A Hybrid Object Detection Technique from Dynamic Background Using Gaussian Mixture Models

Mahfuzul Haque, Manzur Murshed, and Manoranjan Paul

Gippsland School of Information Technology, Monash University
Victoria 3842, Australia
mahfuzul.haque@infotech.monash.edu.au
manzur.murshed@infotech.monash.edu.au
manoranjan.paul@infotech.monash.edu.au

***Abstract*—Adaptive background modelling based object detection techniques are widely used in machine vision applications for handling the challenges of real-world multimodal background. But they are constrained to specific environment due to relying on environment specific parameters, and their performances also fluctuate across different operating speeds. On the other side, basic background subtraction (BBS) is not suitable for real applications due to manual background initialization requirement and its inability to handle repetitive multimodal background. However, it shows better stability across different operating speeds and can better eliminate noise, shadow, and trailing effect than adaptive techniques as no model adaptability or environment related parameters are involved. In this paper, we propose a hybrid object detection technique for incorporating the strengths of both approaches. In our technique, Gaussian mixture models (GMM) is used for maintaining an adaptive background model and both probabilistic and basic subtraction decisions are utilized for calculating inexpensive neighbourhood statistics for guiding the final object detection decision. Experimental results with two benchmark datasets and comparative analysis with recent adaptive object detection technique show the strength of the proposed technique in eliminating noise, shadow, and trailing effect while maintaining better stability across variable operating speeds.**

I. Introduction

Object detection is the first task in any machine vision application for extracting moving objects, making it the most critical part of the system. Two types of object detection techniques are commonly used, one using a stored background image of the scene and then identifying the moving regions by finding its difference from current image with a threshold. This technique is known as basic background subtraction (BBS) [1] and the simplest technique of detecting moving objects, however it is unrealistic for real-world applications as the background may change over time, and the system needs to be initialized with clear background which is impossible in most cases. The second type of detection techniques use adaptive background modelling to cope with the challenges associated with the dynamics of real-world background with sudden and gradual illumination variations, intrinsic repetitive background motions, and global motions due to camera displacements.

Gaussian mixture models (GMM) is commonly used by adaptive techniques for background modelling, where each pixel is modelled independently by a mixture of at most K Gaussian distributions, and each Gaussian represents the colour/intensity distribution of one of the different environment components e.g., moving objects, shadow, illumination changes, sky, tree leaves, and static background, observed by the pixel over time. These techniques are appropriate for real-world applications, as the background model maintained by them can automatically evolve over time with the change of operating environment and handle multimodal background. However they are constrained to specific environment due to inherent dependency on environment related parameters, and their performances widely fluctuate across different operating speeds. On the other hand, basic subtraction can better eliminate noise, shadow, and trailing effect than adaptive techniques for using intensity thresholding, and also shows better stability across different operating speeds since no model adaptability or environment related parameter is involved.

In this paper, we propose a hybrid object detection technique for incorporating the strengths of both techniques. Our technique maintains a GMM-based adaptive background model for each pixel and determines two pixel level detection decisions. One decision is taken using a probabilistic formulation based on the learned mixture while the other is based on basic background subtraction considering multimodal background. Since, computation time is critical for real-time applications, multiple believed-to-be backgrounds for the basic subtraction are generated from the Gaussian mixture model without maintaining a separate model. Then, the final detection decision is guided by two inexpensive neighbourhood statistics computed based on those decisions.

The proposed technique is evaluated by extensive experiments with two benchmark datasets. Both qualitative and quantitative comparisons with recent GMM-based object detection technique clearly show its better stability across different operating speeds and the strengths in eliminating shadow, noise, and trailing effect. The results also validates the improvement in computational complexity in maintaining the underlying background model by prohibiting redundant model induction in the mixture.

II. Related Work

A classical and the most widely cited object detection technique was introduced by Stauffer and Grimson (S&G)[2] using adaptive Gaussian mixture models [3]. In this technique, each pixel is modelled using a separate Gaussian mixture, which is continuously learnt by an online approximation. Object detection at the current scene is then performed at pixel-level by comparing its value against the most likely background Gaussians, determined by a threshold T, representing the proportion by which the pixel is going to observe the background. However, simplicity of this technique in separating moving objects from multimodal background has attracted many researchers to enhance this technique further, primarily to improve its adaptability, computational complexity, and detection quality.

Lee [4] proposed an adaptive learning rate for each Gaussian model to improve the convergence rate without affecting the stability. He also incorporated a Bayesian framework to isolate the most likely background Gaussians and generate an intuitive representation of the believed-to-be background. The user-defined threshold T in the original work is replaced with two parameters of the sigmoid function modelling the posterior probability of a Gaussian to be background. Although these parameters are trained from some commonly observed surveillance videos, both are inherently relying on the proportion by which a pixel is going to observe the background. KaewTraKulPong and Bowden [5] also addressed the slow learning rate with a shadow detection algorithm.

Shimada *et al.*[6] proposed an approach for improving the computational time of the (S&G) technique by reducing the number of concurrent models for a pixel through merging.

Several multi-stage techniques are proposed to improve the detection quality. Zeng and Lai [7] developed a two stage background/foreground classification procedure where the pixel-based GMM classifier is augmented with a region-based classifier to remove undesirable subtraction due to shadow, automatic white balance, and sudden illumination changes. Huang *et al.*[8] addressed the same issue inversely, by first dividing each scene into a set of motion coherent regions, then constructing pixel-based background models, and finally using these models to classify each region into background/foreground. Zhang and Chen [9] introduced support vector machine to further classify foreground pixels into motion/non-motion classes to reduce false motion detection in complex background.

Allili *et al.*[10] improved the detection quality in the presence of sudden illumination changes and shadows by generalising the Gaussian pdf to accommodate better fitting of the background model.

In general, GMM-based adaptive object detection techniques can be broadly categorized into single stage and multi-stage techniques. Single stage techniques use only pixel level information while multi-stage techniques utilize the pixel level decisions for further improvement at higher level for region level classification.

III. The proposed technique

In the proposed technique, each pixel of a scene is modelled independently by a mixture of at most K Gaussian distributions. Let the kth Gaussian in the mixture be denoted as η_k with mean μ_k, variance σ_k^2, the most recently observed pixel value m_k, the number of observed pixel values c_k, and weight ω_k such that $\sum_{\forall k} \omega_k = 1$. Let $\eta_k(x)$ denotes the probability pixel intensity x in Gaussian η_k.

A. Model learning

The system starts with no model in the mixture of a pixel and then for every new observation x_t of the pixel at time t, it is first matched against each of the existing models where x_t is no further than 3 standard deviations or S from the mean. Here S is a constant, typically used in basic background subtraction as a background-foreground separating threshold. As setting S low has shown guaranteed high quality object detection for a wide range of surveillance test sequences in [11], we set a fixed value (S = 20) for all operating environments determined from a sensitivity analysis. Of all the matched models, the one (say η_i) with the maximum weight times the probability of x_t in the model is selected as follows:

$$i = \arg\max_{\forall k:|x_t-\mu_k|\leq \max(3\sigma_k,S)} \{\omega_k \eta_k(x_t)\}; \quad (1)$$

and its associated parameters are updated as follows:

$$m_i \leftarrow x_t; \quad (2)$$

$$c_i \leftarrow c_i + 1; \quad (3)$$

$$\beta_i \leftarrow (1-\alpha)/c_i + \alpha; \quad (4)$$

$$\sigma_i^2 \leftarrow (1-\beta_i)\sigma_i^2 + \beta_i (x_t - \mu_i)^2; \quad (5)$$

$$\mu_i \leftarrow (1-\beta_i)\mu_i + \beta_i x_t; \quad (6)$$

$$\omega_i \leftarrow (1-\alpha)\omega_i + \alpha. \quad (7)$$

If no match is found, a new Gaussian (say η_i) is introduced with $m_i = \mu_i = x_t$, $\sigma_i = 30$, $c_i = 1$, and $\omega_i = \alpha$. The weights of the remaining Gaussians are updated as

$$\forall k \neq i : \omega_k \leftarrow (1-\alpha)\omega_k \quad (8)$$

in both the cases. Finally weights of all the models are normalized such that $\sum_{\forall k} \omega_k = 1$. In the proposed technique, high quality mixtures are maintained than existing GMM-based techniques by not allowing redundant models in the mixtures with a relaxed model matching threshold. We propose to use $\max(3\sigma_k, S)$ instead of $3\sigma_k$ as model matching threshold for preventing near-duplication model induction when model variance becomes very small due to stable observations over time.

B. Object detection

Two independent detection decisions $D_i(x,y)$ $(1 \leq i \leq 2)$ are made for each pixel (x,y), where $D_i(x,y)$ can be 1 and 0 for foreground and background, respectively.

1) Probabilistic background subtraction, $D_1(x,y)$: First, we use a probabilistic formulation [4] to classify the current observation x_t:

$$P(B|x_t) = \frac{\sum_{\forall k} \eta_k(x_t)\omega_k P(\eta_k)}{\sum_{\forall k} \eta_k(x_t)\omega_k}; \quad (9)$$

where B represents the background class and x_t is classified as foreground if $P(B|x_t) < 0.5$. Here $P(\eta_k)$ is the background probability estimated using following sigmoid function:

$$P(\eta_k) = 1/(1 + e^{-a\omega_k/\sigma_k + b}); \quad (10)$$

where the constants $a = 96$ and $b = 3$ are suggested in [4] after sensitivity analysis on commonly observed surveillance sequences. The background probability $P(\eta_k)$ increases either with the increase of weight or decrease of variance.

2) Basic background subtraction with multi-background, $D_2(x,y)$: Multiple dominating backgrounds are automatically identified from the mixture for intrinsic repetitive background motion in the environment.

First, all existing models in the mixture are sorted in descending order of their background probabilities $P(\eta_k)$'s such that after sorting $P(\eta_1) \geq P(\eta_2) \geq \cdots \geq P(\eta_K)$. Now, η_1 always represents the most dominating background. Two different statistics are utilized for identifying multiple dominating backgrounds. One is obviously the weight of the Gaussian, as the existence of similarly weighted models corresponds to the existence of a repetitive multimodal background. The other is the observation stability. Standard deviation σ_k of model η_k is a good measure of its stability as low σ_k corresponds to stable observation and high σ_k indicates varying intensities. However, this is not true when the static background is revealed after a long observation of varying intensities due to moving foregrounds as it may introduce a new model η_i with very high σ_i. To avoid this situation, we use an alternative measure $d_k = |m_k - \mu_k|$ for each Gaussian η_k. This measure is always closer to zero and shorter in Gaussians with stable observations than those representing fluctuating observations. The test for Gaussian η_k, $k = 2, 3, \ldots$ is carried out as follows:

$$|\omega_1/d_1 - \omega_k/d_k|d_1/\omega_1 < f; \quad (11)$$

where the constant $f = 0.05$ is determined from a sensitivity analysis.

If B models are identified in the set of most dominating background models, a pixel with value x_t at time t is considered background if

$$\exists i_{i=1,\ldots,B} : |m_i - x_t| \leq S/i. \quad (12)$$

Note that the test threshold S/i decreases linearly with i to make sure that enough intensity band is left to represent the foreground even for large B. This measure is found to reduce false negative error where a foreground pixel is undetected.

3) Hybrid detection decision, $D(x,y)$: Final detection $D(x,y)$ for a pixel (x,y) is determined using the hybrid detection algorithm (Algorithm 1). For each pixel-level detection decision $D_i(x,y)$, a neighbourhood weight W_i is computed

(a) (b) (c)

Fig. 1. Object detection on *PETS2006-B1* sequence; (a) test frame; (b) ideal result; and (c) detection result. Red and green pixels represent detections by probabilistic subtraction and basic subtraction, respectively, while yellow pixels represent detections by both approaches.

representing the proportion of foreground pixels within the neighbourhood including that pixel:

$$W_i(x,y) = \forall_{(m,n)\epsilon N} D_i(m,n)/N_c; \quad (13)$$

here N is the neighbourhood of the pixel (x,y) and N_c is total number of pixels in N. Only a spatial 8-connected neighbourhood is considered without any temporal pixels for avoiding increased space and computational complexity.

Algorithm 1: Hybrid Detection

```
Input: D_i(x,y) and W_i(x,y) for all (x,y) (1 ≤ i ≤ 2)
Output: D(x,y) for all (x,y)
foreach (x,y) do
    if W_1(x,y) > 0.5 And W_2(x,y) > 0.5 then
        D(x,y) = Foreground
    else if W_1(x,y) > 0.5 And W_2(x,y) < 0.5 then
        if D_1(x,y) ∧ D_2(x,y) then
            D(x,y) = Foreground
        else
            D(x,y) = Background
        end
    else
        if W_2(x,y) > 0.5 then
            D(x,y) = Foreground
        else
            D(x,y) = Background
        end
    end
end
```

A pixel within a dense foreground region is identified by majority voting when the corresponding neighbourhood weight $W_i(x,y)$ is greater than 0.5. When both probabilistic and basic subtraction decisions identify a pixel within a dense foreground region ($W_1(x,y) > 0.5$ And $W_2(x,y) > 0.5$), the corresponding pixel is classified as foreground irrespective of it's own classification, as this will improve the detection quality inside object regions. The scenarios of moving shadows and trailing effect are identified when $W_1(x,y) > 0.5$ and $W_2(x,y) < 0.5$. These are also shown visually in Figure 1(c) by the red pixels, which are eliminated by ANDing of pixel level decisions ($D_1(x,y) \wedge D_2(x,y)$). For default case, priority is given to basic subtraction decision $W_2(x,y)$ for ensuring detection in dense foreground region and eliminating sparse noises in the scene.

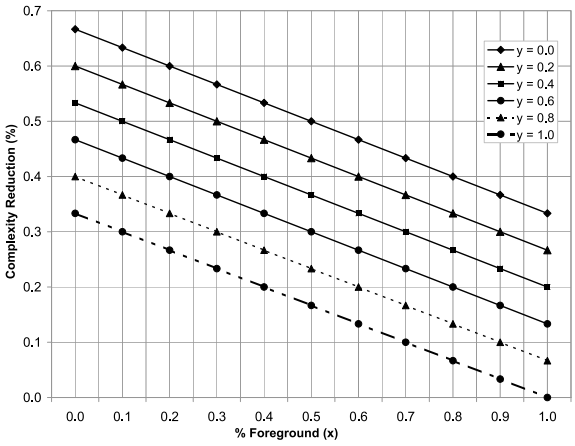

Fig. 2. Computational complexity reduction in the proposed model learning stage for $K = 3$. x and y represent the average proportions of foreground and multimodal pixels, respectively.

TABLE I
ERROR RATES AT MEDIUM LEARNING RATE ($\alpha = 0.01$) AND THE STANDARD DEVIATION OF THE ERROR RATES OVER THREE LEARNING RATES ($\alpha = 0.1$, $\alpha = 0.01$, AND $\alpha = 0.001$).

Test Sequence	**%Error Rate (FP + FN)**			
	$\alpha = 0.01$		**Stdev**	
	Lee	Proposed	Lee	Proposed
1. PETS2000	4.1	**1.6**	1.3	**0.0**
2. PETS2006-B1	10.3	**3.9**	1.2	**0.5**
3. PETS2006-B2	3.8	**2.7**	**0.3**	**0.3**
4. PETS2006-B3	5.6	**2.4**	1.1	**0.3**
5. PETS2006-B4	11.3	**5.6**	1.1	**0.9**
6. Bootstrap	13.3	**11.8**	2.1	**1.3**
7. Camouflage	29.8	**12.1**	9.6	**2.3**
8. Fground Aper.	67.2	**15.8**	7.4	**0.0**
9. Light Switch	86.1	**85.0**	32.9	**14.8**
10. Moved Object	0.5	**0.1**	**3.3**	3.8
11. Time Of Day	**4.1**	5.7	7.0	**0.6**
12. Waving Trees	19.2	**13.0**	0.5	**0.1**
13. Football	33.4	**21.2**	10.8	**2.4**
14. Walk	0.5	**0.2**	0.6	**0.1**

C. Computational complexity reduction in model learning, C_R

As mentioned before, the mixture quality of the proposed technique is improved by the modified model matching threshold $\max(3\sigma, S)$. However the complexity reduction will depend on the exposure of foreground and degree of multimodality in a particular operating environment. Only two models are sufficient for unimodal pixels with foreground, and a single model is enough for unimodal background pixels. In case of multimodal foreground pixels, no complexity gain is possible as all of the K models will be utilized. If x and y be the average proportions of foreground and multimodal pixels in a scene, then the total complexity reduction (C_R) can be expressed using the following equation:

$$C_R = \frac{x(1-y)(K-2)}{K} + \left(\frac{1-x}{K}\right)[y + (1-y)(K-1)] \quad (14)$$

Figure 2 plots the percentage of complexity reduction with different proportions of x and y for $K = 3$. The gain decreases with the increase of multimodality for a fixed foreground proportion. It also decreases for a constant multimodality with increased foreground proportion. At fast learning rate, on average 33% gain will be possible with different operating environments for $K = 3$.

IV. EXPERIMENTS

The proposed detection technique is evaluated by extensive experiments with 14 test sequences from *PETS* [12] and *Wallflower* [13] datasets. The results are compared qualitatively and quantitatively with the recent GMM-based object detection technique proposed by Lee [4].

A. Qualitative evaluation

We used a novel visualisation method for qualitative evaluation of mixtures by representing both the number of models and average mean (μ) distance among the models, where a unique colour is assigned to a mixture. Figure 3 shows the distance colour mapping for different number of models. Based on this mapping, a single RGB image can visualise the overall mixture quality of all pixels of a frame.

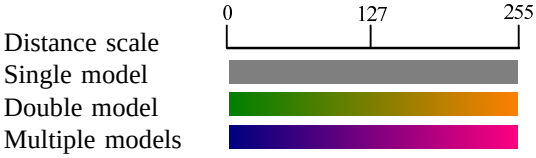

Fig. 3. Visualisation of model quality and number: distance colour mapping for single (gray), double (green-orange), and multiple (blue-pink) models.

Qualitative comparison results of both object detection and model quality are presented in Figure 4. For each sequence, first frame, test frame, ideal detection result, and actual detection result along with model quality visualisation are presented at medium learning rate, $\alpha = 0.01$ for both techniques. Due to space limitation, only a small subset of the qualitative results are presented here.

In general, we observed significant improvement in the detection result of the proposed technique in noise elimination, which is more prominent in *PETS2000* (Figure 4(a)) and *Waving Trees* (Figure 4(g)). *PETS2006-B1* (Figure 4(b)) and *PETS2006-B2* (Figure 4(c)) clearly show the strengths of the proposed technique in eliminating moving shadows and trailing effects, and getting the detection results almost close to ground truths. The mixture quality visualisations also reflect the high quality of the mixtures maintained by the proposed technique with less number of redundant models. In *PETS* sequences (Figure 4 (a)-(c)), these improvements are clearly apparent with higher proportion of gray regions.

B. Quantitative evaluation

Table I presents the error rates of test sequences at medium learning rate ($\alpha = 0.01$) and the standard deviation of error rates at three different learning rates ($\alpha = 0.1$, $\alpha = 0.01$, and

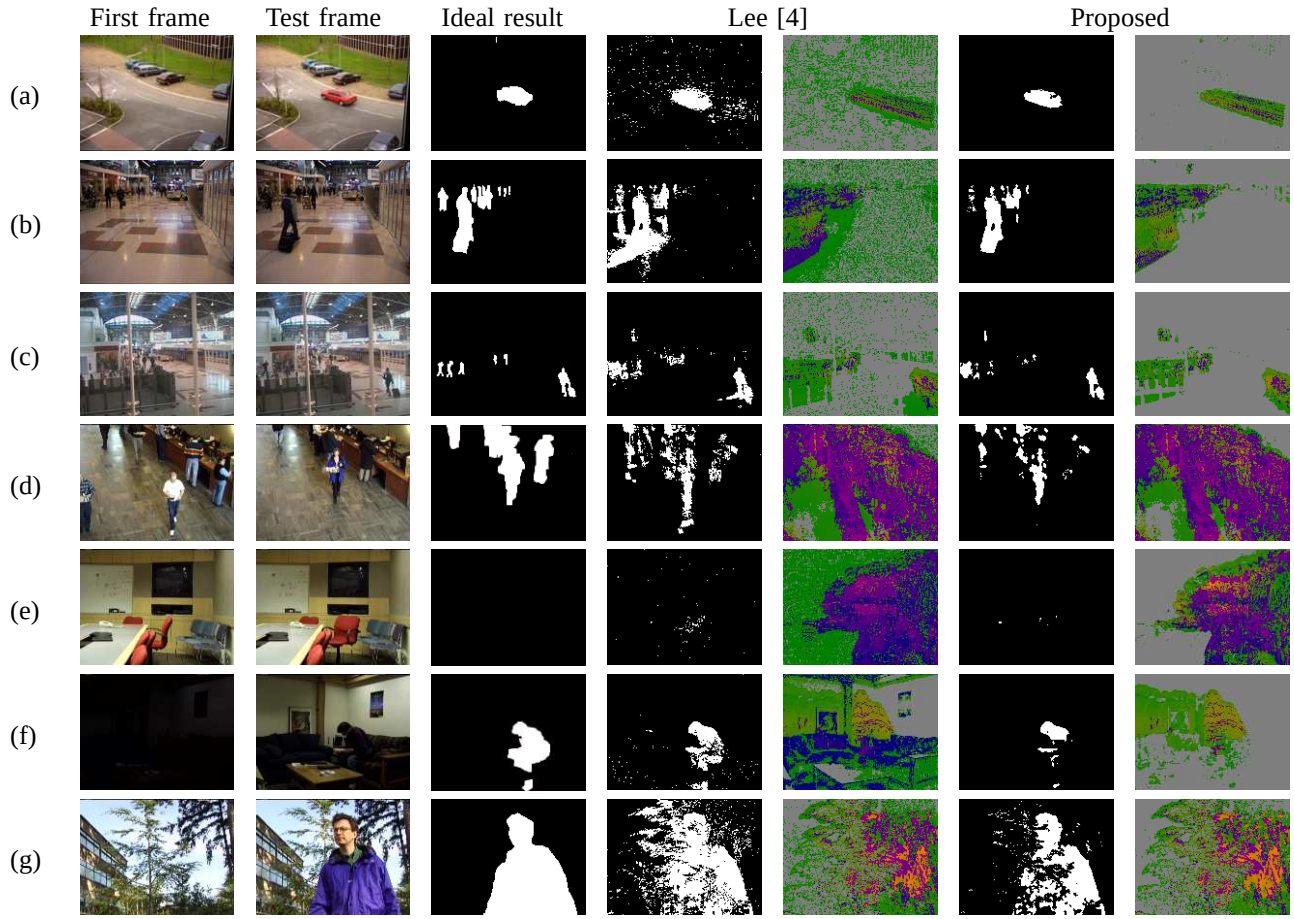

Fig. 4. Qualitative comparison results of object detection and model quality at learning rate, $\alpha = 0.01$ for test sequence (a) *PETS2000*; (b) *PETS2006-B1*; (c) *PETS2006-B2*; (d) *Bootstrap*; (e) *Moved Object*; (f) *Time Of Day*; and (e) *Waving Trees*.

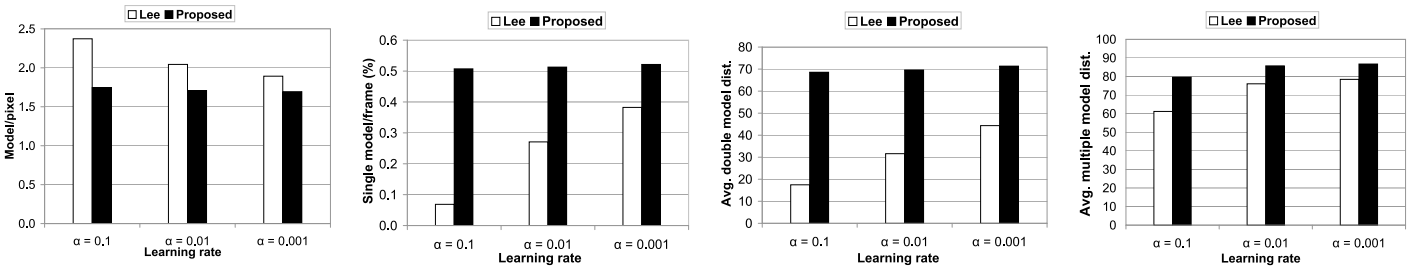

Fig. 5. Plots of four model quality measures: (a) average number of models per pixel; (b) proportion of single model per frame; (c) average mean distance for double model mixtures; and (d) average mean distance for multiple model mixtures.

$\alpha = 0.001$) for both the techniques (pairwise best shown in bold). The proposed technique was more robust and stable as its error rates were far less sensitive to learning rates.

In Figure 5, four different measures for mixture quality evaluation are plotted at three different learning rates. These measures are computed at the test frame of each sequence and the average on sequences are plotted for comparison. Both average number of models per pixel and proportion of single model per frame improved significantly at all learning rates, which are consistent with the analytical complexity reduction presented in Section III-C. The average mean (μ) distance among the models also increased, indicating the high quality of mixtures with distinct models.

The comparative quantitative results of detection errors are presented in Figure 6. The proposed technique outperformed the Lee's technique in 13 test sequences out of 14 as shown in the combined error (FN+FP) plot (Figure 6(c)).

V. Conclusion

The proposed object detection technique not only outperformed the recent GMM-based technique in more than 90%

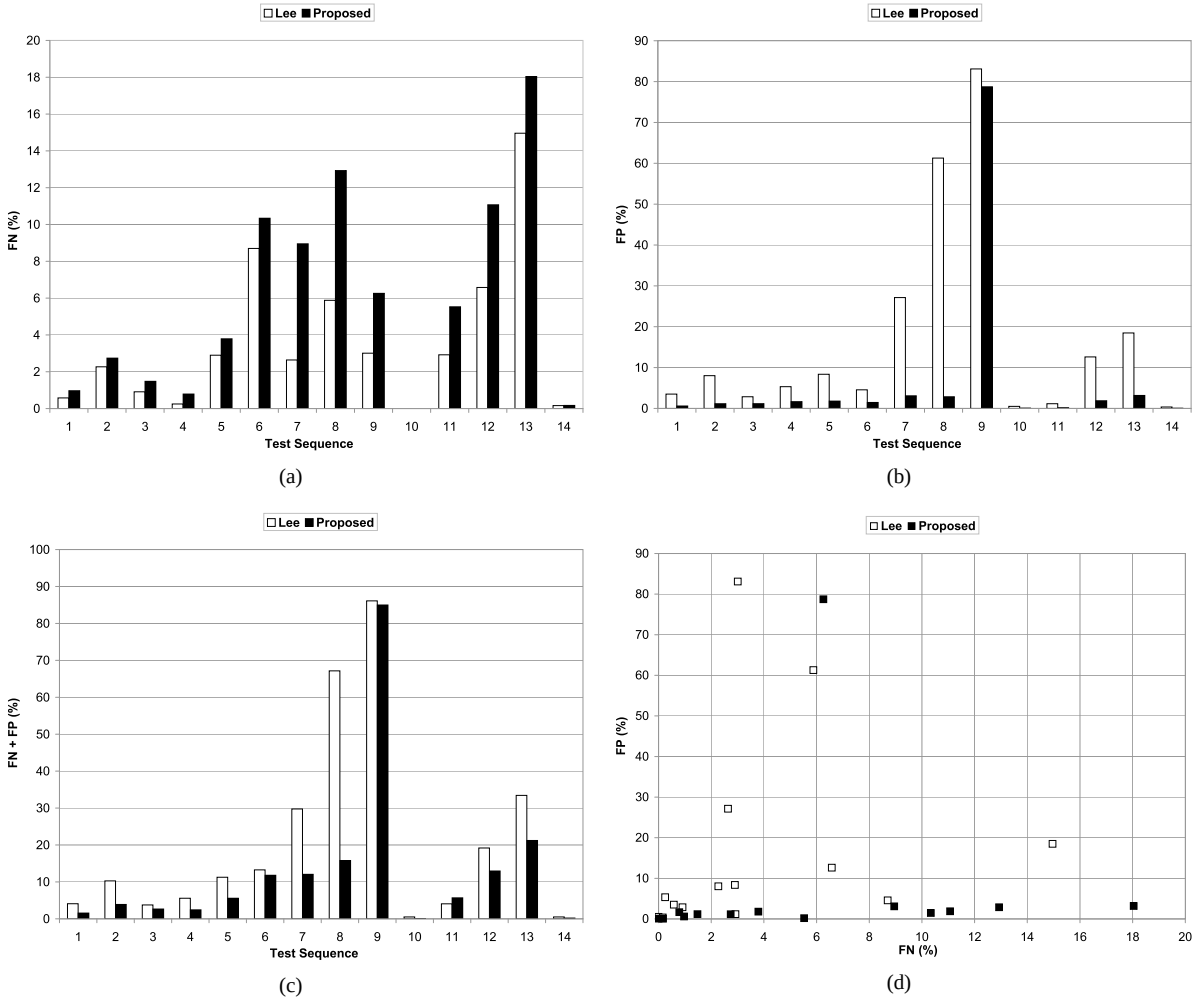

Fig. 6. Quantitative comparison results of the proposed technique presented in this paper and the technique of Lee [4] at learning rate, $\alpha = 0.01$ for 14 test sequences: (a) FN; (b) FP; (c) FN + FP; and (d) FP/FN plot. FN and FP stand for false negatives and false positives, respectively.

of the test sequences, but also showed better stability across different operating speeds with almost no shadow, noise, and trailing effect. All these attributes make the proposed technique an ideal choice for remote surveillance, where no prior information about operating environment is available.

REFERENCES

[1] S. S. Cheung and C. Kamath, "Robust techniques for background subtraction in urban traffic video," in *Video Communications and Image Processing. SPIE Electronic Imaging*, vol. 5308, 2004, pp. 881–892.

[2] C. Stauffer and W. E. L. Grimson, "Adaptive background mixture models for real-time tracking," in *IEEE Computer Society Conf. on Computer Vision and Pattern Recognition*, vol. 2, 1999, pp. 246–252.

[3] P. W. Power and J. A. Schoonees, "Understanding background mixture models for foreground segmentation," in *Image and Vision Computing New Zealand*, 2002, pp. 267–271.

[4] D. S. Lee, "Effective gaussian mixture learning for video background subtraction," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 27, pp. 827–832, 2005.

[5] P. KaewTraKulPong and R. Bowden, "An improved adaptive background mixture model for realtime tracking with shadow detection," in *2nd European Workshop on Advanced Video Based Surveillance Systems*, 2001.

[6] A. Shimada, D. Arita, and R. Taniguchi, "Dynamic control of adaptive mixture-of-gaussians background model," in *IEEE Int. Conf. on Video and Signal Based Surveillance*, 2006, pp. 5–5.

[7] H. C. Zeng and S. H. Lai, "Adaptive foreground object extraction for real-time video surveillance with lighting variations," in *IEEE Int. Conf. on Acoustics, Speech, and Signal Processing*, vol. 1, 2007, pp. 1201–1204.

[8] S. S. Huang, L. C. Fu, and P. Y. Hsiao, "Region-level motion-based background modeling and subtraction using mrfs," *IEEE Trans. Image Process.*, vol. 16, pp. 1446–1456, 2007.

[9] J. Zhang and C. H. Chen, "Moving objects detection and segmentation in dynamic video backgrounds," in *IEEE Conf. on Technol. for Homeland Security*, 2007, pp. 64–69.

[10] M. S. Allili, N. Bouguila, and D. Ziou, "A robust video foreground segmentation by using generalized gaussian mixture modeling," in *Fourth Canadian Conf. on Computer and Robot Vision*, 2007, pp. 503–509.

[11] J. C. Nascimento and J. S. Marques, "Performance evaluation of object detection algorithms for video surveillance," *IEEE Trans. Multimedia*, vol. 8, pp. 761–774, 2006.

[12] http://www.cvg.rdg.ac.uk/slides/pets.html. (2007, oct) Pets: Performance evaluation of tracking and surveillance.

[13] K. Toyama, J. Krumm, B. Brumitt, and B. Meyers, "Wallflower: Principles and practice of background maintenance," in *Seventh IEEE Int. Conf. on Computer Vision*, vol. 1, 1999, pp. 255–261.