Robust Foreground Segmentation Using Improved Gaussian Mixture Model and Optical Flow

Hajer Fradi
EURECOM
Sophia Antipolis, France
Email: Hajer.Fradi@eurecom.fr

Jean-Luc Dugelay
EURECOM
Sophia Antipolis, France
Email: Jean-Luc.Dugelay@eurecom.fr

***Abstract*—In automatic video surveillance applications, one of the most popular topics consists of separating the moving objects from the static part of the scene. In this context, Gaussian Mixture Model (GMM) background subtraction has been widely employed. It is based on a probabilistic approach that achieves satisfactory performance thanks to its ability to handle complex background scenes. However, the background model estimation step is still problematic; the main difficulty is to decide which distributions of the mixture belong to the background.**
To achieve an improved overall performance, motion cue could provide a rich source of information about the scene. Therefore, in this paper, we propose a new approach based on incorporating an uniform motion model into GMM background subtraction. By considering these both cues, high accuracy of foreground segmentation is obtained. Our approach has been experimentally validated showing better segmentation performance by comparisons with other approaches published in the literature.

I. Introduction

Segmenting foreground objects from a video sequence is a fundamental step in many computer vision applications such as video understanding, video conferencing and traffic monitoring. In this context, there are many issues (such as illumination changes, reflections, shadows, objects that have been moved, sleeping foreground, and so on) that make obtaining high accuracy foreground segmentation difficult and still subject to errors.
To tackle these problems, intensive research has been conducted. Among the proposed methods, GMM based background subtraction includes some robustness against changes in the background. Nevertheless the popularity of this method, the background/foreground discrimination still leaves rooms for further improvements. Actually, the decision on which of the Gaussians are most likely belonging to the background is made based on selecting the ones having the most supporting evidence and the least variance, which is not always correct. This ambiguity left by GMM method makes the estimation of the background model still hard to be properly addressed. To overcome the mentioned shortcoming, we propose incorporating an uniform motion model into GMM background subtraction. Considering these both information over time into a single overall system has the potential to detect foreground objects more reliably.
The rest of the paper is organized as follows: Section II reviews relevant works to foreground segmentation. Section III describes our motivation for the proposed approach. Then, Section IV and Section V detail our approach. Section VI shows the experimental results to demonstrate the effectiveness of our method. Finally, we give a brief conclusion in Section VII.

II. Related works

The simplest method to handle the problem of segmenting foreground objects is based on computing the difference between consecutive frames in order to detect moving objects. This technique is referred as *temporal differencing* [1], it is useful only in dynamic environments because as soon as an object stopped moving, it will be considered as foreground entity. Also, it is unable to extract the complete shape of moving objects.

Another commonly used technique to achieve this segmentation is based on building a representation of the scene called *background model* and comparing each incoming frame to this model. This process is termed as *background subtraction* and it aims at separating the expected part of the scene from the moving objects. The technique of background subtraction is widely used in real-time video processing using stationary cameras. Especially in video surveillance, it is considered as the basic low-level operation since many techniques can be carried out after performing background subtraction.

In this context, there are many difficulties that decrease the reliability of foreground segmentation. Precisely, the task becomes more challenging to deal with videos containing complex background [2]. One central problem is the changes that occur in the background such as water waves, waving trees, sudden lights switch and illumination changes caused by time of day evolution. The pixels belonging to the background are classified wrongly as foreground pixels. There is another category of issues making the distinction between the foreground and the background difficult, for example, sleeping (or motionless) foreground may be labeled as a part of the background. Added to that, shadows and reflections have not to interfere with foreground entities.

In the literature, many methods have been proposed to perform background subtraction which is mainly oriented towards the segmentation of moving objects from a video sequence. At the beginning, most of these methods assumed that the background is static and is composed of stationary objects. This constraint generates erroneous segmentation in

many cases. As a result, an efficient solution to handle the background variations was required.
The simplest manner proposed to deal with these variations is to smooth the color of background pixels with a Kalman filter or with an Infinite Impulse Response [3], [4].
After that, a better solution was proposed to handle smooth variation in the background; it is based on using Gaussian function [5]. For each pixel, the parameters of the Gaussian (the mean and the variance) are learned from the color observations in several adjacent frames. Afterwards, for an input frame, each pixel deviating from the background model is classified as foreground entity.
Although this technique allows learning gradual background changes in the time; it is restricted to model only one background. To handle this limitation, Gaussians Mixture Model (GMM) has been proposed by Stauffer and Grimson [6]. It consists of modeling each pixel as a mixture of Gaussians and using an on-line approximation for updating the model. Thanks to its ability to model various background distributions, this method showed a substantial progress to handle complex scenes. Therefore, until nowadays, GMM based background subtraction is considered as a standard method and it has become the basis for a large number of related methods [7], [8], [9].

Another way to segment foreground objects is by computing the optical flow between adjacent frames in a video sequence. To address this problem, many approaches have been proposed in the literature. Among them, the most popular ones are block matching and differential methods. Block matching methods are based on establishing the correspondences between each two successive frames. Differential based methods attempt to calculate the motion between two frames taken at times t and $t+\delta t$. Depending on the additional constraint used to estimate the flow velocity, differential methods can be also categorized into two subgroups which are local and global methods.
The difficulty of these methods is that its performance is low in poor textured regions where objects are not clearly defined. In addition, another problem in using optical flow is caused by the constant brightness assumption. This latter makes the technique unable to handle any variation in the lighting conditions. As a result, applying only optical flow is not so much used for foreground segmentation.

III. Motivation

After the analysis of the above different techniques for foreground segmentation, we demonstrate that GMM based background subtraction is powerful method to handle complex background scenes compared to temporal differencing and optical flow methods. But, the estimation of the background model using GMM still leave some ambiguities. At the same time, motion cue can provide additional and important information about the scene structure. That is why, we consider the combination of motion information and GMM background subtraction as a promising research direction.

In the last decade (after the method conducted by Stauffer and Grimson about GMM for background subtraction), many extensions of GMM have been proposed. However, there have been only few works on using optical flow with GMM background subtraction. A brief review on these works will be discussed. First, Zhou and Zhang proposed a combination of background subtraction, optical flow and temporal differencing [10]. The method starts by extracting foreground objects using GMM background subtraction. Then, optical flow supported by frame differences is used as post-processing step. The fusion is based on considering foreground objects only that are in motion. Another way to use optical flow within background subtraction was presented by Cai et al. [11]. This paper used optical flow to indicate the movement of a pixel's neighborhood. Therefore, two events can be distinguished: covering and revealing events which simplifies the task of determining the background value at a pixel position. After that, Kermouche and Aouf proposed a new technique [12] by integrating flow information with RGB color information in the same feature vector to statistically estimate the background model. Recently, a new background subtraction has been proposed by Wolf and Jolion [13], it integrates a discrete approximative optical flow by spatio-temporal regularization.

These works represent preliminary tentative attempting to overcome the limitations of background subtraction by using optical flow. In the same trend, we propose a new approach using both of these cues: the appearance models and the motion. Unlike the previous works, our proposed method applied an improved GMM for the background subtraction. The difference between the original and the improved versions of GMM will be discussed in Section IV.
In addition, we deal with the problem of the fusion differently. Actually, the integration of both cues is based on the assumption that pixels moving together (with the same velocity and orientation of the optical flow) have to get the same label (foreground or background). For this purpose, a measure for uniformity of motion is defined in Section V . Then, the incorporation of these two cues is done by favoring similar labels for pixels moving together. Moreover, compared to the previous woks, the present study has the advantage of comparing the results to other approaches in the literature with quantitative evaluation using public dataset, see Section VI.

IV. Background subtraction by Gaussian Mixture Model

The most popular background subtraction algorithm is based on Gaussian mixture model proposed by Stauffer and Grimson [6]. This method uses a mixture of K Gaussian distributions to model the recent history $\{X_1, ..., X_t\}$ of each pixel. The probability of observing the current pixel value is defined by a sum of weighted Gaussian distributions :

$$P(X_t) = \sum_{i=1}^{K} w_{i,t} * \eta(X_t, \mu_{i,t}, \Sigma_{i,t}) \quad (1)$$

where K is the number of distributions (typically between 3 and 5), $w_{i,t}$, $\mu_{i,t}$ and $\Sigma_{i,t}$ are respectively, an estimate of the weight, the mean value and the covariance matrix of the

i^{th} Gaussian in the mixture at time t. And $\eta(X_t, \mu, \Sigma)$ is the Gaussian probability density function.
Then, incoming pixels are compared against the corresponding Gaussian mixture model in order to find a Gaussian within 2.5 standard deviations. If a matching is found, the mean and the variance of the matched Gaussian are updated accordingly. However, if there is no match, the least probable component of the mixture is replaced by a new one modeling the incoming pixel. The prior weights of the K distributions at time t are defined as follows:

$$w_{k,t} = (1 - \alpha)w_{k,t-1} + \alpha(M_{k,t}) \tag{2}$$

where α is a learning rate, and $M_{k,t}$ is equal to 1 for the matched model and equal to 0 for the remaining models.
The updated parameters of the distribution that matches the new observation are defined by:

$$\sigma_t^2 = (1 - \rho)\sigma_{t-1}^2 + \rho(X_t - \mu_t)^T(X_t - \mu_t) \tag{3}$$

$$\mu_t = (1 - \rho)\mu_{t-1} + \rho X_t \tag{4}$$

where σ_{t-1}, μ_{t-1} are the last mean and variance values of the matched Gaussian and X_t is the value of the new pixel. ρ is the second learning rate defined by:

$$\rho = \alpha(X_t|\mu_k, \sigma_k) \tag{5}$$

The last step aims at estimating the background model from the mixture. For this purpose, the algorithm assumes that Gaussian distributions having the most supporting evidence and the least variance are most likely produced by background processes. As a result, the Gaussians are ordered by w/σ and the first B of the ranked distributions whose accumulated weights exceed T are deemed to be the background:

$$B = \arg\min_b(\sum_{k=1}^{b} w_k > T) \tag{6}$$

where T is the minimum fraction of the background model.

The detailed adaptive modeling background is robust enough for illumination changes; it can also deal with the movement in the background due to its multimodality. Many improvements for this method can be found in the literature to solve different limitations. One of these limitations arises due to the use of fixed learning rate all the time. Therefore, the parameters stabilize slowly which leads to problems with the initialization. That is why, the original version of GMM background subtraction has been further enhanced to improve its learning rate. This modification was proposed by Kaewtrakulpong and Bowden [14]. For the initialization, they improved the slow learning problem by using online Expectation Maximization algorithm and switching to the L-recent window update equations in order to give priority over recent data. This makes the convergence on a stable background model faster and also the tracker adapted to changes in the environment.

Another limitation of GMM method is caused by using a fixed number K of components over the time. Also, this number remains the same for each pixel which is not optimal in terms of computational time and segmentation accuracy. To address this problem, Zivkovic [15] proposes to constantly update not only the parameters but also the number of components of the mixture for each pixel. Using the Dirichlet prior, an online algorithm estimates the parameters of the GMM and selects the appropriate number of Gaussians simultaneously. As a result, K is dynamically adapted to the multimodality of each pixel. This method is called improved adaptive Gaussian mixture model and it is developed with shadow detection [16] to remove moving shadow pixels upon pixels labeled as foreground. A pixel is detected as shadow if it is considered as darker version of the background. For this purpose, a threshold is used to specify the darkness of the shadow.

Even if these modifications proposed in [15] showed improvement comparing to the original algorithm, the separation between foreground and background distributions is still problematic. Actually, the distinction is based on selecting as background components the Gaussians that are more frequently matched. In other words, it assumes that the often occurring pixels are deemed to model the background, which it is not always the case. That is why; we propose combining the improved GMM background subtraction with a uniform motion model into a single framework. This observation leads to better segmentation of the scene into foreground and background entities.

V. Uniform motion estimation

The second cue of the proposed approach is motion information. It is obtained by computing the optical flow between consecutives frames, then a measure for uniformity of motion is applied. For optical flow computation, several algorithms exist in the literature. A popular one was proposed by Lucas and Kanade [17], but this algorithm is classified as a sparse method since it is more suitable to be applied to a subset of points.

For dense optical flow computation, the algorithm conducted by Horn and Schunck [18] was the first proposed, it is based on *the brightness constancy assumption*. Then, another dense method was proposed by Farneback [19], it consists of computing optical flow based on polynomial expansion. This method uses quadratic polynomial model to approximate each neighborhood of both frames. Then, it estimates displacement fields from the polynomial expansion coefficients by observing how an exact polynomial transforms under translation. Also, it uses Gaussian to smooth the neighboring displacements. The evaluation of this method shows good results in terms of accuracy and low computation burden. Therefore, for optical flow computation, we utilize this method.

Figure 1 shows the optical flow across two adjacent frames at times t and $t + 1$. For each point P located at the 2D image coordinate $\vec{x} = [x\ y]^T$, the dense optical flow field provides a motion vector which is expressed as $2D$ velocities $\vec{V} = [v_x\ v_y]^T$. From these $x-$ and $y-$ components of the 2D velocity field, the optical flow of each point P in the origin image can be also defined by its magnitude and its direction

as follows:

$$Optical\ Flow(P_{x,y,t}) = \begin{pmatrix} Magnitude\ (P_{x,y,t}) \\ Direction\ (P_{x,y,t}) \end{pmatrix} \quad (7)$$

(a) *First frame* (b) *Second frame* (c) *Optical flow*

Fig. 1. *Dense optical flow computation for two consecutive frames*

After computing the optical flow on the current frame, the detection of uniform motion is performed. It works as follows: only pixels having non-zero optical flow velocities are considered. For the remaining values in the magnitude of the motion, neighbor pixels having similar direction are grouped in the same component. Therefore, four directions corresponding to these quadrants $\{[-\pi/2, 0], [0, \pi/2], [\pi/2, \pi], [-\pi, -\pi/2]\}$ can be distinguished.

For each direction d, N_d connected components are obtained with different brightness values for the magnitude. To measure the uniformity of the motion inside each component, a mean motion value is computed. If we denote Ω_k one component of the current frame, where k varies from 1 to N (N is the total number of components expressed as: $N = \sum_{d=1}^{4} N_d$), the mean motion value inside Ω_k is defined as follow:

$$v_k = 1/p \sum_{i \in \Omega_k} v_i \quad (8)$$

where p is the total number of pixels inside Ω_k and v_i is the magnitude of the motion for a pixel i. The difference between each magnitude value and the mean value inside the component is used as an error measure:

$$\epsilon = v_i - v_k \quad (9)$$

Then, an adequate threshold is chosen empirically for measuring the uniformity of motion. Only pixels belonging to Ω_k and regarding this uniformity will be considered. After the distinction between the different new components $\Omega_k^{'}$ (with the same velocity and orientation), the label of each component $\Omega_k^{'}$ (whether it belongs to background BG or foreground FG process) is defined as follows:

$$label(\Omega_k^{'}) = \begin{cases} FG \text{ if } (\frac{\sum_{\forall P_i \in \Omega_k^{'}} E(P_i)}{M_k} \geq R) \\ BG \text{ otherwise} \end{cases} \quad (10)$$

where M_k is the total number of pixels inside $\Omega_k^{'}$, R is a ratio in the range of [0,1] and

$$E(P_i) = \begin{cases} 0 & \text{if } label(P_i) = BG \\ 1 & \text{if } label(P_i) = FG \end{cases}$$

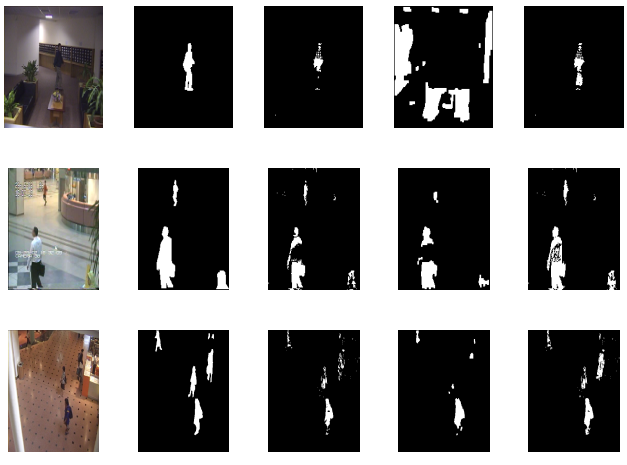
The goal of this integration is to improve the detection rate of GMM background subtraction without deteriorating the precision. Actually, pixels that belong to the background and undergo changes are correctly classified as background entities by GMM. However, these pixels are prone to be classified as foreground entities using optical flow. Therefore, we chose to start with the labels provided by GMM, then, by using the measure defined for uniform motion, the label of each pixel is updated. This integration adds spatial and temporal coherence since labeling process using GMM is done only at pixel level. It is an efficient way to improve the results and to avoid outliers caused by optical flow as well. Experimental results reported in the next section demonstrate the effectiveness of our proposed approach.

VI. Experimental Results

The proposed algorithm is compared with the improved adaptive GMM [15], and the foreground object detection method [20]. Unfortunately, most recent works cited in section III do not report results on public datasets.

To evaluate these methods, we used the i2r dataset (http://perception.i2r.a-star.edu.sg/bk_model/bk_index.html) with available ground truths. The dataset is composed of nine video sequences captured in challenging environments. For each sequence, ground-truth foreground masks are provided for 20 randomly selected frames.

Using this dataset, both of qualitative and quantitative analysis of the results are presented with comparisons to the already cited methods. Figure 2 shows results on three frames from different sequences. The results of the proposed method are qualitatively better than those obtained by the other methods.



(a) (b) (c) (d) (e)

Fig. 2. *Foreground segmentation results (a) Evaluation frames (b) Ground-truth foreground masks (c) Results of improved adaptive GMM [15] (d) Results of foreground object detection method [20] (e) Results of our proposed approach*

For quantitative evaluation, we use these metrics to compare the foreground mask to the ground truth:

Video sequences	Metrics	Improved adaptive GMM [15]	Foreground object detection method [20]	Our proposed approach
Restaurant	recall	55.09	48.48	63.41
	precision	99.61	97.38	99.03
Curtain	recall	38.92	41.66	70.62
	precision	99.95	99.89	99.23
Escalator	recall	71.65	40.02	74.36
	precision	98.75	98.31	98.24
Fountain	recall	44.42	40.40	54.87
	precision	99.28	99.41	99.23
Water Surface	recall	67.72	50.05	79.39
	precision	99.77	99.19	99.44
Trees	recall	73.99	63.49	88.14
	precision	97.45	99.63	97.19
Shopping center	recall	52.18	59.50	66.60
	precision	99.69	98.33	99.28
Lobby	recall	40.14	38.87	73.34
	precision	99.97	94.1	99.88
Hall	recall	39.10	47.37	63.18
	precision	99.71	99.08	99.27

TABLE I
QUANTITATIVE EVALUATION OF OUR PROPOSED APPROACH COMPARED TO OTHER METHODS

- Recall: It is called also True Positive Rate or Detection Rate defined by:

$$Recall = \frac{TP}{TP + FN} \tag{11}$$

- Precision: it corresponds to $1 - FAR$ where FAR is called False Acceptance Rate, it is defined by:

$$Precision = \frac{TN}{FP + TN} \tag{12}$$

where TP, TN, FP and FN denote respectively the total number of true positives, true negatives, false positives and false negatives.

Quantitative results using these metrics are reported in Table I. These results show that the improved GMM [15] reaches 99% for the precision (we compute the average of different results), however, the detection rate is neither sufficient nor satisfactory for many applications (only 53%). That is why, our proposed method showed substantial improvement over GMM by increasing the detection rate (by 17%) and the precision remains roughly the same (around 99%). Also, by means of comparison to the method proposed in [20], our method gives better results. For the detection rate, it achieved a noteworthy improvement of about 23% compared to [20]. For the precision, the method proposed in [20] achieved 98%.

From these comparisons, we conclude that our proposed method outperforms the other methods. In addition, as it is shown is Figure 2, our results are able to detect full object or in a shape that can be useful in many other applications. Since foreground segmentation is a key step in automatic video surveillance, the superior results that we obtained can deeply affect many applications such as people detection and tracking, person counting, and so on.

VII. CONCLUSION

In this paper a new approach is proposed for robust and online foreground segmentation using Gaussian Mixture Model and motion cue. The proposed approach succeeds to harness the advantages of both cues by improving the detection rate without deteriorating the precision. Our approach has been also tested on dataset of complex background scenes. The results demonstrate its ability to improve significantly the accuracy of the foreground segmentation compared to other existing approaches in the literature. The obtained superior results are significant since many applications can be carried out after performing reliable foreground segmentation.

REFERENCES

[1] R. Jain and H. H. Nagel. On the analysis of accumulative difference pictures from image sequences of real world scenes. *IEEE Trans. Pattern Analysis and Machine Intelligence*, 1(2):206–214, 1979.

[2] M. Cristani, M. Farenzena, D. Bloisi, and V. Murino. Background subtraction for automated multisensor surveillance: a comprehensive review. *EURASIP Journal on Advances in Signal Processing*, 2010, 2010.

[3] T. J. Ellis, P. L. Rosin, and P. Golton. Model-based vision for automatic alarm interpretation. *IEEE Aerospace and Electronic Systems Magazine*, 6(3):14–20, 1991.

[4] S. B. Gray. Local properties of binary images in two dimensions. *IEEE Trans. Computers*, 20(5):551–561, 1971.

[5] C. R. Wren, A. Azarbayejani, T. Darrell, and A. P. Pentland. Pfinder: real-time tracking of the human body. *IEEE Trans. Pattern Analysis and Machine Intelligence*, 19(7):780–785, 1997.

[6] C. Stauffer and W. E. L. Grimson. Adaptive background mixture models for real-time tracking. *Proc. of CVPR*, 246–252, 1999.

[7] Y. T. Chen, C. S. Chen, C. R. Huang, and Y. P. Hung. Efficient hierarchical method for background subtraction. *Pattern Recognition*, 40(10):2706–2715, 2007.

[8] E. Hayman and J. O. Eklundh. Statistical background subtraction for a mobile observer. *International Conference on Computer Vision*, 2003.

[9] Z. Zivkovic and F. van der Heijden. Efficient adaptive density estimation per image pixel for the task of background subtraction. *Pattern Recognition Letters*, 27(7):773–780, 2006.

[10] D. Zhou and H. Zhang. Modified GMM background modeling and optical flow for detection of moving objects. *International Conf. on Systems, Man and Cybernetics*, 3(5):2224–2229, 2005.

[11] X. Cai, F. H. Ali, and E. Stipidis. Robust online video background reconstruction using optical flow and pixel intensity distribution. *IEEE International Conference on Communications-Signal Processing for Communications Symposium*, 1–5, 2008.
[12] M. S. Kemouche and N. Aouf. A gaussian mixture based optical flow modeling for object detection. *ICDP*, 1–6, 2009.
[13] C. Wolf and J. M. Jolion. Integrating a discrete motion model into gmm based background subtraction. *ICPR*, 9–12, 2010.
[14] P. Kaewtrakulpong and R. Bowden. An improved adaptive background mixture model for realtime tracking with shadow detection. *2nd European Workshop on Advanced Video Based Surveillance Systems*, 2001.
[15] Z. Zivkovic. Improved adaptive gaussian mixture model for background subtraction. *ICPR*, 2:28–31, 2004.
[16] A. Prati, I. Mikic, M. Trivedi, and R. Cucchiara. Detecting moving shadows: Algorithms and evaluation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, (25):918–923, 2003.
[17] B. Lucas and T. Kanade. An iterative image registration technique with an application to stereo vision. *Proc. of the 7th International Joint Conference on Artificial Intelligence*, 674–679, 1981.
[18] B. K. P. Horn and B. G. Schunck. Determining optical flow. *Artificial Intelligence*, 17:185–203, 1981.
[19] G. Farneback. Two-frame motion estimation based on polynomial expansion. *Proc. of 13th Scandinavian Conference on Image Analysis*, 363–370, 2003.
[20] W. Huang, I. Y. H. Gu, and Q. Tian. Foreground object detection from videos containing complex background. *ACM MM*, 2003.