EURASIP Journal on Image and Video Processing
a SpringerOpen Journal

RESEARCH Open Access

Moving object detection using keypoints reference model

Wan Mimi Diyana Bt. Wan Zaki*, Aini Hussain and Mohamed Hedayati

Abstract

This article presents a new method for background subtraction (BGS) and object detection for a real-time video application using a combination of frame differencing and a scale-invariant feature detector. This method takes the benefits of background modelling and the invariant feature detector to improve the accuracy in various environments. The proposed method consists of three main modules, namely, modelling, matching and subtraction modules. The comparison study of the proposed method with a popular Gaussian mixture model proved that the improvement in correct classification can be increased up to 98% with a reduction of false negative and true positive rates. Beside that the proposed method has shown great potential to overcome the drawback of the traditional BGS in handling challenges like shadow effect and lighting fluctuation.

1. Introduction

Today, every state-of-the-art security system must include smart video systems that act as remote eyes and ensure the security and safety of the environment. One of the main challenges in any visual surveillance systems is to identify objects of interest from the background. Background subtraction (BGS) is the most widely used technique for object detection in real-time video application [1,2].

There are various approaches in BGS modelling. Running Gaussian average (RGA) [3], Gaussian mixture model (GMM) [4,5], kernel density estimation [6] and median filtering [7,8] are the most common methods due to their reasonable accuracy and speed. Although all these techniques work moderately well under simple conditions, because they treat each pixel independently without considering its neighbouring area, their performance depends strongly on environmental variation like illumination change.

Recently, affine region detectors have been used in quite varied applications that deal with extracting the natural features of objects. These detectors identify similar regions in different images regardless of their scaling, rotation or illumination. In this article, we propose a new method by combining the affine detector with a simple BGS model to detect moving-object for real-time video surveillance.

The rest of this article is organized as follows: Section 2 reviews some previous work on BGS and affine region detectors; Section 3 describes our approaches for keypoint modelling; Section 4 compares GMM with our proposed model and discusses the final result; and, finally, Section 5 concludes and provides recommendations based on the results.

2. Background

2.1 Background subtracting methods

For the past decades, various BGS approaches have been introduced by researchers for different challenging conditions [1]. Frame differencing is the most basic method in BGS. This method subtracts a frame at (t - 1) from a frame at time (t) to locate the foreground object. Median modelling [7] is another simple and popular approach in which the background is extracted based on the median value of the pixel sequence. In a complement to median filtering, McFarlane and Schofield [9] use a recursive filter to estimate median filtering to overcome the drawback of the previous model. The famous RGA was proposed later in [3]. This recursive technique modelled colour distribution of each pixel as a single Gaussian and updated the background model with the aim of adaptive filtering (Equation 1).

$$\mu_{t+1} = \propto F_t + (1-\propto)\mu_t \tag{1}$$

* Correspondence: wmdiyana@eng.ukm.my
Smart Engineering System Research Group, Department of Electrical, Electronic and Systems Engineering, Faculty of Engineering and Built Environment, Universiti Kebangsaan Malaysia, 43600 UKM Bangi, Selangor, Malaysia

where μ is mean and α is variance and F is pixel at time t.

Owing to the simplicity and reasonable accuracy of a single Gaussian, Stauffer and Grimson [5] proposed one of the most reliable BGS modelling techniques [10], the GMM. This method clusters each uniform object into k different Gaussian distributions (Equation 2).

$$P(x_t) = \sum_{i=1}^{K} \omega_{i,t} \cdot \eta(u; \mu_{i,t}, \sigma_{i,t}) \tag{2}$$

In Equation 2, $\eta(u; \mu_{i,\ t}$ i, t $\sigma_{i,\ t})$ is the ith Gaussian component, $\sigma_{i,\ t}$ is the standard deviation and $\omega_{i,\ t}$ is the weight of each distribution and u is the distribution model. The parameter K is the number of the distribution.

2.2 Scale invariant feature detectors

Regarding scale-invariant feature detectors, recently there have been several approaches proposed in various pieces of literature, but undoubtedly, most of today's state-of-the-art detectors rely on the Harris detector, which was introduced by Harris and Stephens [11]. As an enhanced feature detector, the popular scale-invariant feature transform (SIFT) algorithm [12] combines the Harris operator with a staged filtering approach to extract the scale-invariant feature. The scale-invariant feature is constant with respect to image translation, scaling and rotation, and partially invariant to illumination. The main drawback of SIFT is that it suffers from high computational time. Two related methods, the Hessian-affine detector and the Harris-affine detector, were proposed by Mikolajczyk et al. [13,14], and are another well-known set of algorithms that rely on the Harris measurement. As a matter of fact, the Hessian- and the Harris-affine detectors are identical in most cases because both detect points of interest in scale-space and use Laplacian operators for scale selection. In addition, they use the second moment of the matrix for describing the local image structure.

The second moment matrix describes the gradient distribution on a local neighbourhood of each feature and the eigenvalues of this matrix represent the signal changes neighbouring the point. Therefore, the extracted points are more stable in arbitrary lighting changes and pixel variations.

Another common technique is the speed up robust feature (SURF) [15], which is inspired by the SIFT detector and is based on the determinant Hessian matrix.

The most important feature of this detector is the computational time. An almost real-time computation can be gained without any loss in performance [15]. This computation improvement is caused using integral images [16], which drastically reduce the number of operations in the filtering step (Figure 1). Agrawal et al. [17] introduced scale-invariant centre-surround detectors CenSurE, which are the newest scale-space detectors. These detectors give almost the same accuracy as SURF and SIFT detectors, but are even faster computationally than SURF. To achieve this performance, CenSurE computes all features at all scales and selects the extremes across each scale using a bi-level kernel as the centre-surround filter. In addition, CenSurE achieves full spatial resolution at every scale. It also uses the Harris operator for edge filtering and takes advantage of an approximation to the Laplace operators for better scale selection.

Features from an image which is independent from scale, rotation and lighting invariants in the scene extracted from information around a keypoint are called descriptors. Once the keypoint is found, neighbouring information of the keypoint can be extracted to uniquely identify each keypoint with respect to the local image patch. These descriptors are highly distinctive, and they are resistant to illumination change and pixel variation. Basically, the descriptors show how do the intensities are distributed around the neighbouring of each keypoint. The SIFT and SURF are two well-known methods used for extracting the descriptors. In the SIFT, local image gradients are measured at different selected scales in a region around each keypoint to extract the descriptors [18].

The SURF uses a similar approach to the SIFT, but instead of gradients, integral image with Haar wavelets filter are used. It is to speed up the extraction time and improve its robustness. The Haar wavelets act as simple filters to find gradient in the x and y directions, as illustrated in Figure 1a. On the other hand, the integral image will significantly decrease computational time of the gradients in which only four memory accesses and four summation operations involve (Figure 1b).

To determine the orientation of each feature, the Haar wavelets responses within certain radius area of each keypoints are calculated. Then, x and y responses of each area are summed to form a new vector in which the longest vector will show the orientation of each interest point. The descriptor components are extracted based on a square window built around the interest point. This window is then divided into 4 × 4 sub-windows in which each sub-window has four features from Haar wavelet calculated as dx, dy , |dx| and |dy|. In total, 64 features 4 × 4 × 4 are extracted for each keypoint where each feature is invariant to rotation, scale, brightness.

3. Our Approach

As mentioned previously, pixel independence is the main drawback in almost all BGS techniques because BGS algorithm does not consider the neighbouring pixels in the modelling stage. Obviously, they became sensitive to

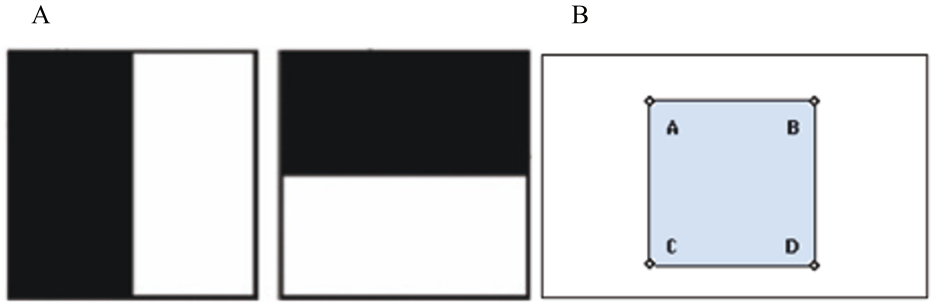

Figure 1 **(a) Harr responds on the *y*-direction (left) and on the *x*-direction (right); (b) If we consider a rectangular box with A, B, C, and D vertices, it only takes four memory accesses to calculate the sum of intensities inside the box as ∑A + D - (C + B)**.

environmental challenges, such as illumination changes and shadow effects. Scale-invariant features prove to have accurate results in various lighting conditions and scaling changes (Figure 2). Therefore, in this study, we combine a simple background difference model with the newly proposed scale-invariant centre-surround detectors (CenSurE) to decrease the difficulty of the BGS model.

As one of the states-of-art scale-invariant feature detectors, the CenSurE is chosen for matching correspondence between two images of the same scene. The CenSurE computes a simplified box filtering using integral images, as illustrated in Figure 1, at all locations with different scale-spaces. The scale-space is a continuous function which is used to find extrema across all possible scales. To achieve real-time performance, the CenSurE performs the scale-space by applying bi-level kernels to the original image.

In our approach, rather than modelling the pixel intensity to obtain a foreground mask, we use image features and their descriptor to extract significant changes in the scene. This model is divided into three main module names: modelling, matching and thresholding (Figure 3).

3.1. Modelling

The first stage of this system deals with setting the background in the scene which is similar to all other BGS techniques. Unlike traditional background modelling, which deals with all the pixels in the frame without considering their neighbouring pixel, only the selected area of the keypoints of interest and their neighbouring pixels are considered in this system. The general flow diagram of the proposed model is as shown in Figure 3.

Before modelling the background based on keypoints, we first need to initialize the background. Median filtering is a non-recursive approach that is widely used in background initialization. This model assumes that the background is more likely to appear in a scene in the sequence

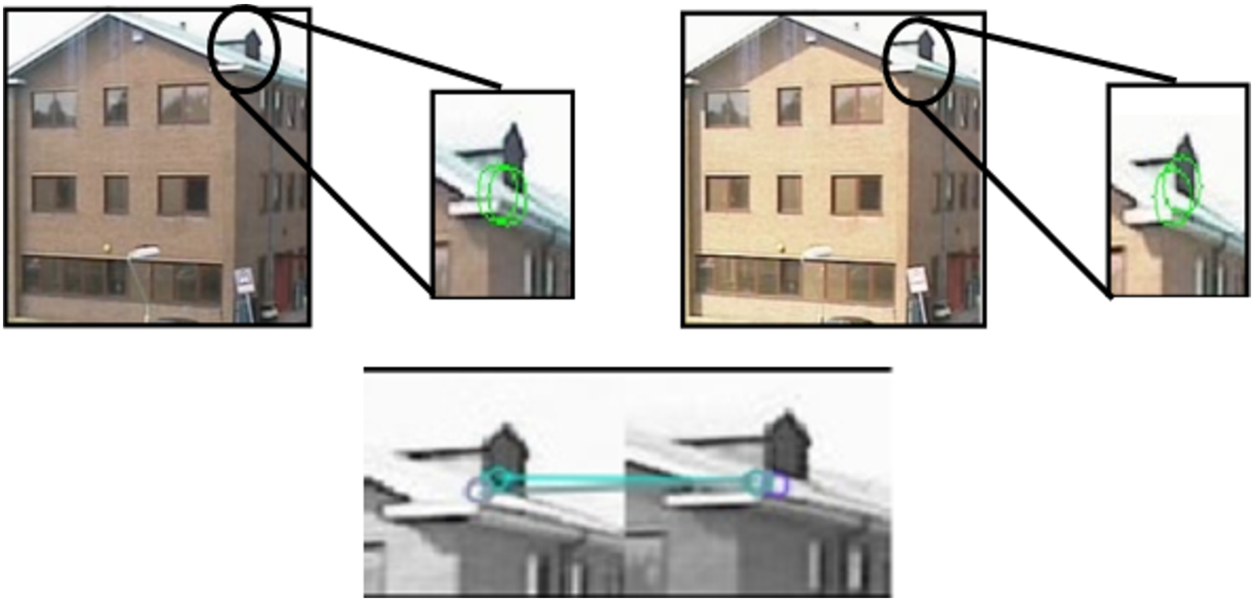

Figure 2 **Point detection in different lighting conditions using the CenSurE detector**.

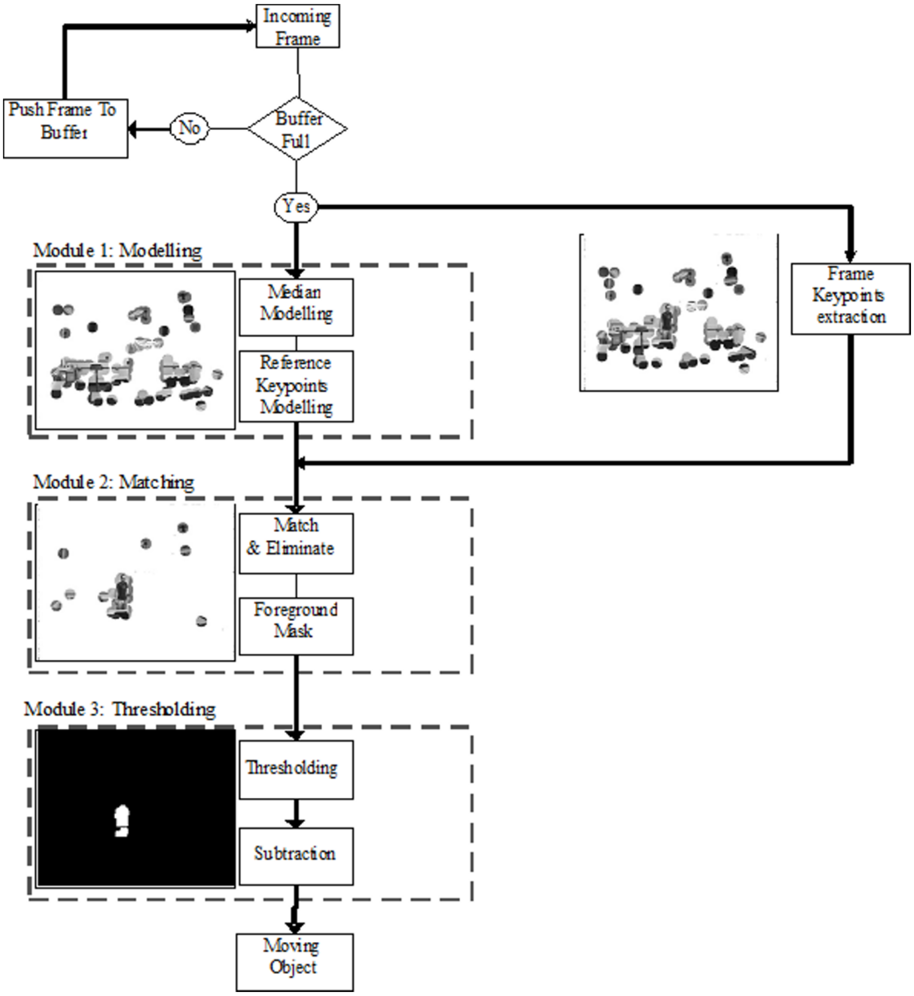

Figure 3 General flow diagram of the proposed model.

of frame, so it uses the median of the previous n frames I as the background model (Equation 3).

$$B_t(x;y) = \text{Median} \sum_{n=t-\Delta t}^{n=t} I_n(x;y) \tag{3}$$

In median filtering, the correct selections of the buffer size n and frame time rate Δt are critical issues that affect the performance of median filtering. It has been shown by Cucchiara et al. [7] that with proper selection of the observation time window ($n\Delta t$), median filtering gives the best overall performance for real-time application as compared to mean and mode filtering.

After building the reference background, we need to extract a significant keypoint from the reference image. To achieve this goal, the CenSurE detector is applied to both backgrounds as well as the incoming frame to extract a reference keypoint K_r and frame keypoint K_f as shown by module 1 in Figure 3. Because the keypoint itself is not efficient enough to give us information about the scene and the lighting condition, SURF descriptors have to be extracted to gain a more stable and recognisable point.

3.2. Matching

With given reference and frame descriptors, we can compare and match this descriptor to find any changes in the scene. Here, we have used a simple brute force matcher technique that simply matches the descriptor in one set with the closest descriptor in another set by making a guess for each one based on a distance metric. Results of the implementation are shown in Figure 4.

To achieve maximum elimination, we use a Euclidean distance with a high value to assign the most probable incoming feature to correspond to the reference feature. After matching all possible descriptors, we are now able to eliminate unwanted keypoints and their neighbours to locate an area of interest in the incoming frame based on Equation 4, where D_r and D_f represent the reference

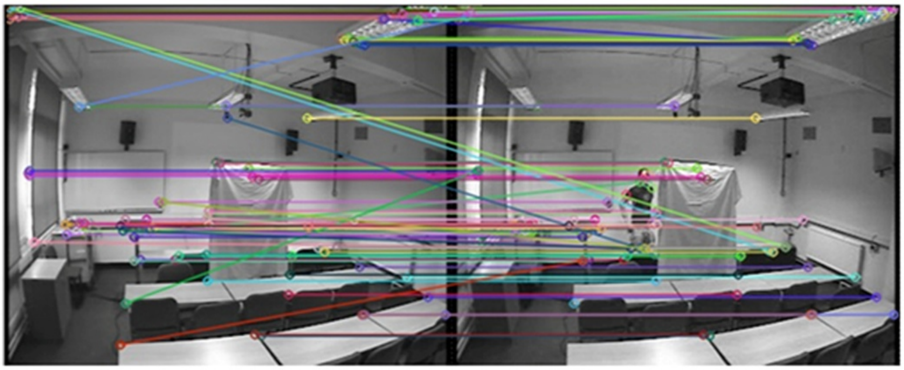

Figure 4 Matching result from brute force matcher.

descriptor and frame descriptor, respectively. This is done in module 2 shown in Figure 3.

$$k_m = k_\mathrm{r} - k_\mathrm{f} \text{ If } D_\mathrm{r} \rightarrow D_\mathrm{f} \quad (4)$$

3.3. Thresholding

After going through the procedure of module 2 in Figure 3, there are still some false blobs coming from the matching module. Thus, for each blob, a local thresholding method is applied to remove them using certain threshold values. For this experimental study, the threshold values are manually set and they are greyscale values varying between 40 and 50.

The local thresholding is one of the techniques which can be useful particularly when the scene illumination varies locally over time [19]. In modules 2 and 3, the interest's pixels and their neighbouring areas are masked so that vast amount of pixel intensity from each frame can be automatically eliminated. Correspondingly, using global thresholding over this mask, we can obtain the same result as the local thresholding.

4. Comparison and discussion

In this article, we have proposed a new method for moving object detection using a keypoint model and compared it to the GMM [1,2,5,10], which is considered to be one of the best BGS models available. The Intel (R) core (TM) i7-960 @ 3.2 GHz CPU with 5 GB RAM is chosen as the hardware platform. Algorithm implementation is done using a C-based computer vision library, "OPENCV," to carry out real-time performance for these two models.

The datasets are selected from the Internet, based on various challenges of indoor and outdoor environments such as camera variation, lighting difference and shadow effect (Figure 5) The ground truth data were segmented manually with the help of Photoshop and Adobe after the effect. The sample visual result of our comparison can be seen in Figure 5.

To produce a quantitative analysis, 11 frames were selected randomly from each dataset and the following measures comprising: false positive (FP), false negative (FN) and percentage of correct classification (PCC) are computed for each dataset. The FP parameters represent the accuracy of correct detection of a changed pixel in the frame. Conversely, a FP, FN or false alarm rate shows the number of changed pixels that incorrectly detect as no change, and finally, the PCC represents the overall rate of correct detection, which can be determined from FN and FP according to Equation 5:

$$\mathrm{PCC} = \frac{(\mathrm{CD})}{(\mathrm{CD} + \mathrm{FP} + \mathrm{FN})} \quad (5)$$

In Equation 6, CD is correct detection and can be calculated as:

$$\mathrm{CD} = \text{Total pixel} - (\mathrm{FN} + \mathrm{FP}) \quad (6)$$

Here, we discuss different properties of GMM and a keypoint model based on the final results from Table 1 and Figure 6. For the purpose of comparison, all quantity values (FN, FP and PCC) are normalized based on the size of the image databases 384 × 288.

1. GMM uses weighted Gaussian distributions of pixels over sequences of a frame. Therefore, it is not able to properly handle the condition where unwanted noise abides in the scene for a long period of time. This drawback can be seen from the shadow effect database where the shadow stays in the video for a long period of time. As a result, the FN rate of GMM became twice as larger than FN rate of the keypoint model.
2. In the last case, both the algorithms were tested under different lighting conditions, and there are once more big FP rate differences between these two methods with a value of 0.001700 for the keypoint model and 0.019453 for the GMM algorithm. The reason

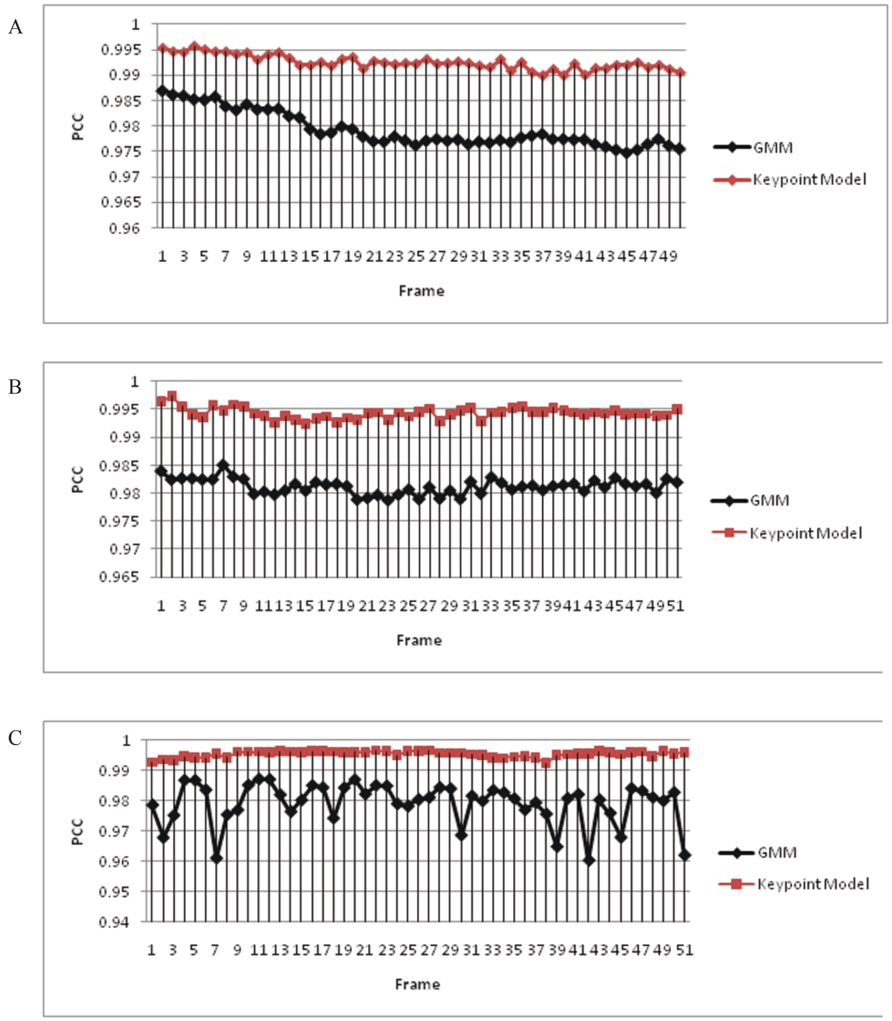

Figure 5 (a) Shadow effect (the first and second row shows the original image and ground truth consecutively, and the third and last row shows the GMM and keypint model's output). **(b)** Waving tree(the first and second row shows the original image and ground truth consecutively, and the third and last row shows the GMM and keypint model's output). **(c)** Lighting difference(the first and second row shows the original image and ground truth consecutively, and the third and last row shows the GMM and keypint model's output).

behind this improvement can be found in the thresholding technique used by the keypoint model, which treats each blob independently from the others and adjusts the thresholding parameter with respect to the intensity value of each individual blob.

Table 1 FP and FN rate for three selected databases

Frame/scenario	Average FP	Average FN	Average PCC
GMM			
Waving tree	0.016708	0.002135	0.981155
Shadow	0.004854	0.015982	0.979178
Light	0.019453	0.001347	0.979194
Keypoint model			
Waving tree	0.001345	0.004337	0.994320
Shadow	0.004988	0.002526	0.992494
Light	0.001700	0.002865	0.995441

3. The graphs from Figure 6a-c illustrate the PCC for a waving tree, shadow effect and lighting difference consecutively. As these graphs show, the proposed model gives accuracy improvements in all three cases with 99.2% in shadow effect, 99.4% in waving tree and 99.5% in lighting difference.

4. In addition, as the graph in Figure 6 shows, the keypoint model gives a more stable performance in comparison with GMM, with less variation in PCC rate.

5. From Figure 5, it can be observed that qualitatively the GMM gives comparable or slightly better pixel recognition results. However, in some cases that the

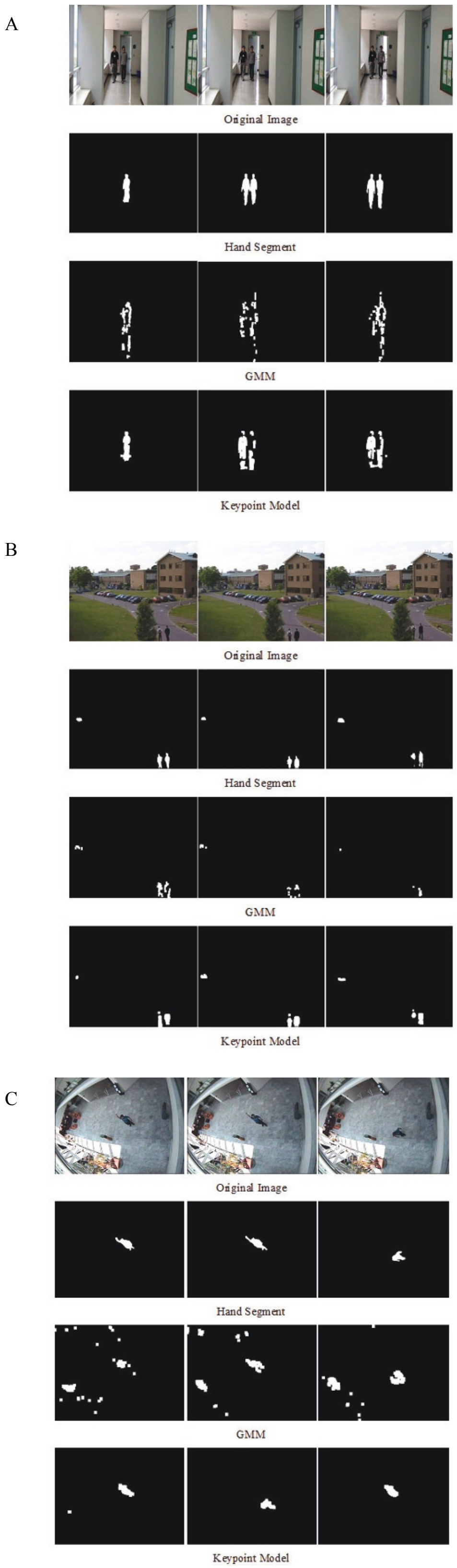

Figure 6 PCC for three different scenarios. **(a)** Shadow effect. **(b)** Waving tree. **(c)** Lighting difference.

Table 2 Computational comparison between keypoint model and GMM

				Computational time frame per second (fps)	
Video	**Number of frame**	**Format**	**Size**	**Keypoint**	**GMM**
Waving tree	735	Avi (Xvid–mpeg4)	640 × 512	4.4fps	2.4fps
Light	600	Avi (Xvid–mpeg4)	384 × 288	12.8fps	7fps
Shadow	414	Avi (Xvid–mpeg4)	640 × 512	8.3fps	2.4fps

pixels are not compact, the object recognition or tracking is not good as our proposed keypoint model.

6. Table 2 presents the computational comparison of the keypoint model and the GMM, in which the proposed model gives better computational speed. For the first two cases (waving tree and shadow effect) and the last case (lighting difference) respectively, the keypoint models are 1.8 and 3.5 faster than the GMM.

7. Speed of the keypoint model is dependent on the number of keypoints recognized in the scene and is not based on individual pixels. Thus, the data from Table 2 prove that the keypoint model gives more variant computational speed in different cases due to the nature of this algorithm.

5. Conclusion and future work

In this article, we have presented a keypoint reference model for object detection under various conditions. For the purpose of comparison, we investigated the proposed method with the well-known GMM in three challenging situations: pixel variation, illumination changes and a shadow effect. The overall evaluation shows that the keypoint modelling gives higher accuracy in all the different situations because of the reduction of TP and FN error rates.

This improvement is achieved by two main factors. First, through the use of keypoint model that considers the pixel dependency in the modelling stage. Hence, it is less sensitive to illumination changes and shadow effects. Second is due to the fact that the individual blob thresholding technique used by the keypoint model significantly helps reduce the FP rate in the final stage. The fastest and more accurate model can be gained by combining the newest matching technique and faster descriptor extractor with that in a specific environment. In addition, machine learning can be used to improve the matching accuracy.

Acknowledgements
This research was supported in part by the eScience Fund grant 01-01-02-SF0563 by MOSTI and OUP-UKM-ICT-36-184/2010 grant from the Centre for Research & Innovation Management (CRIM) of Universiti Kebangsaan Malaysia (UKM).

Competing interests
Mohammad Hedayati receives financial support from the eScience Fund grant by MOSTI.

Received: 3 March 2011 Accepted: 3 October 2011
Published: 3 October 2011

References

1. Cristani M, Farenzena M, Bloisi D, Murino V: **Background subtraction for automated multisensor surveillance: a comprehensive review.** *EURASIP J Adv Signal Process* 2010, Article ID 343057, 24 (2010).
2. Lopez-Rubio E, Luque-Baena RM: **Stochastic approximation for background modelling.** *Comput Vis Image Understand* 2011, **115(6)**:735-749.
3. Wren C, Azarhayejani A, Darrell T, Pentland AP: **Pfinder: real-time tracking of the human body.** *IEEE Trans Pattern Anal Mach Intell* 1997, **19(7)**:780-785.
4. Bouttefroy PLM, Bouzerdoum A, Phung SL, Beghdadi A: **On the analysis of background subtraction techniques using Gaussian mixture.** *Acoustics Speech and Signal Processing (ICASSP)* 2010.
5. Stauffer C, Grimson WEL: **Adaptive background mixture models for real-time tracking.** *IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'99)* 1999, **2**:2246.
6. Elgammal AM, Harwood D, Davis LS: **Non-parametric model for background subtraction.** *ECCV* 2000, 751-767.
7. Cucchiara R, Grana C, Piccardi M, Prati A: **Statistic and knowledge-based moving object detection in traffic scenes.** *2000 IEEE Proceedings of Intelligent Transportation Systems* 2000.
8. Cucchiara R, Grana C, Piccardi M, Prati A: **Detecting moving objects, ghosts and shadows in video streams.** *IEEE Trans PAMI* 2003, **25(10)**:1337-1342.
9. McFarlane N, Schofield C: **Segmentation and tracking of piglets in images.** *Mach Vis Appl* 1995, **8(3)**:187-193.
10. Moeslunda TB, Hiltonb A, Krügerc V: **A survey of advances in vision-based human motion capture and analysis.** *J Comput Vis Image Understand* 2006, **104(2)**:90-126.
11. Harris C, Stephens M: **A combined corner and edge detector.** *Proceedings of the 4th Alvey Vision Conference* 1988, 147-151.
12. Lowe DG: **Object recognition from local scale-invariant features.** *Proceedings of the International Conference on Computer Vision* 1999, **2**:1150-1157.
13. Mikolajczyk K, Schmid C: **An affine invariant interest point detector.** *Proceedings of the 8th International Conference on Computer Vision, Vancouver, Canada* 2002.
14. Mikolajczyk K, Schmid C: **Scale & affine invariant interest point detectors.** *Int J Comput Vis* 2004, **60(1)**:63-86.
15. Bay H, Ess A, Tuytelaars T, Van Gool L: **SURF: speeded up robust features.** *Comput Vis Image Understand* 2008, **110(3)**:346-359.
16. Viola P, Jones M: **Rapid object detection using a boosted cascade of simple features.** *CVPR* 2001, **1**:511.
17. Agrawal M, Konolige K, Blas MR, CenSur E: **center surround extremas for realtime feature detection and matching.** In *ECCV LNCS. Volume 5305.* Edited by: Forsyth D, Torr P, Zisserman A. Springer, Heidelberg, 2008; 2008:102-115, Part IV.
18. Lowe DG: **Distinctive image features from scale-invariant keypoints.** *Int J Comput Vis* 2004, **60(2)**:91-110.
19. Rosin P, Ioannidis E: **Evaluation of global image thresholding for change detection.** *Pattern Recogn Lett* 2003, **24(14)**:2345-2356.

doi:10.1186/1687-5281-2011-13
Cite this article as: Wan Zaki *et al.*: **Moving object detection using keypoints reference model.** *EURASIP Journal on Image and Video Processing* 2011 **2011**:13.