BAYESIAN FOREGROUND SEGMENTATION AND TRACKING USING PIXEL-WISE BACKGROUND MODEL AND REGION BASED FOREGROUND MODEL

Jaime Gallego, Montse Pardàs

Universitat Politècnica de Catalunya

Gloria Haro

Universitat Pompeu Fabra

ABSTRACT

In this paper we present a segmentation system for monocular video sequences with static camera that aims at foreground/background separation and tracking. We propose to combine a simple pixel-wise model for the background with a general purpose region based model for the foreground. The background is modeled using one Gaussian per pixel, thus achieving a precise and easy to update model. The foreground is modeled using a Gaussian Mixture Model with feature vectors consisting of the spatial (x, y) and colour (r, g, b) components. The spatial components of this model are updated using the Expectation Maximization algorithm after the classification of each frame. The background model is formulated in the 5 dimensional feature space in order to be able to apply a Maximum A Posteriori framework for the classification. The classification is done using a graph cut algorithm that allows taking into account neighborhood information. The results presented in the paper show the improvement of the system in situations where the foreground objects have similar colors to those of the background.

Index Terms— Foreground Segmentation, space-color models, tracking.

1. INTRODUCTION

Accurate and robust segmentation and tracking of moving objects in dynamic and cluttered visual scenes is a big challenge in computer vision. It is used in video surveillance applications in order to allow a correct object identification and tracking. In 3D multi-camera environments, robust foreground segmentation allows a correct 3-dimensional reconstruction without background artifacts. In addition, such systems are the building blocks of higher-level intelligent vision-based or assisted information analysis and management systems with a view to understanding the complex actions, interactions, and abnormal behaviors of objects in the scene. Another application is to use it as a video editing tool to combine objects from different video data. In this paper we focus on applications with a fix camera and our objective is to obtain an accurate segmentation and tracking of the foreground objects.

Thank to project CENIT-VISION 2007-1007.
Thanks to Ramon y Cajal Program.

Over the recent years there have been extensive research activities in proposing new ideas, solutions and systems for robust object segmentation and tracking to address the above situations. Most of them adopt the background subtraction as a common approach for detecting foreground moving pixels, whereby the background scene structures are modeled pixel-wise by various statistically-based learning techniques on features such as intensities, colours, edges, textures etc. The models employed include parallel uni-modal Gaussians [6], Gaussian Mixture Model (GMM) [12], nonparametric Kernel density estimation [4], or simply temporal median filtering [15]. A connected component analysis (CCA) is then followed to cluster and label the foreground pixels into meaningful object blobs, from which some inherent appearance and motion features can be extracted. Finally, there is a blob-based tracking process aiming to find persistent blob correspondences between consecutive frames.

If a foreground model is available, a Bayesian approach for foreground segmentation and tracking can be performed. In order to create the models, an initial segmentation is usually performed using an exception to background method, and once there is sufficient evidence that the foreground entities are in the scene, foreground models are created. Several foreground models have been proposed in the past for different purposes including the mentioned foreground segmentation task [10, 7, 8] and also in object and person trackers where the foreground has been previously segmented [9, 4]. Similarly as with background models, foreground models are Gaussian-based in most of the cases. For instance, single-Gaussians have been used in [13], GMMs have been used in [9, 7] and nonparametric models with Gaussian kernels, in [10, 11]. In [7] people are first segmented with the exception to background approach and tracked by segmenting them into classes of similar color (initialized by Expectation Maximization, EM). Each pixel is assigned in the following frames to the class that maximizes the probability of the pixel to belong to that class (including a class for the background). Means and variances of the classes are updated after classification. However, the partition of the object in regions modeled by independent Gaussians is too rigid and prone to errors. [9] uses a GMM to model the color distribution of the objects to track and EM to update its distribution. Since the objective is to track a single object, a background model is not used and thus

a complete segmentation is not achieved. In [14] a GMM for modeling both the foreground and background, in spatial and color domains, is used. The models are first initialized using a reference frame and the background and foreground models are adjusted using the EM algorithm. The classification is made in a Bayesian framework using the graph cuts algorithm [1]. The adaptation of the models to the next frame is done combining the background and foreground models in a generative model of the image. Finally, the EM algorithm is used to adjust the spatial components of the models before doing the classification, thus implicitly tracking the foreground regions. However, in case of a complex background, even if a GMM with a very high number of Gaussians is used, the foreground occupies background regions of similar color which become close to its position as the object moves along the scene.

The system that we propose follows the workflow used in other works like [11] or [14]. That is, the classification is made in a Bayesian framework, introducing a prior that contains neighborhood information. A graph cut is used to make the classification in this context. For every frame I_t, the foreground and background models are constructed. We propose to use the more complete GMM model in the joint color-space domain for the foreground regions, initialized with an initial foreground object mask [14]. But, in contrast with [11] and [14], we model the background with a pixel based model that allows a more precise description of it and it is computationally much less expensive to update. We thus combine a pixel-wise background model with a region based model. The models are used for the classification of the pixels of frame I_t, which is performed comparing the probabilities of foreground and background of every pixel within the graph cut algorithm. Based on the classification performed on the current frame, the models are updated. The EM algorithm is used for updating the foreground model and the Gaussian model of every pixel assigned to background is updated recursively. These updated models are then used for the classification of the next frame, I_{t+1}

The remainder of the paper is organized as follows. Section 2 describes the foreground probabilistic model and its update. The background model is explained in Section 3. Section 4 is devoted to the classification method. Finally, some results are presented in Sect. 5 and conclusions in Sect. 6.

2. FOREGROUND MODEL

A better classification of the pixels in foreground (fg) and background (bg) can be done if a probabilistic model for the foreground is also constructed. Since the foreground is constantly moving and changing, an accurate model at a pixel level is difficult to build and update. For this reason, we propose to use a Spatial Color Gaussian Mixture Model (SCGMM), as in [14], because foreground objects are better characterized by color and position, and GMM is a parametric model that describes accurately multi-modal probability density functions. Moreover, it can be easily estimated using an initialization frame. Thus, the foreground pixels are represented in a five dimensional space. The feature vector for pixel i, $zi \in \mathbb{R}^5$, is a joint domain-range representation, where the space of the image lattice is the domain, (x, y), and the color space, (r, g, b), is the range [11]. The likelihood of pixel i is then,

$$\begin{aligned} P(z_i|fg) &= \sum_{k=1}^{K_{fg}} \omega_k G_{fg}(z_i, \mu_k, \Sigma_k) \\ &= \sum_{k=1}^{K_{fg}} \omega_k \frac{1}{(2\pi)^{5/2}|\Sigma_k|^{1/2}} e^{-\frac{1}{2}(z_i-\mu_k)^T \Sigma_k^{-1}(z_i-\mu_k)} \end{aligned}$$

where ω_k is the mixture coefficient, μ_k and Σ_k are, respectively, the mean and covariance matrix of the k-th Gaussian distribution, $|\Sigma_k|$ is the determinant of matrix Σ_k. It is commonly assumed that the spatial and color components of the SCGMM models are decoupled, i.e., the covariance matrix of each Gaussian component takes the block diagonal form,

$$\Sigma_k = \begin{pmatrix} \Sigma_{k,s,} & 0 \\ 0 & \Sigma_{k,c} \end{pmatrix}$$

where s and c stand for the spatial and color features respectively. With such decomposition, each foreground Gaussian component has the following factorized form:

$$G_{fg}(z_i, \mu_k, \Sigma_k) = G(x_i, \mu_{k,s}, \Sigma_{k,s}) G(v_i, \mu_{k,c}, \Sigma_{k,c}), \quad (1)$$

where $x_i \in \mathbb{R}^2$ is the pixel's spatial information and $v_i \in \mathbb{R}^3$ is its color value. The parameter estimation can be reached via Bayes' development, with the EM algorithm [2]. For this estimation an initialization frame is needed, containing a first segmentation of the foreground object. This initialization can be performed with an exception to the background scheme.

Updating

While we assume a static background, the foreground objects usually perform a displacement within the scene. Thus, the spatial components of the Gaussian Mixture need to be updated after the classification in foreground and background of each frame. The pixels classified as foreground form a mask that is used for the updating. In order to avoid error propagation, as in [14], only the spatial components of the Gaussian Mixture are updated, and not the color ones. This updating is performed using an Expectation Conditional Maximization algorithm. In the E-step, the posteriori of the pixels belonging to each Gaussian component is computed, and in the M-step, the spatial mean, spatial variance, and mixture coefficient of each Gaussian component are refined based on the updated posteriori probability of pixels performed in the E-step.

3. BACKGROUND MODEL

For static backgrounds applications, a precise pixel model can be learnt. Although more complex models for each pixel

could be used, we propose to use a Gaussian distribution in the RGB color space [13] that has proved to work efficiently in most considered scenarios. We consider non-correlated components and the same variance for every color:

$$P(v_i|bg) = G(v_i, \mu_{i,c}, \sigma_{i,c}) = \frac{1}{(2\pi)^{3/2}\sigma_{i,c}^3} e^{-\frac{||v_i - \mu_{i,c}]||_2^2}{2\sigma_{i,c}^2}} \quad (2)$$

where $v_i \in \mathbb{R}^3$ is the i-th input pixel's value ($i = 1, ..., N$) in the RGB space, $\mu_{i,c} \in \mathbb{R}^3$ is the pixel mean value, $||.||_2$ denotes the Euclidean distance, and $\sigma_{i,c} \in \mathbb{R}$ is the color variance. We first initialize each background Gaussian ($\mu_{i,c}$ and $\sigma_{i,c}$) with initial training values learned from a set of frames with no foreground.

Since we want to combine the range background model with the joint range-domain foreground model we need to extend the pixel-based model (2) to a five dimensional model by using a SCGMM, analogously to the foreground model. For that, we use a mixture of N five dimensional gaussians, one representing each pixel in the image and thus having equal mixture proportions,

$$P(zi|bg) = \sum_{k=1}^{N} \frac{1}{N} G_{bg}(z_i, \mu_k, \sigma_k),$$

where

$$\begin{aligned} G_{bg}(z_i, \mu_k, \sigma_k) &= \delta(x_i - \mu_{k,s}) G(v_i, \mu_{k,c}, \sigma_{k,c}) \\ &= \delta(x_i - \mu_{k,s}) P(v_i|bg). \end{aligned}$$

Thus, we are using N gaussians, each one centered (in space) at each pixel position ($\mu_{k,s}$) with a zero spatial variance. This is sufficient for indoor scenarios with a static camera, although a small spatial variance can be used in order to allow for small outdoor background motions or camera shaking.

Updating

When a pixel value is classified as background, its model is updated following the Running Gaussian average model [13], in order to adapt it to progressive image variations. The update for a pixel i classified as background at frame t, $v_{i,t}$, is

$$\begin{aligned} \mu_{i,c,t} &= (1-\rho)\mu_{i,c,t-1} + \rho v_{i,t} \\ \sigma_{i,c,t}^2 &= (1-\rho)\sigma_{i,c,t-1}^2 + \rho(v_{i,t} - \mu_{i,c,t})^2 \end{aligned}$$

where ρ is the update rate (typically we use $\rho = 0.01$).

4. CLASSIFICATION

Once the foreground and background models have been computed, at frame t, the labeling can be done, assuming that we have some knowledge of foreground and background prior probabilities, $P(fg)$ and $P(bg)$ respectively, using a Maximum A Posteriori (MAP) decision. The priors can be approximated by using the foreground and background areas in the previous frame, $t-1$,

$$P(fg) = \frac{Area_{fg}|_{t-1}}{N}; \quad P(bg) = \frac{Area_{bg}|_{t-1}}{N}.$$

A pixel i is assigned to the class $c_i = \{fg, bg\}$ that maximizes $P(c_i|z_i) \propto P(z_i|c_i)P(c_i)$ (since $P(z_i)$ is the same for both classes and thus can be disregarded). However, analogously to [11], [14], we choose to consider the spatial context also for taking the segmentation decisions, instead of making an individual classification of the pixels. We consider for this aim a MAP-MRF framework in order to take into account neighborhood information. If we denote by c the labeling of all pixels of the image: $c = \{c_1, c_2, ..., c_N\}$, and by Nb_i the four connected neighborhood of pixel i, then:

$$P(c|z) \propto \prod_{i=1}^{N} P(z_i|c_i)P(c_i) e^{\sum_{i=1}^{N}\sum_{j \in Nb_i} \lambda(c_i c_j + (1-c_i)(1-c_j))}$$

Taking logarithms in the above expression leads to an standard form of the energy function that can be solved for global optimum using a standard graph-cut algorithm [1].

5. RESULTS

Tests have been performed in order to compare the proposed method with a pixel based background segmentation method and a region based foreground and background segmentation. The pixel based method used for the comparison is the state of the art Stauffer and Grimson method [12] and the region based method is based on [14]. In figures 1 and 2, sequences that performed poorly using pixel-based methods [12] have been selected. In particular, two sequences are shown in this paper where the colors of the foreground objects are in the same range than a part of the background. This generates many misses in the foreground detection when only the background model is used. Results improved when using a region based model [14]. However, a high computational load is required and some errors still appear. These are normally due to the poor modeling of the background due to the limited number of gaussians used (in figure 1: twenty gaussians for foreground and fourty gaussians for background; in figure 2: ten gaussians for foreground and twenty gaussians for background). These areas of the background poorly modeled are eventually captured by the foreground gaussians. In the method proposed the computational load is substantially reduced (diveded by a factor of seven in our tests) and less errors can be observed. Foreground regions in this method are represented by twenty gaussians in figures 1 and 2. Note that no filtering is applied on the resulting masks of each method, and the errors obtained could be easily avoided adding a filtering as a final step. The complete sequences are available in our web page *http://gps-tsc.upc.es/imatge/_Jgallego/icip09_results/* .

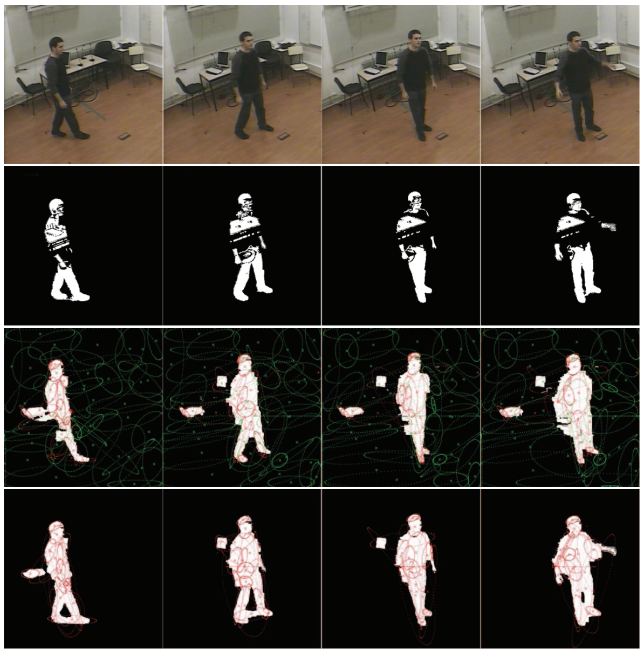

Fig. 1. *First sequence. Rows, from top to bottom, are: original frames, pixel-based foreground detection [12], region-based detection [14], combined pixel/region-based detection.*

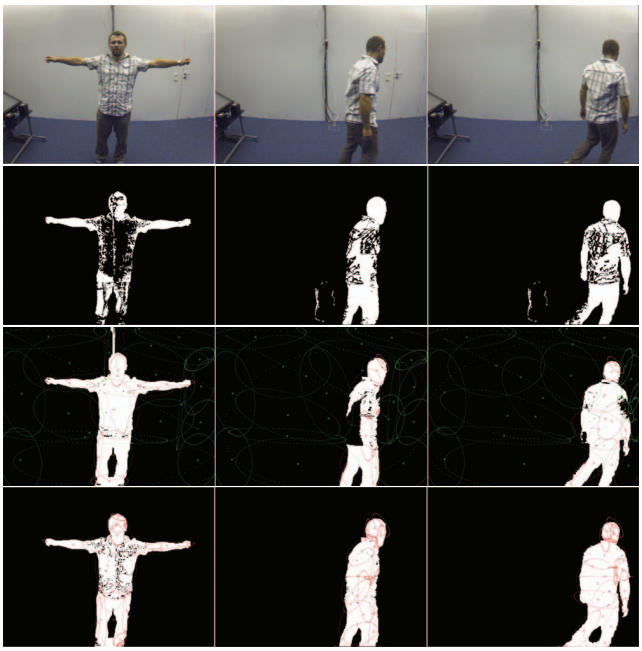

Fig. 2. *Second sequence. Rows, from top to bottom, are: original frames, pixel-based foreground detection [12], region-based detection [14], combined pixel/region-based detection.*

About the computational cost, analyzing an input video sequence of 400x400 pixels with one object in scene and using an Intel Xeon X5450 3.0GHz processor, our system allows a speed of 12 seconds/frame, while the pixel-based foreground detection [12] and region-based detection [14] allow a speed of 0.2 seconds/frame and 40 seconds/frame respectively. Note that no computational optimization has been taken in any of these foreground segmentation methods.

6. CONCLUSIONS

We have introduced a method for foreground segmentation and tracking which combines simple pixel-wise model for the background with a region based model for the foreground. Compared to region based methods for both background and foreground and to pixel-wise exception to background methods, our method performs better in situations of similar colors in foreground and background. In future work we will consider an updating of the models before decision, robust updating of the color components of the foreground model and algorithm simplifications to reduce the computational cost of the system.

7. REFERENCES

[1] Y. Boykov, O. Veksler, and R. Zabih, Fast Approximate Energy Minimization via Graph Cuts, IEEE Trans. on Pattern Analysis and Machine Intelligence, 23(11), 2001.

[2] A. P. Dempster, N. M. Laird, and D. B. Rubin, Maximum likelihood from incomplete data via the EM algorithm, Journal of the Royal Statistical Society, 39(1), 1977.

[3] A. Elgammal, R. Duraiswami, D. Harwood, and L.S. Davis, Background and foreground modeling using nonparametric Kernel density estimation for visual surveillance, Proc. of the IEEE, 90(7), 2002.

[4] A. Elgammal, D. Harwood, and L.S. Davis, Nonparametric model for background subtraction. In IEEE ICCV Frame-rate workshop, 1999.

[5] I. Haritaoglu, D. Harwood, and L.S. Davis, W4: Real time surveillance of people and their activities, IEEE Trans. on Pattern Analysis and Machine Intelligence, 22(8), 2000.

[6] S. Jabri, Z. Duric, H. Wechsler, and A. Rosenfeld, Detection and location of people in video images using adaptive fusion of color and edge information, Proc. of ICPR'2000.

[7] S. Khan and M. Shah, Tracking people in presence of occlusion, Proc. of Asian Conference on Computer Vision, 2000.

[8] L. Li, W. Huang, I.Y. H. Gu, and Q. Tian, Statistical modeling of complex backgrounds for foreground object detection, IEEE Trans. on Image Processing, 13(11), 2004.

[9] S. J. McKenna, Y.Raja, and S. Gong, Tracking colour objects using adaptive mixture models, Image and Vision Computing, 17(3), 1999.

[10] A. Mittal and L.S. Davis, M2tracker: A multi-view approach to segmenting and tracking people in a cluttered scene using region-based stereo. In Proc. of the 7th European Conference on Computer Vision, 2002.

[11] Y. Sheikh and M. Shah, Bayesian Modeling of Dynamic Scenes for Object Detection, IEEE Trans. on Pattern Analysis and Machine Intelligence, 27(11), 2005.

[12] C. Stauffer and W.E.L. Grimson, Learning patterns of activity using real-time tracking, IEEE Trans. on Pattern Analysis and Machine Intelligence, 22(8), 2000.

[13] C. R. Wren, A. Azarbayejani, T. Darrell, and A. Pentland, Pfinder: Real-time tracking of the human body, IEEE Trans. on Pattern Analysis and Machine Intelligence, 19(7), 1997.

[14] T. Yu, C. Zhang, M. Cohen, Y. Rui, and Y. Wu, Monocular Video Foreground/Background Segmentation by Tracking Spatial-Color Gaussian Mixture Models, Proc. of the IEEE Workshop on Motion and Video Computing, 2007.

[15] Q. Zhou and J.K. Aggraval, Tracking and classifying moving objects from video, Proc. of 2nd IEEE Int. Workshop on Performance Evaluation of Tracking and Surveillance, 2001.