SPATIOTEMPORAL ALGORITHM FOR JOINT VIDEO SEGMENTATION AND FOREGROUND DETECTION

S. Derin Babacan, Thrasyvoulos N. Pappas

EECS Department, Northwestern University
2145 Sheridan Rd, Evanston, IL 60208, USA
phone: + (1) 847-491-3039, fax: + (1) 847-491-4455
email: sdb@northwestern.edu, pappas@ece.northwestern.edu

ABSTRACT

We present a novel algorithm for segmenting video sequences into objects with smooth surfaces. The segmentation of image planes in the video is modeled as a spatial Gibbs-Markov random field, and the probability density distributions of temporal changes are modeled by a Mixture of Gaussians approach. The intensity of each spatiotemporal volume is modeled as a slowly varying function distorted by white Gaussian noise. Starting from an initial spatial segmentation, the pixels are classified using the temporal probabilistic model and moving objects in the video are detected. This classification is updated by Markov random field constraints to achieve smoothness and spatial continuity. The temporal model is updated using the segmentation information and local statistics of the image frame. Experimental results show the performance of our algorithm.

1. INTRODUCTION

A traditional representation of a video sequence is a collection of two-dimensional color images, where each image is a rectangular plane of pixels in a color space. With the improvements in image and video processing, a representation that is semantically more meaningful and useful is necessary for a wide range of applications, such as object-based video coding, video indexing, retrieval and video summarization. For example, MPEG-4 standard [1] requires a semantically higher presentation of video frames than traditional pixel presentation. In MPEG-4, the users are able to interact with objects in a video sequence, for example, different coding schemes can be applied to separate regions in a video. To be able to use the full extent of this standard, a system to extract regions of arbitrary shape is needed.

On the other hand, content description of video is standardized with the MPEG-7 standard [2]. In order to correctly identify the contents of a video, an initial segmentation into objects is crucial, since the semantic content is mostly created by moving/background objects and classification/identification of regions in video frames.

Various algorithms are proposed in the literature attempting to solve different aspects of video segmentation. An overview of the segmentation tools and representation schemes can be found in [3]. Most of the approaches are attempting to solve change detection and moving object localization problems. These approaches emphasize the moving (or foreground) objects in the content of the video and can be considered as a high-level binary segmentation. This problem is also known as background subtraction in the computer vision literature and is extensively studied (see [4] for a review).

On the other hand, approaches based on motion field segmentation, texture classification and segmentation and combinations of the above are also proposed. Background subtraction and foreground detection module by IBM [5] is a good example of combining color, motion, and texture information. However, most of the approaches focus in some subset of the video like a moving target and therefore do not give a complete object-based representation of the video-sequence. To exploit and analyse the full semantic content of the video, identifying and classifying objects might be very helpful in a number of applications.

This paper presents a novel clustering algorithm to segment color video sequences into spatio-temporal objects of uniform or slowly varying intensity over time and space. The algorithm clusters the pixels based on the color information and relative location. The intensity of each spatio-temporal volume is considered to be a slowly varying function plus Gaussian noise. Each image plane is modeled by a Markov random field (MRF), and temporal interactions between image planes are controlled by a probabilistic framework. Our algorithm involves moving object detection as well as segmentation of background scene in the video.

This paper is organized as follows: In Sec. 2, we describe the spatial and temporal models used for the segmentation. We explain the Markov random field approach applied to image frames and temporal classification of pixels using a probabilistic framework. The overview of the segmentation algorithm using this model is discussed in Sec. 3. We show the performance of our segmentation algorithm with foreground detection results in Sec. 4. We provide our conclusions in section 5.

2. MODEL

A video sequence can be considered to be a 3-D volume consisting of spatio-temporal regions of arbitrary shape, which evolve with time through the volume. Modeling the segmentation distribution of the video as a 3-D Gibbs-Markov random field is a very powerful tool since it considers all relations that exist between the pixels in the video. However, optimization of this framework is computationally inefficient, therefore rendering it impractical for most applications. In this work, we present an approximation to a 3-D Gibbs-Markov random field by applying Markov random field constraints spatially in each image plane and temporal classification using a statistical approach. As in [6], each spatio-temporal object is a region with uniform or slowly varying intensity. The sharp transitions only occur at the object boundaries, thus we are assuming no complex texture is present in the video.

The following subsections present the methods of modeling spatial and temporal interactions between pixels:

2.1 Modeling Spatial Interactions

Let $\mathbf{y}$ be the observed video sequence , $\mathbf{y_t}$ be an image plane at time t and $\mathbf{s}$ be the location of a pixel on $\mathbf{y_t}$. Thus, a pixel intensity $\mathbf{y_{s,t}}$ is indexed by spatio-temporal coordinates in a Cartesian grid as (s,t). A segmentation of the video sequence into regions is denoted by $\mathbf{x}$, where $\mathbf{x_t}$ is a segmentation of the frame $\mathbf{y_t}$ at time t, and $\mathbf{x_{s,t}}$ denotes the region where the pixel at location (s,t) belongs. The number of clusters in the video is K, where K is normally a small number like 3-5. Note that $K = 2$ results in a binary segmentation of the video. The segmentation information of a pixel is therefore given by

$$\mathbf{x_{s,t}} = i, \text{where } i = 1,2,..,K \tag{1}$$

Given the observed video sequence $\mathbf{y}$, the *a posteriori* probability density function for a segmentation can be formulated by using Bayes' theorem :

$$p(\mathbf{x}|\mathbf{y}) \propto p(\mathbf{y}|\mathbf{x})\, p(\mathbf{x}) \tag{2}$$

where $p(\mathbf{x})$ is the *a priori* density of the region distribution and $p(\mathbf{y} \mid \mathbf{x})$ is the observed video sequence given the distribution of the regions.

The region process $\mathbf{x}$ is modeled by a Markov random field, where $N_{s,t}$ is a neighborhood of the pixel at location (s,t) in the image plane at time t, and

$$p(x_{s,t}|x_{q,r}, \text{all } (q,r) \neq (s,t)) = p(x_{s,t}|x_{q,r}, (q,r) \in N_{s,t}) \tag{3}$$

The form of $p(\mathbf{x})$ can be derived from the equivalence of Gibbs and Markov random fields according to the Hammersley-Clifford theorem [7]:

$$p(\mathbf{x}) = \frac{1}{Z} \exp\left\{ \sum_C V_C(\mathbf{x}) \right\} \tag{4}$$

where $V_C(x)$ is a clique potential and Z is a normalizing constant. The summation is over all cliques C. The clique potentials V_C depend only on the pixels that belong to clique C. A *clique* C is a subset of the neighborhood system defined on the Cartesian grid, where every pair of distinct pixels in C are neighbors. In our model, we assume a neighborhood of 8 nearest pixels. A more extensive discussion of Markov-Gibbs random fields can be found in [8] and [9].

We assume that the only nonzero clique potentials are those that correspond to the two-point cliques. One point clique potentials are set to zero so that all regions are equally likely. The spatial two-point clique potentials are defined as:

$$V_C(\mathbf{x}) = \begin{cases} -\beta, & \text{if } x_{s,t} = x_{q,t} \text{ and } (s,t),(q,t) \in S \\ +\beta, & \text{if } x_{s,t} \neq x_{q,t} \text{ and } (s,t),(q,t) \in S \end{cases} \tag{5}$$

The parameter β controls the size of the regions and smoothness of the region boundaries. Larger β values favor bigger regions and smoother boundaries whereas smaller values result in more clusters and sharper edges. In this work, we use only spatial clique potentials that control the interaction between pixels in a single frame $\mathbf{y_t}$. As in [10], temporal interactions between frames can be controlled by introducing temporal clique potentials. However, as will be explained in Section 2.2, this requires optimization over a huge amount of data. In this work, we propose a novel probabilistic approach to classify pixels in the temporal direction.

Given a segmentation $\mathbf{x}$ of regions, the conditional density $p(\mathbf{y} \mid \mathbf{x})$ is modeled as a white Gaussian process with mean $\mu^i_{s,t}$ and variance σ^2. Each spatio-temporal region i is characterized by a slowly varying mean $\mu^i_{s,t}$ plus white Gaussian noise with variance $\boldsymbol{\sigma^2}$:

$$p(\mathbf{y}|\mathbf{x}) = \exp\left\{ -\sum_{s,t} \frac{1}{2\sigma^2} \left[y_{s,t} - \mu^{x_{s,t}}_{s,t} \right]^2 \right\} \tag{6}$$

The combined conditional density function in Equation 2 has the form :

$$p(\mathbf{x}|\mathbf{y}) \propto \exp\left\{ -\sum_{t,s} \frac{1}{2\sigma^2} \left[y_{s,t} - \mu^{x_{s,t}}_{s,t} \right]^2 - \sum_C V_C(\mathbf{x}) \right\} \tag{7}$$

We should note that the probability density function consists of two main components. The first constrains the region intensity function to be close to the observation y, and the second preserves spatial continuity. The temporal continuity is achieved implicitly by probabilistic modeling and spatial smoothing as shown in Section 2.2.

2.2 Modeling Temporal Interactions

In [10], temporal interactions between pixels are controlled by using temporal cliques as well as spatial cliques. However, this requires optimizing the probability density function over all the video sequence (or trying to attain a suboptimal solution with a subset of the sequence). Moreover, introducing temporal cliques preserves temporal continuity, but can cause fragmentation (segmentation of one object in multiple parts), whereas segmenting moving (foreground) objects as a whole is desired in many applications. Also, the motion generally has to be modeled explicitly and a classification between foreground (moving) and background (stationary, semantically uninteresting) objects has to be performed.

To overcome these problems, we introduce a probabilistic framework to classify pixels in a image $\mathbf{x_t}$ based on the previous segmentation information $\mathbf{x_{t-1}}$. Given the segmentation $\mathbf{x_{t-1}}$, a pre-classification is done based on local statistics of each pixel $\mathbf{y_{s,t}}$ to get an initial estimate of $\mathbf{x_t}$, and then the spatial constraints are applied using the Markov random field framework. To identify which pixels are foreground, after the initial segmentation $\mathbf{x_1}$ with K levels, each consecutive layer is segmented with $K+1$ levels, where the additional region label denotes the foreground (or moving object) regions. Thus, starting with a still background the algorithm detects the moving objects and incorporates them in the Markov random field.

The temporal distribution of each pixel is modeled as a mixture of K spatiotemporal Gaussian distributions where K is the number of clusters as in Section 2.1. In the neighborhood of $N_{s,t}$ of pixel (s,t), there exists maximum of K regions. Since each region samples a Gaussian distribution, the estimated probability of an observation $y_{s,t}$ is

$$P(y_{s,t}) = \frac{1}{K} \sum_{i=1}^{K} \eta(y_{s,t}; \mu_{i,s,t}, \Sigma_{i,s,t}) \tag{8}$$

where $\mu_{i,s,t}$ is the mean value of the i^{th} region at time t inside the neighborhood of pixel (s,t) , $\Sigma_{i,s,t}$ is the associated covariance matrix, and η is a Gaussian probability density function defined as:

$$\eta(y_{s,t};\mu,\Sigma) = \frac{1}{(2\pi)^{\frac{n}{2}}|\Sigma|^{\frac{1}{2}}} e^{-\frac{1}{2}(y_{s,t}-\mu)^T\Sigma^{-1}(y_{s,t}-\mu)} \quad (9)$$

The color channels are assumed to be independent with different variances, so the covariance matrix reduces to a diagonal matrix. Therefore, each region in the neighborhood of a pixel is represented by a single Gaussian with spatiotemporal parameters ($\mu_{i,s,t}$,$\Sigma_{i,s,t}$). This model is close to the assumption made on the region intensities in Section 2.1, so that the probabilistic model and Markov random field model are consistent.

The mixture of Gaussians in (8) is used to model the background pixel intensities. The foreground pixel distribution within this neighborhood is modeled as a single Gaussian as:

$$P(x_{s,t} = foreground) = \eta(y_{s,t};\mu^f_{s,t},\Sigma^f_{s,t}) \quad (10)$$

where $\mu^f_{s,t}$ and $\Sigma^f_{s,t}$ are calculated similarly by using the foreground pixel intensities in the neighborhood. We assume that only changes in lighting and small texture differences are present in a small neighborhood of a pixel (s,t), therefore a single Gaussian distribution is sufficient to represent the characteristics of the foreground object.

The probability of a new pixel intensity $\mathbf{y_{s,t}}$ is calculated using the K background distributions and the foreground distribution and the highest probability is thresholded. If this probability is higher than the threshold, the pixel label $\mathbf{y_{s,t}}$ is assigned to the region i which the distribution belongs. If the probability is lower than the threshold, it is considered to be a foreground pixel.

We assume that only background objects are present in the first image plane, so that spatial clustering of the first frame gives us the distribution of regions that characterize the scene. Foreground distributions are created in the future image planes as the algorithm processes. The model is initialized using the spatial segmentation of the first frame and the parameters of the distributions are initialized using the local statistics calculation shown in Figure 1. If the number of pixels that belong to region i within the neighborhood is too small, the estimates of $\mu_{i,s,t}$ and $\Sigma_{i,s,t}$ are not very reliable and therefore no distribution is created for this region. In this way, outliers can be discarded and distribution statistics can be more accurately estimated. Also, in the case of occlusion, background pixel distributions are preserved.

Temporal adaptation is very important to track the changes in dynamic scenes for a more accurate classification. In our model, adaptation in each image plane $\mathbf{y_t}$ can be done by reestimating the distribution parameters using the segmentation data $\mathbf{x_t}$, however, errors in the segmentation can lead to estimation problems and therefore to propagation of segmentation errors. Also, more weight has to be given to the present data than past. To achieve this, we use an autoregressive model to update the distributions :

$$\mu_{i,s,t} = (1-\alpha)\mu_{i,s,t-1} + \alpha\hat{\mu}_{i,s,t} \quad (11)$$

$$\mu^f_{s,t} = (1-\alpha)\mu^f_{s,t-1} + \alpha\hat{\mu}^f_{s,t} \quad (12)$$

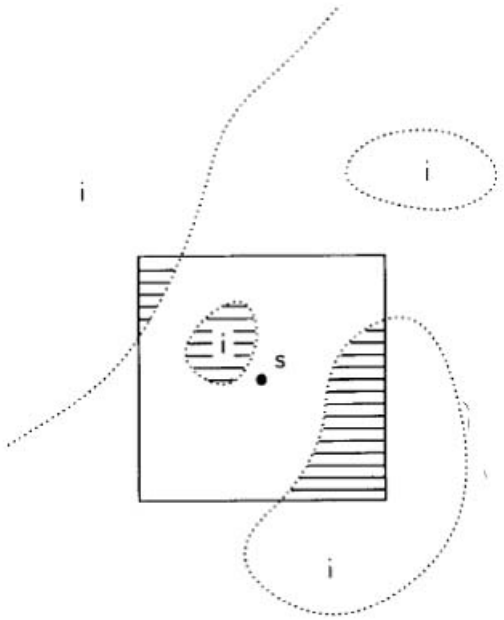

Figure 1: Local Statistics Estimation. The overlapped regions of the window and cluster i is used to estimate the statistics. This estimation is used both in the spatial segmentation and the temporal model.

$$\Sigma_{i,s,t} = (1-\alpha)\Sigma_{i,s,t-1} + \alpha(\hat{\mu}_{i,s,t}-\mu_{i,s,t})^T(\hat{\mu}_{i,s,t}-\mu_{i,s,t}) \quad (13)$$

$$\Sigma^f_{s,t} = (1-\alpha)\Sigma^f_{s,t-1} + \alpha(\hat{\mu}^f_{s,t}-\mu^f_{s,t})^T(\hat{\mu}^f_{s,t}-\mu^f_{s,t}) \quad (14)$$

where α is the learning rate with $0 \leq \alpha \leq 1$, $\hat{\mu}_{i,s,t}$ and $\hat{\mu}^f_{s,t}$ are the local intensity mean estimates calculated for background and foreground, respectively, within the neighborhood as shown in Figure 1. This updating filter (Eqs. 11-14) is mainly an exponential temporal weighting, so that current pixel intensities has more effect on the distribution statistics. The learning parameter α determines the rate of adaptation to the new pixel intensities.

It should be noted that although the update process is similar to the one used in [11], there is one fundamental difference in the data used to update the distributions: In our model, the distributions represent neighboring regions and they are updated by the average of pixel intensities over the corresponding regions after the classification is done, while in [11] each pixel is modeled contex-independent and adaptation is done using new pixel intensities. Smoothing the pixel intensities over each separate region results in more reliable estimates without sacrificing accuracy, which is one of the strengths of our model.

The proposed segmentation model differs from the 3D ACA proposed in [10] in the sense that, once the initial segmentation of the first frame is obtained, the MRF constraints in the temporal direction are eliminated to obtain more accurate foreground detection. After the foreground regions are detected, the MRF constraints are applied spatially to obtain smooth region boundaries and to prevent instability. In addition, for computational efficiency, the distribution function parameters are computed recursively by spatial averaging and temporal exponential weighting, instead of spatiotemporal averaging and optimization in 3-D space.

Control variables: K , α , W , β
Initialization: Obtain initial segmentation x using ACA
Calculate local statistics $\mu_{i,s,t}, \sigma_{i,s,t}$
Initialize Gaussian pdf's p_i using $\mu_{i,s,t}, \Sigma_{i,s,t}$
Initialize foreground Gaussian pdf p^f to NULL
while new data $\mathbf{y_t}$ **do**
$i = \text{argmax}_i \frac{|y_{s,t}-\mu_{i,s,t-1}|}{\sigma_{i,s,t-1}}$ *// Obtain new region label i*
// Temporal Labeling
if $\frac{|y_{s,t}-\mu_{i,s,t-1}|}{\sigma_{i,s,t-1}} > threshold$ **then**
pixel = *foreground*
pixel label = $K+1$
else if $\frac{|y_{s,t}-\mu^f_{i,s,t-1}|}{\sigma^f_{i,s,t-1}} < threshold$ **then**
pixel = *foreground*
pixel label = $K+1$
else
pixel = *background*
pixel label = i
end if
// Update
Calculate local statistics $\hat{\mu}_{s,t}$ using $\mathbf{x_t}$
Update temporal model by applying Eqs. 11-14
// Smoothing
Smooth $\mathbf{x_t}$ by MRF until convergence using β
end while

Figure 2: Proposed Algorithm

3. ALGORITHM

In this section we provide the general outline of the algorithm to segment the video in spatiotemporal volumes. We obtain an initial estimate of $\mathbf{x}$ of the first frame using the Markov random field model discussed in Section 2.1. This segmentation is obtained by iteratively estimating local averages μ^i_s and the segmentation $\mathbf{x}$ until convergence. Since we assume that no foreground objects are present in the initial frame, this segmentation gives the underlying region distribution in the overall video. Getting this initial estimate can be done using [6] with one initial frame or using [10] with a few initial frames.

After the initial segmentation result is obtained, the parameters of the Gaussian distributions are estimated per pixel as explained in Section 2.2 and shown in Figure 1. At each frame $\mathbf{y_t}$, the foreground is detected using these distributions and each pixel is assigned to a label $1, 2, ..K+1$, where the label $K+1$ denotes the foreground pixels. After the classification is done, Markov random field constraints are applied to smooth the region boundaries and to prevent oversegmentation. This smoothing is performed iteratively by alternating between local statistics estimation and relabeling the regions until convergence. Although in most cases the convergence is achieved within a few iterations, we limit the number of iterations to 10 for computational efficiency and to prevent oversmoothing. After the final segmentation is obtained, the distribution function parameters are updated as explained in Section 2.2. The overall algorithm is summarized in Figure 2.

4. EXPERIMENTAL RESULTS

In this section we show the performance of the algorithm. Figure 3 shows a frame of a MPEG-4 test sequence Hall Monitor with a spatial resolution of 176 x 144 (QCIF) and a temporal resolution of 30 frames/second. This video sequence is extracted from a low bitrate compressed bitstream, which involves major blocking artifacts. This is major obstacle for video segmentation algorithms since blocking artifacts create artificial edges and abnormal intensity distributions which may result in classification errors. We show that our algorithm is robust and gives good results even in this case.

Figure 3: A frame from MPEG-4 Hall Monitor sequence.

Figure 4: Initial segmentation result from the first frame.

The initial segmentation result with $K = 4$ from the first frame is shown in Figure 4. We can observe that this segmentation provides a sketch of the background objects, where each region with uniform (or smoothly varying) intensity is separated well and isolated pixels are eliminated. This initial segmentation provides regions with uniform pixel statistics which leads to skew probability distributions which lead to better foreground/background classification.

The foreground detection result is shown in Figure 5. Notice that although the video involves a significant amount of intensity changes in the background objects due to noise (see Figure 3), the foreground object is detected accurately and completely using the temporal model. Additionally, the parts of the object that are very close to the surrounding background objects are also detected by the use of the foreground distributions, which keep track of the moving region. At this phase, the algorithm assigns initial labels to regions.

Figure 7 shows the final segmentation result after the Markov random field smoothing is applied to the initial region classification. Notice that the segmentation of the moving object is preserved after the smoothing. The reader

Figure 5: Foreground detection result of the frame in Figure 3.

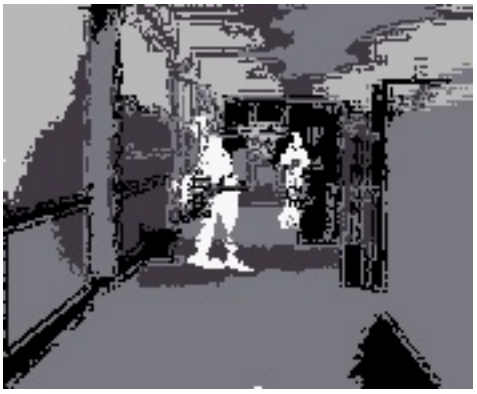

Figure 6: Initial segmentation result of the frame in Figure 3.

should note that although the foreground detection result in Figure 5 is accurate, region labeling using the temporal model removes the classification dependencies between pixels, and depending on the window size used to calculate the local statistics, discontinuities can occur in the segmentation (Figure 6). However by employing the iterative MRF smoothing, the regions are relabeled to obtain the spatial continuity and the necessary smoothness.

5. CONCLUSIONS

In this paper, we presented a novel algorithm for segmenting video sequences into spatiotemporal volumes. The segmentation of the video is modeled by a spatio-temporal hybrid model: The spatial interactions are controlled by a Gibbs-Markov random field and the temporal interactions are modeled by a mixture of Gaussians approach. Moving objects in the video are detected in an accurate and robust way, and the segmentation model adapts to scene dynamics very efficiently. Experimental results show that our algorithm gives a good classification of the objects and a robust object-based representation of the video. For the future work, we plan to incorporate texture and motion information both in the temporal and spatial model to obtain segmentation of spatiotemporal volumes with complex texture.

Figure 7: Final segmentation result of the frame in Figure 3.

REFERENCES

[1] T. Sikora, "MPEG-4 video standard verification model," *IEEE Trans. Circuits and Systems for Video Technology*, vol. 7, no. 1, pp. 19–31, 1997.

[2] T. Sikora, "The MPEG-7 visual standard for content description - An overview," *IEEE Trans. Circuits and Systems for Video Technology*, vol. 11, no. 6, pp. 696–702, 2001.

[3] P. Salembier and F. Marques, "Region-based representations of image and video: segmentation tools for multimedia services," *IEEE Trans. Circuits and Systems for Video Technology*, vol. 9, no. 8, pp. 1147–1169, Dec 1999.

[4] M. Piccardi, "Background subtraction techniques: a review," in *Proc. of IEEE SMC 2004 International Conference on Systems, Man and Cybernetics*, The Hague, The Netherlands, Oct 2004, vol. 4, pp. 3099–3104.

[5] Y. Tian, M. Lu, and A. Hampapur, "Robust and efficient foreground analysis for real-time video surveillance," in *Proc. of the 2005 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'05)*, Washington, DC, USA, 2005, vol. 1, pp. 1182–1187.

[6] T. N. Pappas, "An adaptive algorithm for image segmentation," *IEEE Trans. Signal Process.*, vol. 40(4), pp. 901–913, April 1992.

[7] J. Besag, "Spatial interaction and the statistical analysis of lattice systems," *Journal of the Royal Statistical Society*, vol. B36, pp. 192–236, 1974.

[8] H. Derin and H. Elliott, "Modeling and segmentation of noisy and textured images using gibbs random fields," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 9, no. 1, pp. 39–55, 1987.

[9] S. Geman and D. Geman, "Stochastic relaxation, gibbs distributions, and the bayesian restoration of images," *IEEE Trans. Pattern Anal. Machine Intell.*, vol. PAMI-6, no. 6, pp. 721–741, Nov. 1984.

[10] R.O. Hinds and T.N. Pappas, "An adaptive clustering algorithm for segmentation of video sequences," in *International Conference on Acoustics, Speech, and Signal Processing. ICASSP*, May 1995, vol. 4, pp. 2427–2430.

[11] C.Stauffer and W. E. L. Grimson, "Learning patterns of activity using real-time tracking," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 22, no. 8, pp. 747–757, 2000.