Robust Background Modeling and Foreground Detection using Dynamic Textures

M. Sami Zitouni, Harish Bhaskar and Mohammed Al-Mualla

Department of Electrical and Computer Engineering, Khalifa University of Science, Technology and Research, Abu Dhabi, U.A.E.

Keywords: Background Modeling, Foreground Detection, Dynamic Texture, Gaussian Mixture Model.

Abstract: In this paper, a dynamic background modeling and hence foreground detection technique using a Gaussian Mixture Model (GMM) of spatio-temporal patches of dynamic texture (DT) is proposed. Existing methods for background modeling cannot adequately distinguish movements in both background and foreground, that usually characterizes any dynamic scene. Therefore, in most of these methods, the separation of the background from foreground requires precise tuning of parameters or an apriori model of the foreground. The proposed method aims to differentiate between global from local motion by attributing the video using spatio-temporal patches of DT modeled using a typical GMM framework. In addition to alleviating the aforementioned limitations, the proposed method can cope with complex dynamic scenes without the need for training or parameter tuning. Qualitative and quantitative analysis of the method compared against competing baselines have demonstrated the superiority of the method and the robustness against dynamic variations in the background.

1 INTRODUCTION

Background modeling and hence foreground detection are essentials steps in visual surveillance; particularly for moving object detection and target tracking. Conventional background modeling techniques such as (Stauffer and Grimson, 1999; Stauffer and Grimson, 2000; Bhaskar et al., 2007) assume limited changes in the background, making them unsuitable for capturing the dynamics in the environment caused either due to background movements or motion of the sensor. In addition, background modeling is complicated by the motion dynamics of moving targets, for example stoppages during motion, appearance changes of targets, lighting variations, noise and clutter, that are typical in real-world outdoor scenarios.

1.1 Related Work

A considerable amount of effort has been devoted for developing adaptive background modeling methods, exemplified in the work of (Stauffer and Grimson, 1999) using GMM, and its extensions exploiting various properties such as global consistency (Dalley et al., 2008), local image neighborhoods (Heikkila and Pietikainen, 2006), and density clustering (Bhaskar et al., 2007). Additionally, for handling dynamic background motion, (Zhong et al., 2009) proposed a background subtraction technique based on GMM using a multi-resolution framework, while (Zhang et al., 2009) proposed a spatial-temporal nonparametric background subtraction approach. Furthermore, (Zhang et al., 2009) proposed using an adaptive Local-Patch GMM as the dynamic background model with Support Vector Machiine (SVM) classification for shadow removal applications.

Despite advances, several issues concerning dynamic background continue to remain as challenges to the background modeling community. In recent years, saliency detection in motion and appearance of objects has contributed to significant progress in handling such inadequacies in background modeling. For example, it has been shown in (Tian and Hampapur, 2008) that accurate foreground detection can be facilitated by distinguishing salient motion from background motion. Similarly, the integration of background learning and object detection through Detecting Contiguous Outliers in Low-rank Representation(DECOLOR) in (Zhou et al., 2013) has shown to accommodate global variations. Specif-

Zitouni, M., Bhaskar, H. and Al-Mualla, M.
Robust Background Modeling and Foreground Detection using Dynamic Textures.
DOI: 10.5220/0005724204030410
In *Proceedings of the 11th Joint Conference on Computer Vision, Imaging and Computer Graphics Theory and Applications (VISIGRAPP 2016) - Volume 4: VISAPP*, pages 403-410
ISBN: 978-989-758-175-5

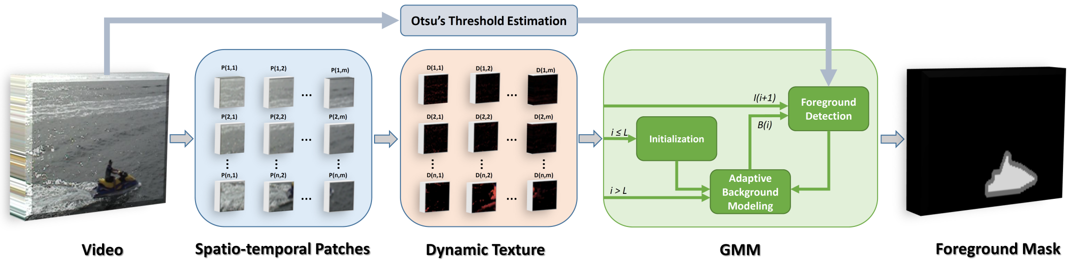


Figure 1: A brief block diagram of the proposed spatio-temporal Gaussian Mixture Model of Dynamic Texture.

ically, background modeling under dynamic background variations has been addressed in the work of DT in (Doretto et al., 2003), where motion is modeled as a linear dynamical system. Several variations of the DT model as in (Mumtaz et al., 2014; Chan et al., 2011; Chan and Vasconcelos, 2009; Zhong and Sclaroff, 2003) have gained recognition within this context, and in particular the work of (Chan et al., 2011) has shown how a generalized formulation of the (Stauffer and Grimson, 1999) algorithm can be accomplished using mixture components of DT with an online learning algorithm. Further, the use of DT along with Kalman filter has been proposed in (Zhong and Sclaroff, 2003) for foreground detection. A layered implementation in (Chan and Vasconcelos, 2009) has been used to model a video represented as stochastic layers using appearance and dynamics, each modeled by a separate DT. In (Chan and Vasconcelos, 2009), it has been demonstrated that over-segmentation can occur in mixed dynamic backgrounds, as each layer's segment corresponds to single motion. Similarly in the work of (Mumtaz et al., 2014), DT is used to jointly model both the foreground and the background. Such techniques have been utilized for modeling a video in (Zhong and Sclaroff, 2003) considering an image frame as a whole, or as spatial patches extracted from the video (Mumtaz et al., 2014; Chan et al., 2011). However, a majority of these techniques for foreground detection require scene-specific parameterization in addition to apriori training before classification.

1.2 Novelty & Contributions

Modeling using mixtures of DT as in (Mumtaz et al., 2014), forces the assumption that the motion encapsulated by the DT is an inherent representation of the background model. However, this limits the method to those dynamic scenes where the DT is a strong representation of the background. For example, in sequences, where the DT is a stronger cue for foreground motion, detection fails. The novelty of the proposed approach is the definition of a GMM of DT that is capable of providing a generic formulation for modeling dynamic motion either as a background or as a foreground. In addition, the proposed method is implemented as a classification of spatio-temporal patches of DT using GMM in a manner that it preserves spatio-temporal homogeneity, ensuring smoother boundaries avoiding under-or-over-segmentation and providing computational advantages. Furthermore, the treatment of DT as a feature space allows reducing the effect of noise and illumination without the need for training or tuning of parameters. Finally, the paper presents multi-resolution analysis on the spatio-temporal patches, exploiting its effect on accuracy and its relationship with the learning rate of the GMM scheme.

2 SPATIO-TEMPORAL GMM OF DT

In this section, the spatio-temporal GMM of DT method is proposed and formulated as a stochastic model using probability density function (PDF) corresponding to the foreground and background. The block diagram in Fig. 1 illustrates the process flow of the foreground detection model proposed in this paper.

According to Fig. 1, the detection process begins by splitting video frames to spatio-temporal texture patches using DT. Further, a decision on whether each patch represents either the foreground or background is formulated using a conventional GMM according to (Stauffer and Grimson, 1999).

2.1 Spatio-temporal Representation

The first step in this approach is to reduce the dimensionality of the video analysis into spatio-temporal blocks. The process begins with the original non-processed video, that is treated as a three dimensional (3D) array of gray pixels $\mathbf{I}_{x,y,t}$ consisting of two spatial dimensions (x,y) and one temporal dimension

t. The spatio-temporal blocks are represented as N-dimensional vectors $\mathbf{b}_{X,Y,t}$, where each block spans $(2T+1)$ frames and contains N_b pixels in each spatial direction per frame, hence producing $N = (2T+1)$ x N_b x N_b. Here block vectors $\mathbf{b}_{X,Y,t}$ can be formally defined according to (Pokrajac and Latecki, 2003) as,

$$\mathbf{b}_{X,Y,t} = [\mathbf{I}_{x,y,t}]_{i=(N_b-1)(X-1)+1,j=(N_b-1)(Y-1)+1,t=t-T}^{i=N_bX,j=N_bY,t=t+T} \quad (1)$$

The key advantage of such a representation is the flexibility to exploit the square of linear block sizes of the vectors to reduce dimension in such a manner that maximum information can be preserved. Dimensionality reduction is typically practiced using the Principal Component Analysis (PCA). However, in this paper, this reduction is imposed during the computation of DT which facilitates estimating low-level appearance and motion features locally and thus allowing to study their impact on global background estimation.

2.2 Dynamic Textures

The model of DT can be written as a linear dynamical system and generated for an image sequence or in this case, for each spatio-temporal patch. Let the array of gray pixels $\mathbf{I}_t$ be decomposed into q number of blocks $\mathbf{b}_t^i$, where $0 \leq i \leq q$. The linear system contains two stochastic processes, the dynamics as state process evolve over time $\mathbf{b}_t^i \in R^n$, and the corresponding appearance $\mathbf{d}_t^i \in R^m$ as a function of current state process and observation noise. The system is defined by:

$$\begin{cases} \mathbf{b}_t^i = A\mathbf{b}_{t-1}^i + \nu_t \\ \mathbf{d}_t^i = C\mathbf{b}_t^i + \omega_t \end{cases} \quad (2)$$

where $A \in R^{n\times n}$ and $C \in R^{m\times n}$ are the state transition matrix and the observation matrix respectively. The state is modeled as Gaussian process $\nu_t \sim N(0,Q)$ as well as the observation noise $\omega_t \sim N(0,R)$. According to (Doretto et al., 2003), the system parameters are calculated with least squares algorithm. Given a spatio-temporal patch, for example, $D_{1:\tau}^i = [d_1^i, \ldots, d_\tau^i]$, the estimated temporal mean of the patch:

$$\bar{d^i} = \frac{1}{\tau}\sum_{t=1}^{\tau} d_t^i \quad (3)$$

which is used to get the mean subtracted sequence:

$$\tilde{D}_{1:\tau}^i = D_{1:\tau}^i - \bar{D}^i = [\tilde{d}_1^i, \ldots, \tilde{d}_\tau^i] \quad (4)$$

where $\bar{D}^i$ is a matrix with τ replications of mean $\bar{d^i}$. For the parameters estimation, singular value decomposition (SVD) is performed on the mean subtracted sequence

$$\tilde{D}_{1:\tau}^i = U^i S^i V^{i'} \quad (5)$$

The n principal components of the largest eigenvalues of V are used to estimate the observation matrix, assuming that diagonal entries of S are ordered in decreasing value. Then $\hat{C}^i = [u_1, \ldots, u_n]$ and the state space is estimated as:

$$\hat{B}_{1:\tau}^i = \hat{C}^{i'}\tilde{D}_{1:\tau}^i = [\hat{b}_1^i, \ldots, \hat{b}_\tau^i] \quad (6)$$

The initial state of the block b_t^i is assumed to be b_1^i. Then, the least square estimation of the transition matrix A is calculated with:

$$\hat{A}^i = \hat{B}_{2:\tau}^i(\hat{B}_{1:\tau-1}^i)^+ \quad (7)$$

given that the Moore-Penrose pseudoinverse of B is $B^+ = B^{'}(BB^{'})^{-1}$. The state space prediction error is used to estimate the state noise:

$$\hat{V}_{1:\tau-1}^i = \hat{B}_{2:\tau}^i - \hat{A}^i\hat{B}_{2:\tau-1}^i \quad (8)$$

$$\hat{Q}^i = \frac{1}{\tau-1}\hat{V}_{1:\tau-1}^i(\hat{V}_{1:\tau-1}^i)^{'} \quad (9)$$

As well, the reconstruction error is used to estimate the observation noise:

$$\hat{W}_{1:\tau}^i = D_{1:\tau}^i - \hat{C}^i\hat{B}_{1:\tau}^i \quad (10)$$

$$\hat{R}^i = \frac{1}{\tau}\hat{W}_{1:\tau}^i(\hat{W}_{1:\tau}^i)^{'} \quad (11)$$

This suboptimal approach for LDS parameter estimation of DT is done q times for the concatenation of each spatial patch.

2.3 Gaussian Mixture Model Framework

It has been shown that an input frame $\mathbf{I}_t$ at time instant t of a given video sequence is decomposed into q number of patches $\mathbf{d}_t^i$, where $0 \leq i \leq q$, using a DT algorithm aforementioned. Henceforth, the decision of whether each texture patch $\mathbf{d}_t^i$ represents a foreground (FG) or a background (BG) can be formulated as the ratio of pdf's as below,

$$\frac{p(BG|\mathbf{d}_t^i)}{p(FG|\mathbf{d}_t^i)} = \frac{p(\mathbf{d}_t^i|BG)p(BG)}{p(\mathbf{d}_t^i|FG)p(FG)} \quad (12)$$

where $\mathbf{d}_t^i = \{d_{1,t}^i, \ldots, d_{N_b,t}^i\}$ characterizes a DT patch $\mathbf{d}^i$ consisting of N_b number of pixels such that the image frame at time t is represented as $\mathbf{I}_t = \bigcup_{i=1}^q \mathbf{d}_t^i = \{\mathbf{d}_t^1, \ldots, \mathbf{d}_t^q\}$. While $p(BG|d_t^i)$ represents the pdf of the background modelled using the DT on patch $\mathbf{d}_t^i$, $p(FG|d_t^i)$ is the pdf of the foreground representing the same DT patch $\mathbf{d}_t^i$. Here, $p(d_t^i|BG)$ denotes the background model whereas $p(d_t^i|FG)$ is the appearance model of the foreground object. The decision of whether any of the DT patches $\mathbf{d}_t^i$ represents the background is according to:

$$p(\mathbf{d}_t^i|BG) > \frac{p(\mathbf{d}_t^i|FG)p(FG)}{p(BG)} \tag{13}$$

The background and foreground in the input video are modeled with the GMM framework, where the DT of the patches are used as a feature for the GMM model with K Gaussian distributions. The probability of a certain DT patch $\mathbf{d}_t^i$ at time t is represented as:

$$p(\mathbf{d}_t^i) = \sum_{k=1}^{K} w_k \mathcal{N}(\mathbf{d}_t^i;\mu_k,\Sigma_k) \tag{14}$$

where w_k is the weight of the k^{th} Gaussian component, and $\mathcal{N}(\mathbf{d}_t^i;\mu_k,\Sigma_k)$ is the Normal distribution of the k^{th} component given as:

$$\mathcal{N}(\mathbf{d}_t^i;\mu_k,\Sigma_k) = \frac{1}{|2\pi\Sigma_k|^{1/2}} e^{-1/2(\mathbf{d}_t^i-\mu_k)^T\Sigma_k^{-1}(\mathbf{d}_t^i-\mu_k)} \tag{15}$$

where μ_k is the mean and Σ_k is the co-variance of the k^{th} component. The distributions are ordered according to the value of w_k/Σ_k, and the first Bg distributions are used to initialize the background model of the scene and it is estimated by:

$$Bg = argmin_b(\sum_{l=1}^{b} w_l > T) \tag{16}$$

The decision threshold T is the minimum prior probability of the background, and in this method its value is obtained using Otsu's method on the mean of a gray version of the input video frames, adding the texture features that contributes in distinguishing the background. The Gaussian components initializing the background model, are updated through an adaptive learning procedure using:

$$\hat{w}_k^{t+1} = (1-\alpha)\hat{w}_k^t + \alpha\hat{p}(\omega_k|\mathbf{d}_{t+1}^i) \tag{17}$$

$$\hat{\mu}_k^{t+1} = (1-\rho)\hat{\mu}_k^t + \rho\mathbf{d}_{t+1}^i \tag{18}$$

$$\hat{\Sigma}_k^{t+1} = (1-\rho)\hat{\Sigma}_k^t + \rho(\mathbf{d}_{t+1}^i - \hat{\mu}_k^{t+1})(\mathbf{d}_{t+1}^i - \hat{\mu}_k^{t+1})' \tag{19}$$

$$\rho = \alpha\mathcal{N}(\mathbf{d}_{t+1}^i;\mu_k^t,\Sigma_k^t) \tag{20}$$

where ω_k is the k^{th} Gaussian component, and $\hat{p}(\omega_k|\mathbf{d}_{t+1}^i)$ is either 1 when ω_k is the first match and 0 otherwise. The value of α determines the learning rate corresponding to the changes in the texture patches.

3 EXPERIMENTS

In this section, experiments conducted to validate the performance of the proposed model and benchmark it against competing baseline methods, are described. The chosen dataset (FBDynSyn from (Mumtaz et al., 2014)) includes video sequences that encapsulate all the challenges of dynamic background variations, and in addition an even more challenging video sequence dubbed as Sailing2 from (Chan et al., 2011) demonstrating complex motion of a boat on water has also been used for validation. The FBDynSyn dataset consists of 7 video sequences with dynamic backgrounds such as water, fountains, and trees and moving targets of interest such as people and boats in the foreground. The video sequence contained (210-601) number of frames at a resolutions in the range of (120x190) - (300x600).

All sequences are annotated with ground truth, which is further used for the qualitative and quantitative evaluations using a range of metrics including True Positive Rate (TPR), False Positive Rate (FPR), Accuracy (ACC), Variation of information (VI) and the Rand Index (RI). All experiments were conducted using MATLAB on an Intel i5 2.6 GHz processor machine with 8 GB RAM. During evaluation, each video frame has been segmented into patches of size 10 x 10 x 3 and extracted the dynamic texture components for each patch. The learning rate of the GMM model for all video sequences was fixed at 0.0001. The (Otsu, 1979) method has been incorporated to automatically estimate the decision threshold within the proposed detection framework. Finally, the proposed method is compared against state-of-the-art methods of (Mumtaz et al., 2014), (Zhou et al., 2013) and (Chan and Vasconcelos, 2009) both quantitatively and qualitatively as described in (Mumtaz et al., 2014) using the same dataset.

In Fig. 2. example frames from each video sequence comparing the proposed method to the ground truth and the results of the state-of-the-art methods for qualitative evaluation through visual inspection are presented. It is clear that the results of detection using the proposed method is more accurate and precise than other methods and moreover, closely matches with its ground-truth counterpart. Table. 1 summarizes the performance comparison of the proposed method against baseline methods (Mumtaz et al., 2014) and (Zhou et al., 2013) using the TPR and FPR metrics estimated at the operating point (OP) using the receiver operating characteristics (ROC) curves. The TPR achieved by the proposed method in all videos is higher than both (Mumtaz et al., 2014) and (Zhou et al., 2013) methods with

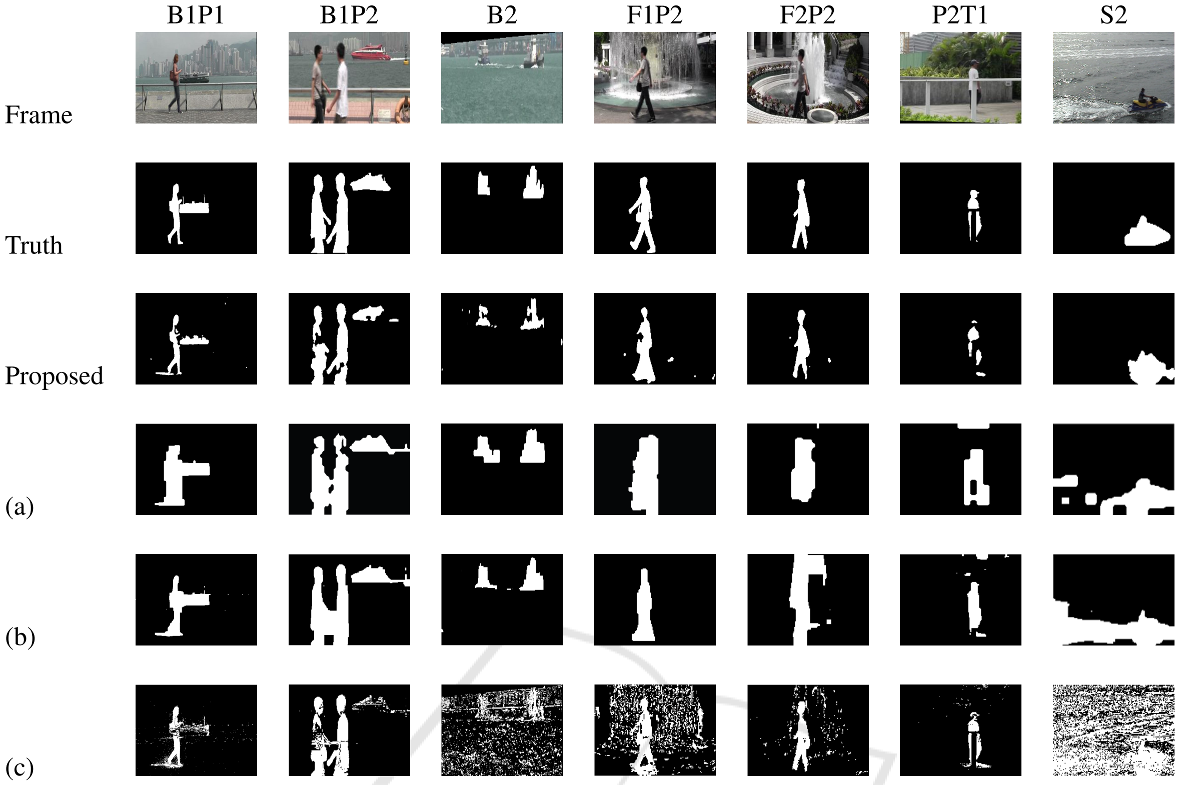


Figure 2: Results of background segmentation of selected a frame from different sequences (across columns) using proposed method (row 3), compared to the ground-truth (row 2) and other baseline methods including (a) (Mumtaz et al., 2014), (b) (Zhou et al., 2013) and (c) (Stauffer and Grimson, 1999).

Table 1: Quantitative Evaluation of proposed method against state of art algorithms (a) (Mumtaz et al., 2014), (b) (Zhou et al., 2013) at their operating point.

Data	TPR			FPR			
	Proposed	(a) at OP	(b)	Proposed	(a)	(a) at OP	(b)
B1P1	0.997	0.973	0.967	0.003	0.004	0.019	0.007
B1P2	0.989	0.919	0.977	0.010	0.009	0.022	0.018
Boat2	0.996	0.955	0.931	0.004	0.004	0.022	0.008
F1P2	0.988	0.972	0.791	0.011	0.034	0.055	0.007
F2P2	0.995	0.892	0.946	0.005	0.064	0.038	0.086
P2T1	0.992	0.953	0.967	0.008	0.030	0.056	0.017
S2	0.976	0.968	0.947	0.023	0.016	0.040	0.164
Average	0.990	0.947	0.932	0.009	0.023	0.036	0.044

a comparative average of 0.990 against 0.947 and 0.932. As well, the measured FPR of the proposed method is lower than that for (Mumtaz et al., 2014) at OP and (Zhou et al., 2013) with an average of 0.009 as against 0.036 and 0.044. However, (Mumtaz et al., 2014) at lower TPR has a similar FPR for Boat1Person1, Boat1Person2, and Boats2 videos, and a lower one at Sailing2 (0.023 vs 0.016). Further, the proposed method is compared to both (Mumtaz et al., 2014) and (Zhou et al., 2013) using the StopPerson1 video sequence. The scenario in this sequence has the target-of-interest (person) stopping for a short duration of time, making it extremely challenging for conventional background modeling techniques to cope with.

In Table 2, the comparison of the proposed method to these baseline method on the StopPerson1 sequence is listed. The results on this sequence indicates the superiority of the proposed technique both in terms of the higher TPR and lower FPR against both (Mumtaz et al., 2014) and (Zhou et al., 2013). The qualitative comparison of results between the

Table 2: Evaluation in stop case video StopPerson1 against (Mumtaz et al., 2014) and (Zhou et al., 2013) methods at their operating points.

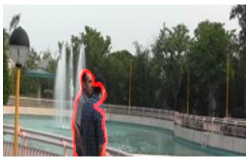
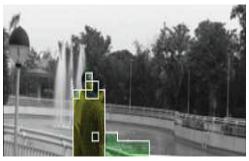
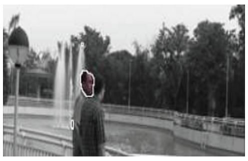
	Proposed	(Mumtaz et al., 2014)	(Zhou et al., 2013)
TPR	0.996	0.945	0.642
FPR	0.004	0.026	0.003
Example			

Table 3: Evaluation of proposed method against state of art (a) (Mumtaz et al., 2014), (b) (Zhou et al., 2013) using Rand Index (RI).

	B1P1	B1P2	B2	F1P2	F2P2	P2T1	SP1	Avg.	S2
Proposed	0.9681	0.9465	0.9589	0.9535	0.9707	0.9825	0.9612	0.9631	0.9314
(a)	0.9632	0.9428	0.9610	0.9156	0.9388	0.9270	0.9482	0.9424	
(b)	0.9524	0.7021	0.7986	0.7769	0.3833	0.8646	0.8668	0.7635	

compared techniques supports the claim of the quantitative results. The superiority in performance can be mainly attributed to the more robust background model built by the proposed algorithm, while in case of (Mumtaz et al., 2014) leads to over-segmentation, as against (Zhou et al., 2013) that under-segments. For further evaluation of the motion segmentation, RI is calculated comparing the proposed method against (Mumtaz et al., 2014) and (Chan and Vasconcelos, 2009) methods in Table. 3. In all videos, the proposed method outperforms LDT at an average of (0.9631 vs 0.7635). In the cases of Boat1Person1(B1P1) and Boat1Person2(B1P2), the proposed method achieved similar or slightly higher RI, and slightly lower in Boat2(B2) compared to (Mumtaz et al., 2014) as some parts of the target are lost due to smoothing throughout the modeling process. In the other cases as in Fountain1Person2(F1P2), Fountain2Person2(F2P2) and StopPerson1 (SP1) sequences, where the background is more complex, the proposed method outperforms (Mumtaz et al., 2014). On average, the performance of the proposed method is higher than (Mumtaz et al., 2014) with a rand index of 0.9631 vs 0.9424. Moreover, in Table. 4, additional quantitative results using precision, ACC, and VI indices are presented. The average precision, ACC, and VI indices achieved are 0.8446, 0.9774 and 0.2107 respectively.

The results described in this section so far, considers DT constituting a majority of the background region as in FBDynSyn dataset. However, without loss of generality, it is equally possible that DT can conveniently represent foreground regions as in the case of crowd motion. In Fig. 3, a plot showing the variations of the DT values of a chosen block from two sequences, one where DT represents the background (Boat1Person1 sequence) and the other where DT represents the foregroun (S1_L1_13-57 PETS crowd sequence) from (Ferryman and Shahrokni, 2009), is illustrated. It can be observed that the variations in DT for the Boat1Person1 sequence represented using red dashed lines, matches the characteristics of the DT changes in the crowd sequence represented as yellow solid line.

Table 4: Quantitative results of proposed technique using precision, accuracy(ACC) and variation of information (VI).

Data	Precision	ACC	VI
Boat1Person1	0.8823	0.9837	0.1628
Boat1Person2	0.8982	0.9695	0.2659
Boat2	0.8311	0.9789	0.2046
Fountain1Person2	0.8187	0.9724	0.2464
Fountain2Person2	0.8977	0.9818	0.1673
Person2Tree1	0.7966	0.9867	0.1070
StopPerson1	0.9495	0.9849	0.1828
Sailing2	0.6850	0.9614	0.3488
Average	0.8446	0.9774	0.2107

Finally, further validation has been done by testing the proposed method on additional dynamic background sequences from changedetection dataset of (Goyette et al., 2012), that also includes the boats, fall, fountain01 and fountain02 sequences. In order to prove beyond doubt that the proposed model can handle dynamic motion characteristics of either the background or foreground, the framework was also tested on different scenarios from the PETS dataset.

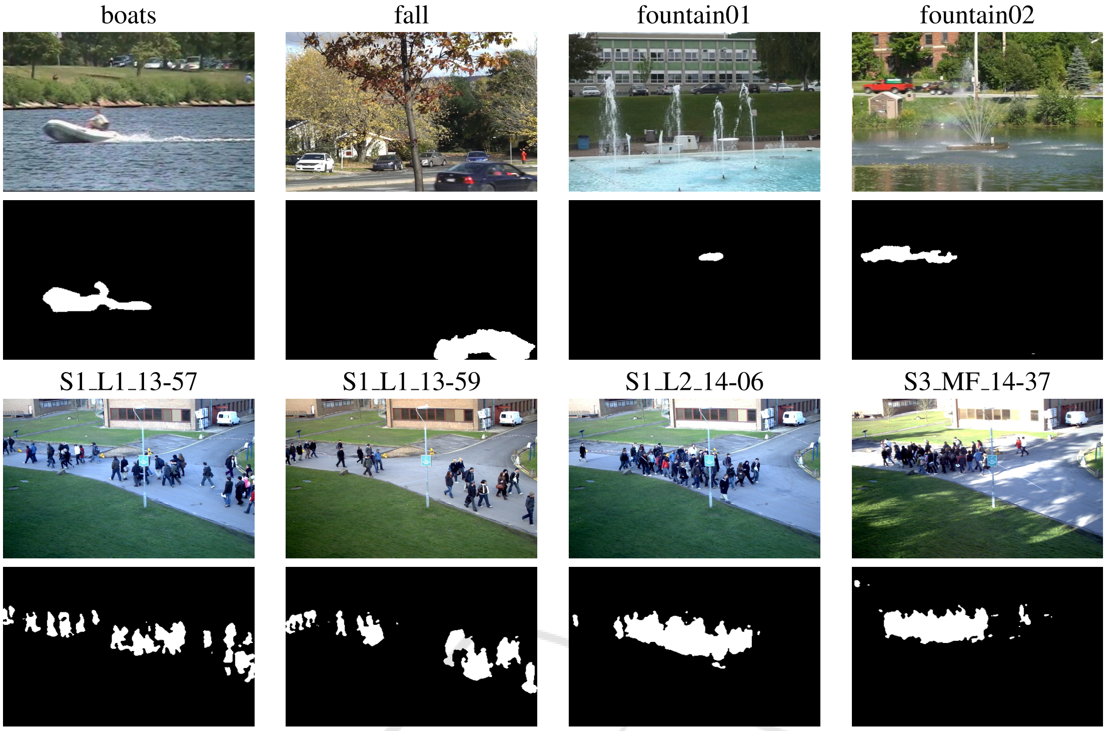


Figure 4: Segmentation examples using the proposed method on changedetection dataset and PETS dataset.

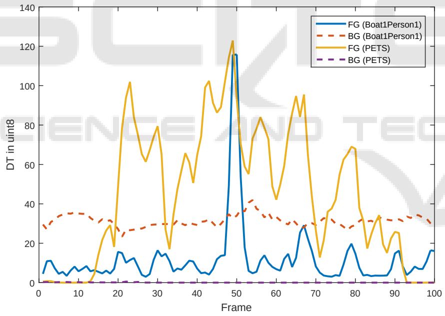


Figure 3: Variations of DT values of chosen blocks from the background (BG) and foreground (FG) of B1P1 sequence and S1_L1_13-57 PETS crowd sequence.

Fig. 4 displays qualitative results on changedetection and PETS datasets.

3.1 Patch Resolution Analysis

One key parameter of the proposed method is the size of the block patches. During empirical study, some close relationship between the patch size and the learning rate used within the GMM model could be observed. In order to formalize this relationship, the results of changing patch size against different learning rates on the recall of the detection process are presented in Fig. 5. The results in Fig.5 has been generated using the Boat1Person1 video sequence, as an example, from FBDynSyn dataset. It can be observed that at low learning rates of the GMM, smaller patch sizes produce better recall. Nevertheless, as higher learning rates are used, this is no longer the case.A qualitative assessment of the impact of learning rate and the patch size using sample frames from the Boat1Person1 sequence is illustrated in Fig.6.

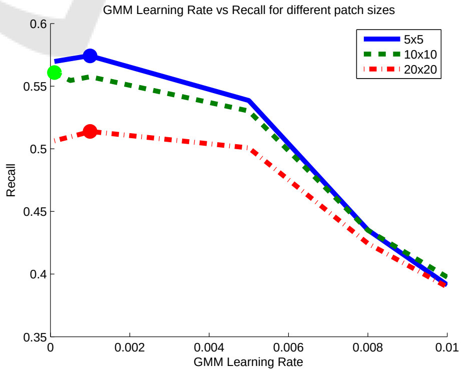


Figure 5: Relation between learning rate and recall for various patch sizes using Person1Boat1 video (the peaks indicated with points).

It can be noticed that with small patches at high learning rates, the amount of false positive detection increases. However, at lower learning rate at small patch sizes, the detection is more accurate. For Boat1Person1 sequence, the processing time was found to be 655 ms per frame at patch size 10x10, while it was 2251 ms per frame at patch size 5x5. Despite better performance at patch size 5x5, the computational overhead of performing background modeling at that patch size is significantly higher than at 10x10 patch size for a small compromise in accuracy.

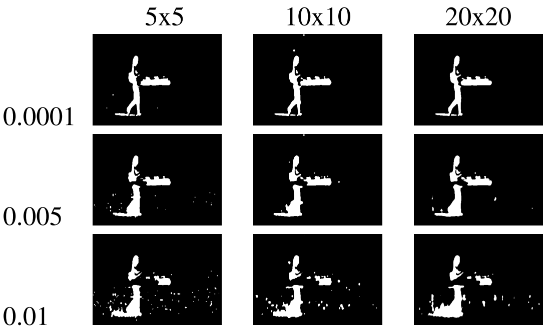


Figure 6: Segmentation example of Boat1Person1 video with different patch sizes and GMM learning rates.

4 CONCLUSIONS

The proposed method of spatio-temporal GMM of DT is an accurate and robust mechanism for background modeling and foreground detection. Evaluative performance follows the hypothesis underpinning the theoretical model. The relationships infused between key parameters of patch size and learning rate indicate that when the patch size is decreased, the number of patches and hence the number of DT components increases, thus yielding higher accuracy of detection even at fixed learning rate. However, on the contrary, with decrease in the patch size, the amount of motion information that is encapsulated within each patch is reduced, thereby causing slower recognition for the motion patterns at that chosen learning rate.

REFERENCES

Bhaskar, H., Mihaylova, L., and Maskell, S. (2007). Background modeling using adaptive cluster density estimation for automatic human detection. *Informatics 2*, pages 130–134.

Chan, A., Mahadevan, V., and Vasconcelos, N. (2011). Generalized stauffer–grimson background subtraction for dynamic scenes. *Machine Vision and Applications*, 22(5):751–766.

Chan, A. and Vasconcelos, N. (2009). Layered dynamic textures. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 31(10):1862–1879.

Dalley, G., Migdal, J., and Grimson, W. (2008). Background subtraction for temporally irregular dynamic textures. *WACV*.

Doretto, G., Chiuso, A., Wu, Y. N., and Soatto, S. (2003). Dynamic textures. *International Journal of Computer Vision*, 51:91–109.

Ferryman, J. and Shahrokni, A. (2009). Pets2009: Dataset and challenge. *Twelfth IEEE International Workshop on Performance Evaluation of Tracking and Surveillance*.

Goyette, N., Jodoin, P., Porikli, F., Konrad, J., and Ishwar, P. (2012). Changedetection.net: A new change detection benchmark dataset. In *IEEE Computer Society Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, pages 1–8.

Heikkila, M. and Pietikainen, M. (2006). A texture-based method for modeling the background and detecting moving objects. *IEEE TPAMI*, 28(4):657–662.

Mumtaz, A., Zhang, W., and Chan, A. B. (2014). Joint motion segmentation and background estimation in dynamic scenes. *CVPR*, pages 368–375.

Otsu, N. (1979). A threshold selection method from gray-level histograms. *IEEE TSMC*, 9(1):62–66.

Pokrajac, D. and Latecki, L. J. (2003). Spatiotemporal blocks-based moving objects identification and tracking. *IEEE Visual Surveillance and Performance Evaluation of Tracking and Surveillance (VS-PETS)*, pages 70–77.

Stauffer, C. and Grimson, E. (2000). Learning patterns of activity using realtime tracking. *IEEE TPAMI*, 22(8):747–757.

Stauffer, C. and Grimson, W. (1999). Adaptive background mixture models for real-time tracking. *CVPR*, 2.

Tian, Y.-L. and Hampapur, A. (2008). Robust salient motion detection with complex background for real-time video surveillance. *WACV/MOTIONS*.

Zhang, S., Yao, H., and Liu, S. (2009). Spatial-temporal nonparametric background subtraction in dynamic scenes. *ICME*, pages 518–521.

Zhong, B., Liu, S., Yao, H., and Zhang, B. (2009). Multl-resolution background subtraction for dynamic scenes. pages 3193–3196.

Zhong, J. and Sclaroff, S. (2003). Segmenting foreground objects from a dynamic textured background via a robust kalman filter. *ICCV*, 1:44–50.

Zhou, X., Yang, C., and Yu, W. (2013). Moving object detection by detecting contiguous outliers in the low-rank representation. *IEEE TPAMI*, 35(3).