A comparative study on mixture of Gaussians for object segmentation

Robina Khatoon [1], Syed Muhammad Saqlain Shah [1, *], Shafina Bibi [1], Imran Khan [1], Muhammad Usman Ashraf [1, 2]

[1]Department of CS & SE, International Islamic University, Islamabad, Pakistan
[2]Institute of Business Management Sciences, University of Agriculture, Faisalabad, Pakistan

A R T I C L E   I N F O

A B S T R A C T

Segmentation is the fundamental step in most of digital image processing and computer vision based applications for feature extraction. The purpose of segmentation is to partition an image into foreground and background. Numerous segmentation algorithms have been proposed for the last four decades ranging from degraded images; high and low contrast images, indoor video, outdoor videos, videos with static background and dynamic backgrounds. This paper presents evaluation and comparison of segmentation techniques used for real-time moving objects through static and adaptive number of Gaussians. The techniques are tested for both indoor and outdoor scenes. The comparison is presented on the basis of qualitative results and computational complexities.

1. Introduction

Segmentation is one of the fundamental steps for many computer vision applications to separate foreground and background. Extracted foreground can be used in a variety of applications related to surveillance, object detection and classification etc. A number of techniques for segmentation have been proposed by different researchers which include histogram methods, thresholding, compression based methods, split-and-merge methods, region growing methods, speaker's vocal identification, Multi resolution imaging segmentation, clustering methods etc. (Balafar, 2014; Douglas, 2015). Each technique offers some advantages and disadvantages based on results and efficiency.

When segmentation is done through mixture of Gaussian it may use k-mean algorithm for clustering of mixture components into foreground and background. It may also use expectation maximization algorithm for classification of foreground and background pixels but it is not cost effective when applied to each pixel of an image. A comparative study of segmentation is presented in this paper that uses mixture of Gaussian (Atev et al., 2004; Carlos et al., 2008). Both of the techniques are based upon adaptive multi-modal background subtraction method (Stauffer and Grimson, 1999).

One of the techniques focuses on the Gaussian parameters used in (Stauffer and Grimson, 1999) in order to improve the results while the other technique focuses on the computational cost of (Stauffer and Grimson, 1999) and presents the method to reduce the computational cost. In this paper two aspects of both techniques have been compared i.e., segmentation evaluation and computational cost.

This study is organized as follows: section two is based on related work and describes in detail the two techniques of segmentation based on mixture of Gaussian and their implementations, while section three is dedicated to the comparison of both the techniques based upon results and computational cost. Finally, section four is focused on the conclusion which is presented on the basis of analysis and experimental results.

2. Segmentation based on mixture of Gaussians

KaewTraKulPong and Bowden (2002) proposed a new technique for image segmentation in which segmentation is achieved through finding edges of a color image by using isotropic edge detector and then applying seeded region growing algorithm. Proposed technique may be used for automatic face detection and content based multimedia applications.

Stauffer and Grimson (1999) presented segmentation technique which was based on mixture of Gaussian. It was adaptive to deal robustly with illumination changes, repetitive motions of scene elements, tracking through cluttered regions, slow-

moving objects, and introducing or removing objects from the scene. Atev et al. (2004) proposed technique which was based on (Stauffer and Grimson, 1999). It performed segmentation by continuous monitoring of the brightness of pixels over time. This technique has been used in many applications related to pattern recognition (Yogameena et al., 2009; Yogameena et al., 2010). Carlos et al. (2008) introduced a new and efficient strategy for segmentation which reduced the computational cost of stauffer's technique (Stauffer and Grimson, 1999) as it used dynamic number of Gaussian per pixel.

Demirci and Karaguuml (2010) introduced segmentation technique for medical images by using phenomena of light refraction. This technique is based on region growing image segmentation method. Proposed technique is computationally more efficient than the existing methods as it does not need to keep knowledge about number of regions in the image.

Zezhi and Ellis (2014) used Gaussian mixture model for modeling the background in a scene and resultantly performed segmentation. They use adaptive learning instead of the static one to anticipate sharp changes in the scene. It anticipates the illumination changes as well. Srikanth et al. (2016) used Gaussian Mixture Model (GMM) for segmenting cervical cells that is a basic step for diagnosing cervical cancer from smear images. Yan and Shui (2015) in their paper used GMM by using spatial and color information provided by user as a marker. They formulated their problem as iterative energy minimization problem. Zeng et al. (2014) presented a three step methodology for image segmentation where they employed GMM for partitioning image into small groups followed by calculating the distance between GMM components through Kullback–Leibler (KL) divergence and finally similar GMM were merged through spectral clustering. Shah and Chauhan (2015) applied Hidden Markov Models (HMM) that used GMM for segmenting brain tumor from MRI images. They employed expectation maximization for achieving their objective and provided comparison with fuzzy c-mean based segmentation. Khatoon et al. (2012) used GMM for object segmentation where the primary task was human counting in crowd scenes.

In the mixture of Gaussian, at each pixel location mixture of K Gaussian distributions is applied to model luminance changes rather than explicitly modeling the values of all the pixels by considering one particular type of distribution. To classify a pixel as background or foreground it is analyzed on the basis of three parameters i.e., mean, variance and weight. Both of the techniques presented below are based on adaptive multi-modal background subtraction method, which was presented by Stauffer and Grimson (1999) which deals strongly with lighting changes, repetitive motions of elements in the scene, tracking through messy regions, slow-moving objects, and introducing or removing of scene objects. As the color of slow moving objects

has a large variance than the background so, slow moving objects remain in the background for a long time. KaewTraKulPong and Bowden (2002) expectation maximization (EM) presented standard method for maximizing the likelihood of the observed. Instead of using EM algorithm, K-means approximation is used because there is a mixture model for every pixel in the image. If exact EM algorithm is applied on each pixel of the image then it would be expensive.

In both the techniques under study sequence of n frames each of dimension W×H are taken as an input. In each frame I, pixel value (pv) at each position is represented by I(x; y), (0 < x < W and 0 < y < H). Probability of each color value of frame I; is represented by (Eq. 1):

$$\rho(pv) = \sum_{j=1}^{k} \rho(M_j)\eta(pv|\mu_j \Sigma_j) \tag{1}$$

Where

$\rho(M_j) \rightarrow$ Probability of jth mixture
$\mu_j \rightarrow$ Mean of jth mixture
Σ_j d×d covariance matrix for jth mixture

$$\eta\left(pv|\mu_j \Sigma_j\right) = \frac{exp^{(pv-\mu_j)\Sigma_j^{-1}(pv-\mu_j)}}{\sqrt{(2\pi)^d \|\Sigma\|}}$$

Techniques presented by both Atev et al. (2004) and Carlos et al. (2008) are based on adaptive multimodal background subtraction method (Stauffer and Grimson, 1999) which adapts itself to deal with illumination changes, tracking through cluttered regions, repetitive motions of scene elements, introducing or removing objects from the scene and slow-moving objects robustly.

The static number of Gaussians (Atev et al., 2004) is updated using (Eqs. 1-3):

$$\rho(M_j) = (1-\alpha)\rho(M_j) + \alpha \tag{2}$$
$$\mu_j = (1-\beta)\mu_j + \beta c \tag{3}$$
$$\Sigma_j = (1-\beta)\Sigma_j + \beta(c-\mu_j)(c-\mu_j)^T \tag{4}$$

While the update for dynamic number of Gaussians (Carlos et al., 2008) is done through following expressions (Eqs. 5-7):

$$\rho = \alpha\eta_j(I_n(x,y); \mu_n(x,y); \sigma_{n,j}(x,y)) \tag{5}$$
$$\mu_n(x,y) = (1-\rho)\mu_{n-1,j}(x,y) + \rho I_n(x,y) \tag{6}$$
$$\sigma_{n,j}^2 = (1-\rho)\sigma_{n-1,j}^2(x,y) + \rho(I_n(x,y) - \mu_{n-1,j}(x,y))^2 \tag{7}$$

The values of parameters used in (1)-(9) are shown in Table 1. Mixture of Gaussian for segmentation, instead of explicitly modeling values of all the pixels as one particular type of distribution, uses a mixture of K-Gaussian distribution to model luminance change at each pixel location. Each pixel is analyzed on mean, variance and weight to classify it as foreground or background.

In Fixed Gaussian case (Atev et al., 2004) once a pixel value matches some jth mixture then its classification towards foreground or background is achieved using (Eq. 8)

if $\sum_{i=1}^{j-1} \rho(M_j(x,y)) \leq \beta$ (8)

$\beta \epsilon [0, \ldots, 1]$

If the above relation (8) holds then pixel at (x,y) is declared as background otherwise it is declared as foreground. For adaptive case (Carlos et al., 2008) every pixel at (x,y) in nth frame is needed to be evaluated whether it belongs to foreground or the background. If the below Eq. 9 holds i.e.,

if $\|NOG_n(x,y) - NOG_{n-1}(x,y)\| = 0$ (9)

Then the pixel corresponding at (x,y) is considered as background otherwise it is assumed to that it belongs to foreground. Where NOG represents number of Gaussians.

3. Evaluation and comparison

In this section details of experimental setup and experimental results are presented.

3.1. Experimental setup

The experimental setup defines dataset, system platform, parameter selection and evaluation metrics Eqs. 10 and 11).

i) Dataset & System Platform: CAVIAR dataset is used to evaluate performance of the two techniques. Total number of tested frames is 4000. Dataset have different sequences of outdoor videos with different resolutions. Algorithms are tested by implementing through MATLAB 7.3 on machine with Pentium IV 1.8 GHZ processor.

ii) Model Parameter selection: Table 1 presents details of the parameters used for both algorithms.

iii) Qualitative Measuring Parameters: Precision and Recall are used to compute Qualitative performance of both techniques i.e.

$Precision = \frac{Number\ of\ true\ positives\ detected}{Total\ number\ of\ positives\ detected}$ (10)

$Recall = \frac{Number\ of\ true\ positives\ detected}{Total\ number\ of\ true\ positives}$ (11)

The total number of positives is calculated from the ground truths. Ground truths of the input frames are created both by manual method and automatic computation. The video frames where background is available, ground truths are created through background subtraction followed by post processing. The video frames where backgrounds are not available, ground truth are created through manual manipulations.

Table 1: Parameter values used by Atev et al. (2004) and Carlos et al. (2008)

Parameters	Symbols	Parameter used in (Atev et al., 2004)	Parameter used in (Carlos et al., 2008)
Number of Gaussians	N	4	Between 1 and 6
Learning Rate	α	10^{-3}	10^{-3}
Learning Rate	β	$\alpha/4$	$\alpha * n(I_n(x,y); \mu_{n,i}(x,y); \sigma_{n,i}(x,y))$
Match Threshold	δ	2.5	2.5
Background Proportion	B	0.25	0.25
Initial weight		α	1/C
Initial variance	320.0	320.0	36
Low variance threshold		49.0	

3.2. Experimental results

The average number of true positives detected and total number of positives detected for technique (Atev et al., 2004) and (Carlos et al., 2008) are for each scenario are shown in Table 2.

The overall precision and recall of technique (Atev et al., 2004) and (Carlos et al., 2008) for each scenario is given in Table 3.

Table 2: Comparison of Atev et al. (2004) and Carlos et al. 2008) on the basis of number of true positives detected and total number of positives detected

Scenario	Avg. No of positives detected (Atev et al., 2004)	Avg. No of True positives detected (Atev et al., 2004)	Avg. No of positives detected (Carlos et al., 2008)	Avg. No of True positives detected (Carlos et al., 2008)
1	1.7089e+004	9.6810e+003	1.3658e+004	6.9194e+003
2	3.7674e+004	1.5260e+004	6.9441e+003	5.8330e+003
3	5.8729e+003	2.3923e+003	1.4487e+004	7.2111e+003

Table 3: Comparative results of precision and recall of Atev et al. (2004) and Carlos et al. (2008) for different scenarios

Scenario	No. of Frames	Avg. Precision (Atev et al., 2004)	Avg. Precision (Carlos et al., 2008)	Avg. Recall (Atev et al., 2004)	Avg. Recall (Carlos et al., 2008)
1	27	0.566	0.470	0.981	0.874
2	25	0.403	0.840	0.781	0.470
3	18	0.498	0.618	0.442	0.631

Scenario 1: Segmentation is applied on a video sequence in which one person is walking in corridor using both techniques. The result of precision and recall for 27 frames of the video sequence is given in Fig. 1.

The precision of the technique by Atev et al. (2004) is shown with green line and technique by

Carlos et al. (2008) is shown with red color in Fig. 1 (b). Precision of the technique by Atev et al. (2004) shows that almost each frame has same precision but precision is low for each frame as on average it is 0.5666 as shown in Table 3. The reason behind this evaluation is that the false segmentation rate is high as the number of positives detected is larger than the number of true positives detected. The technique by Carlos et al. (2008) presents high peaks as well as low peaks in precision graph. The high peaks are observed in the frames in which pixels are segmented almost as true positives detected, especially in frame 3, whereas the low peaks are due to the high rate of false segmentation like in frame 18 and 24. So, the average precision of technique by Carlos et al. (2008) is 0.4702 (Table 3) and on average the total number of positives detected is much greater than number of true positives detected.

Recall of the technique by Atev et al. (2004) is shown with green line and technique by Carlos et al. (2008) is shown with red color in Fig. 1 (c). According to the graph recall of the technique by Atev et al. (2004) is higher than technique by Carlos et al. (2008) as the former correctly segments most of the pixels in comparison with the later technique. So, on average the recall of technique by Atev et al. (2004) is 0.9811 and for the technique by Carlos et al. (2008) is 0.8741 which are shown in Table 3.

Scenario 2: Both techniques of segmentation are applied on a video sequence in which one person is walking in corridor having large shadow. The result of precision and recall for 25 frames of the video sequence is given in Fig. 2.

The precision of technique by Atev et al. (2004) is lower than technique of Carlos et al. (2008) on average, which is 0.4033. It gets lower as low number of true positives is detected than total number of positives detected in almost all frames.

Whereas, for technique of Carlos et al. (2008), we find either high peaks equal to 1 or low peaks equal to zero. The reason behind the low peaks is that it segments only background while leaving the foreground. This behavior of the methodology is for such frames where no moving pixels were present. The overall precision of the technique by Carlos et al. (2008) is 0.8400.

Recall graph for Carlos et al. (2008) contains most of the falls which shows that these frames are not correctly segmented. The recall for Atev et al. (2004) is not very low as compared to technique by Carlos et al. (2008) because the former segments most of pixels which should be segmented as moving pixel. The overall recall of Atev et al. (2004) is 0.7808 and for the technique of Carlos et al. (2008) is 0.4703 as shown in Table 3.

The results of precision and recall show that for such type of scenes in which some frames of video have very slow moving objects or objects remain static for a small chunk of time, technique of Atev et al. (2004) outperforms over the technique of Carlos et al. (2008) as it contains high number of Gaussian

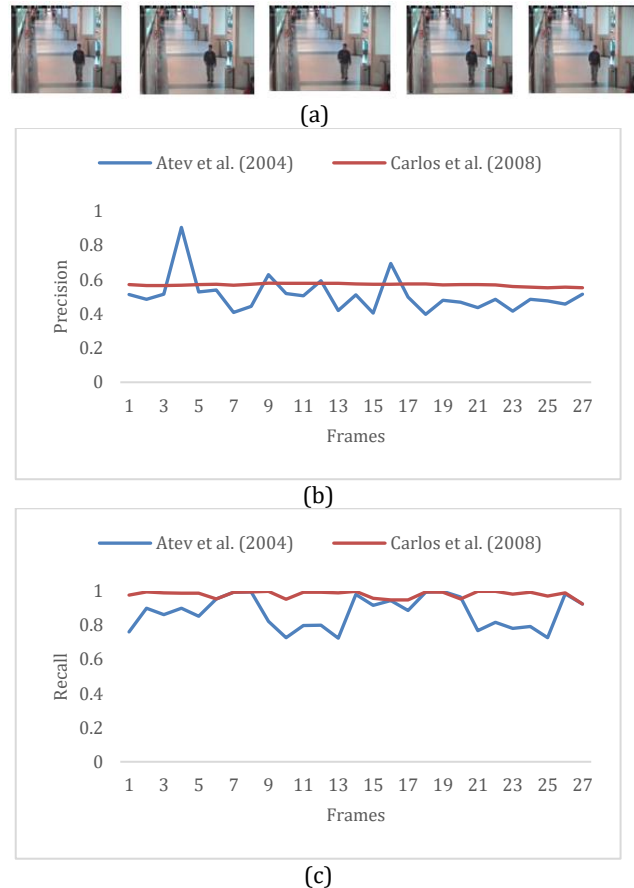for each pixel in comparison with the later technique.

(a)

(b)

(c)

Fig. 1: (a) Images from scenario 1 (b) precision of Atev et al. (2004) and Carlos et al. (2008) (c) recall of Atev et al. (2004) and Carlos et al. (2008)

Scenario 3: Both techniques of segmentation are applied on a video sequence in which one person is walking in outdoor scene having flickering leaves in the background. The results of precision and recall for 18 frames of the video sequence are given in Fig. 3. Precision of the technique by Atev et al. (2004) is 1 for first four frames and then there is an abrupt change. The abrupt change in precision is due to the pixels segmented as foreground which are actually the pixels of background. The technique by Carlos et al. (2008) has zero precision for the first frame as it has segmented first frames as totally background frames with no foreground pixel. Overall precision for Atev et al. (2004) is 0.4982 and for the technique by Carlos et al. (2008) is 0.6185 as given in Table 3.

The recall of technique by Atev et al. (2004) is low as it wrongly segments the flickering leaves in the background whereas the Carlos et al. (2008) copes better with such scenarios so its recall is high.

The result of precision and recall shows that for such type of dynamic background scenes Carlos et al. (2008) outperforms over technique Atev et al. (2004).

3.3. Discussion

In the test images as shown in Fig. 4, the frame 3 and 4 are constant. It means that there is no motion

between these frames. The results obtained from the technique by Atev et al. (2004) cope with such condition whereas the result obtained from the technique by Carlos et al. (2008) do not segment human in such condition where no motion is detected in the frame. Secondly, the blob of human are complete in second row whereas the parts of blob of human is missing in second row so some of the human detection techniques cannot such incomplete blobs as human.
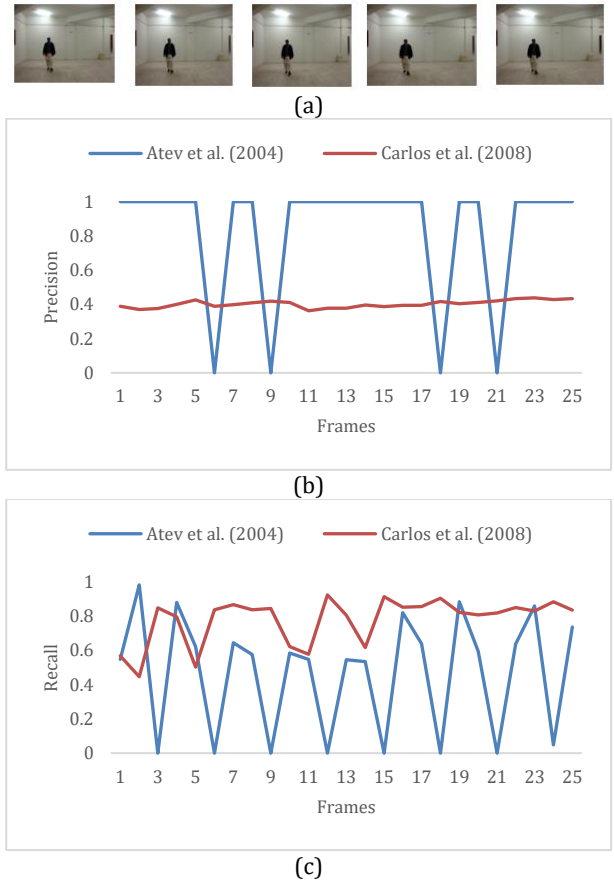

(a)

(b)

(c)

Fig. 2: (a) Images from scenario 2 (b) precision of both techniques (c) recall of both techniques

This difference is due to the reason that technique (Atev et al., 2004) uses same number of Gaussian for each pixel which keep the status of previous frame pixel and these values are used for next frame if pixels are detected as static whereas, technique (Carlos et al., 2008) assign Gaussians dynamically for each frame so it loses the previous information of pixels.

In Fig. 5 test images contain the frames of human interaction in which humans are in standing and moving backwards quickly due to the action of kicking. The results obtained from the technique (Carlos et al., 2008) evaluate the foreground pixels in row 3 better as compare to the technique (Atev et al., 2004) as shown in row 2.

The reason is that the technique (Atev et al., 2004) contains fix number of Gaussian and the technique (Carlos et al., 2008) contains dynamic number of Gaussian so it does not remember the previous state of the pixel and do not evaluate a pixel as foreground or background pixel on the basis of its

state in previous image. So, the technique (Carlos et al., 2008) gives better results of segmentation for actions like fighting and kicking as such actions which give abrupt change in each frame of a video sequence.
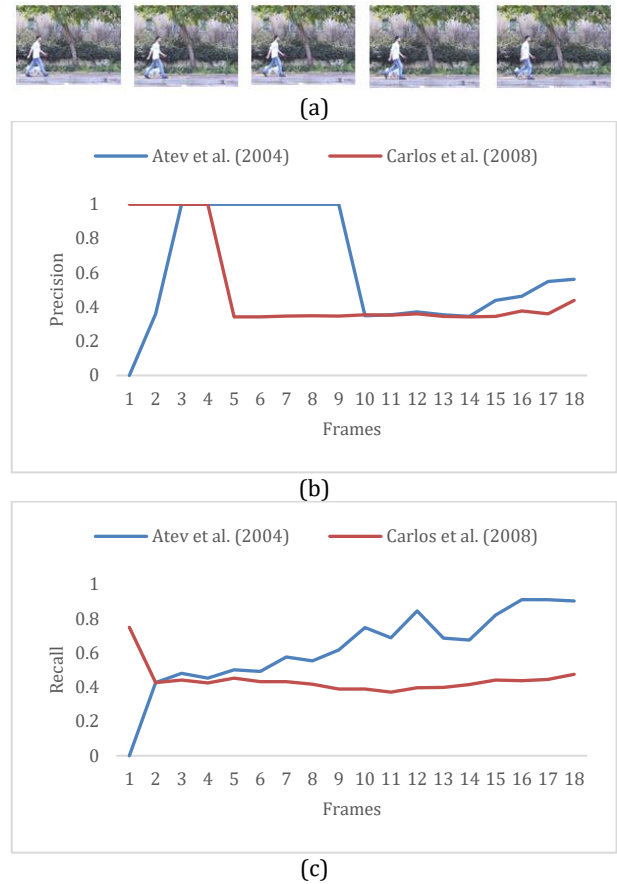

(a)

(b)

(c)

Fig. 3: (a) images from scenario 3 (b) precision of both techniques (c) recall of both techniques

Fig. 4: First row contains the tested frames, Second row contains the result of segmentation obtained from Atev et al. (2004) and third row contains the result of segmentation obtained from Carlos et al. (2008)

Fig. 6 is showing examples of segmentation over two different images. These are taken as example from two different video sequences. It may be observed that for each of the frame, original image along with its ground truth and segmentation results achieved through Atev et al. (2004) and Carlos et al. (2008) are shown. The ground truth for Frame A is manually created while for Frame B it was generated through background subtraction technique as background frames for its video sequences were known. The true positives for frame are 3273 calculated from its ground truth.

The dynamic number of Gaussian technique (Carlos et al., 2008) segmentation gives 4811 pixels as fore ground out of which 3273 are true positives detected and 1538 are false positives while the static Gaussian technique (Atev et al., 2004) detected 5479 pixels as foreground with 3273 true positives detected and 2206 pixels are wrongly detected as positives.
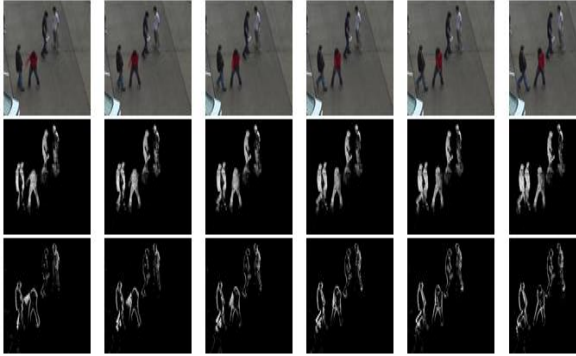

Fig. 5: First row contains the tested frames, Second row contains the result of segmentation obtained from Atev et al. (2004) and third row contains the result of segmentation obtained from Carlos et al. (2008)

For the frame B true positives obtained through its ground truth are 24616. The technique (Carlos et al., 2008) detects 37758 pixels as foreground out of which 18290 are true positives while 19468 are false positives. The methodology (Atev et al., 2004) segments 35407 pixels as foreground from which 18671 are truly classified while 16736 are misclassified as foreground.
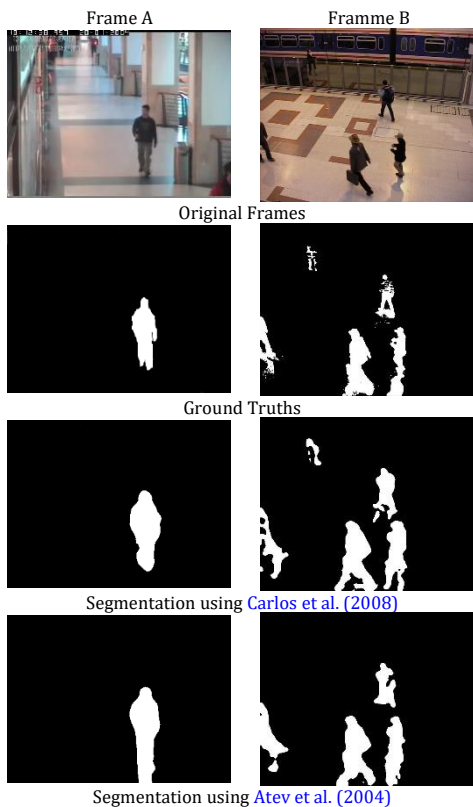

Original Frames

Ground Truths

Segmentation using Carlos et al. (2008)

Segmentation using Atev et al. (2004)

Fig. 6: Comparative results of techniques under study along with ground truths

3.4. Computational cost

The computational cost of both the techniques as can be seen in Table 4 depends upon the number of Gaussians used to model each pixel of an image.

Table 4: Comparative computational cost

Techniqe	Average Case	Best Case	Worst case
(Atev et al., 2004)	n[(wxhx4)+12]	n[(wxhx4)+12]	n[(wxhx4)+12]
(Carlos et al., 2008)	n[(wxhx2)+1]	n[(wxh)]	n[(wxhx6)+30]

The computational cost is directly proportional to the number of Gaussians. The generalized computational cost for techniques of Atev et al. (2004) and Carlos et al. (2008) are shown in Eq. 12 and Eq. 13 i.e.,

$$n[(w * h * 4) + 12] \tag{12}$$
$$n[(w * h * c) + c(c - 1)] \tag{13}$$

where n is number of images, w and h is the width and height of the corresponding image and **c** is number of Gaussians, while the term $c(c - 1)$ is the cost of sorting the values of components for each pixel of an image. Four Gaussians are used for Atev et al. (2004) for each pixel so the average, best and worst case for modeling a pixel as foreground and background is same. Variable number of Gaussians is used in Carlos et al. (2008) for each pixel. In this case c varies from one to six. In best case c=1, for average c=2 and c=6 as worst case.

4. Conclusion

According to the evaluation and comparison of both the techniques we come to know that the results of technique proposed by Atev et al. (2004) are better than the results of the technique proposed by Carlos et al. (2008). But the computational cost of Carlos et al. (2008) is less than the technique proposed by Atev et al. (2004). So, both of the techniques are good from some aspect either according to computational cost or good results. Future work can be the combination of both the techniques in order to combine the advantages of both. In this way we can do segmentation that will give good results and have a low computational cost as well.

References

Atev S, Masoud O, and Papanikopoulos N (2004). Practical mixture of gaussians with brightness monitoring. In The 7th International IEEE Conference on Intelligent Transportation Systems, IEEE, Washington, USA: 423-428. https://doi.org/10.1109/ITSC.2004.1398937

Balafar MA (2014). Gaussian mixture model based segmentation methods for brain MRI images. Artificial Intelligence Review, 41(3): 429-439.

Carlos C, García N, and Salgado L (2008). A new strategy based on adaptive mixture of gaussians for real-time moving objects segmentation. In the Conference on Real-Time Image

Processing, International Society for Optics and Photonics (SPIE), USA. http://dx.doi.org/10.1117/12.768139

Demirci R and Karaguuml T (2010). Light refraction based medical image segmentation. Scientific Research and Essays, 5(10):1127-1132.

Douglas R (2015). Gaussian mixture models. In: Li SZ and Jain AK (Eds.), Encyclopedia of biometrics: 827-832. Springer, USA.

KaewTraKulPong P and Bowden R (2002). An improved adaptive background mixture model for real-time tracking with shadow detection. In: Remagnino P, Jones GA, Paragios N, and Regazzoni CS (Eds.), Video-based surveillance systems: 134-144. Springer, USA.

Khatoon R, Saqlain SM, and Bibi S (2012). A robust and enhanced approach for human detection in crowd. In the 15th International Conference on Multitopic (INMIC'12), IEEE, Islamabad, Pakistan: 215-221. https://doi.org/10.1109/INMIC.2012.6511457

Shah S and Chauhan NC (2015). An automated approach for segmentation of brain MR images using Gaussian mixture model based hidden markov random field with expectation maximization. Journal of Biomedical Engineering and Medical Imaging, 2(4): 58-70.

Srikanth R, Narasimhan S, Basavaraj MG, and Dewar R (2016). Unsupervised segmentation of cervical cell images using Gaussian mixture model. In the 29th IEEE Conference on Computer Vision and Pattern Recognition Workshops (CVPR'16), IEEE, Las Vegas, USA: 70-75. https://doi.org/10.1109/CVPRW.2016.173

Stauffer C and Grimson W (1999). Adaptive background mixture models for real lime tracking. In the IEEE Computer Society Conference on Computer Vision and Pattern Recognition, IEEE, Fort Collins, USA, 2: 246-252. https://doi.org/10.1109/CVPR.1999.784637

Yan M and Shui PL (2015). Interactive image segmentation based on Gaussian mixture models with spatial prior. International Journal of Multimedia and Ubiquitous Engineering, 10(7): 105-114.

Yogameena B, Komagal E, Archana M, and RajuAbhaikumar S (2010). Support vector machine-based human behavior classification in crowd through projection and star skeletonization. Journal of Computer Science, 6: 1008-1013.

Yogameena B, Veeralakshmi S, Komagal E, Raju S, and Abhaikumar V (2009). RVM-based human action classification in crowd through projection and star skeletonization. EURASIP Journal on Image and Video Processing, 2009: 164019. https://doi.org/10.1155/2009/164019

Zeng S, Rui H, Zhen K, and Nong S (2014). Image segmentation using spectral clustering of Gaussian mixture models. Neurocomputing, 144: 346-356.

Zezhi C and Ellis T (2014). A self-adaptive Gaussian mixture model. Computer Vision and Image Understanding, 122: 35-46.