwood, D. and Davis, L., "W^4: Real-Time Surveillance of People and Their Activities," *IEEE PAMI*, vol. 22(8), (2000), 809-830.
7. Khan, S. and Shah, M., "Tracking People in Presence of Occlusion," *Proc. of ACCV*, (2000).
8. Lee, D., Erol, B., Graham, J., Hull, J. and Murata, N., "Portable Meeting Recorder," *Proc. of ACM Multimedia*, (2002).
9. Lee, D., "A Bayesian Framework for Background Segmentation Based on Adaptive Gaussian Mixtures," *Proc. of Conf. On Systemics, Cybernetics and Informatics*, vol. 14, (2002), 76-81.
10. Rui, T., Gupta, A. and Cadiz, J., "Viewing meetings captured by an omni-directional camera", *ACM CHI*, (2001), 450-457.
11. Stauffer, C. and Grimson, W.E.L., "Adaptive background mixture models for real-time tracking," *Proc. CVPR*, vol. 2, (1999), 246-252.
12. Yang, J. and Waibel, A., "A Real-Time Face Tracker," *Proc. of WACV*, (1996), 142-147.
13. Yang, J., Zhu, X., Gross, R., Kominek, J., Pan, Y. and Waibel, A., "Multimodal People ID for a Multimedia Meeting Browser," *Proc. of ACM Multimedia*, (1999), 159-168.
14. Waibel, A., Bett, M., Metze, F., Ries, K., Schaaf, T., Schultz, T., Soltau, H., Yu, H. and Zechner, K., "Advances in automatic meeting record creation and access", *Proc. of ICASSP*, (2001), 597-600.
15. Wren, C., Azarbayejani, A., Darrel, T. and Pentland, A., "Pfinder: Real-Time Tracking of the Human Body," *IEEE PAMI*, vol. 19(7), (1997), 780-785.

Fig. 1 A meeting recorder prototype consisting of an omni-directional camera with 4 microphones at its base (right) and a touch-screen controlled interface on the left. The recording PC is not shown.

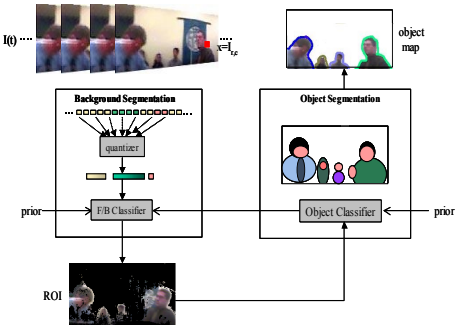

Fig. 2 A segmentation system consisting of two tightly coupled modules for background and object segmentation. The background segmentation module models each pixel with a Gaussian mixture and classifies each Gaussian constituent as foreground or background.

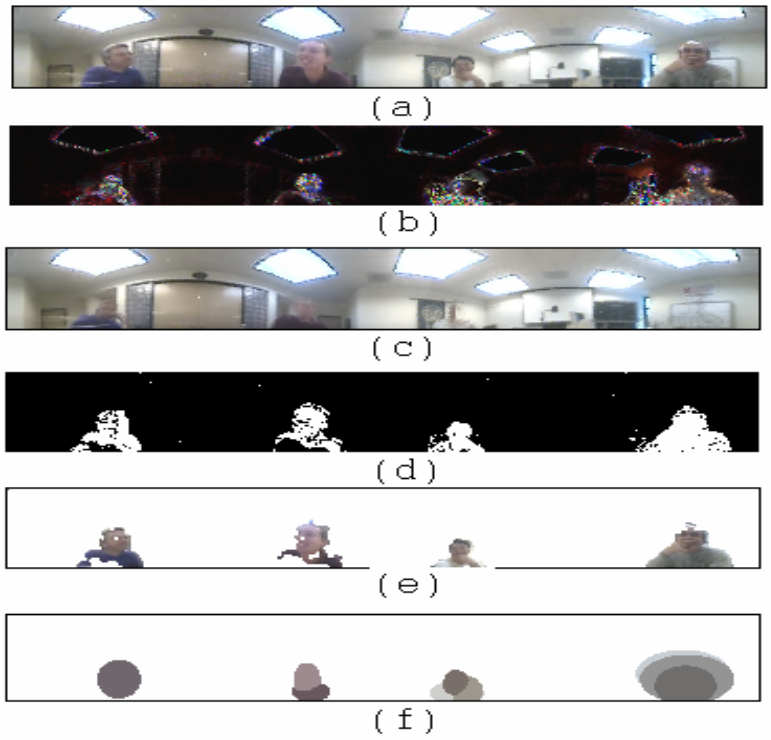

Fig. 3 - From top to bottom showing (a) the original panoramic video frame, (b) motion history prior to this frame, (c) a representation of the background model, (d) segmented foreground region, (e) foreground region after object segmentation, and (f) the object model.