Motion Clustering with Hybrid-sample-based Foreground Segmentation for Moving Cameras

Yi-Chan Wu[1] and Ching-Te Chiu[2]
Email: xdd3732@gmail.com[1], ctchiu@cs.nthu.edu.tw[2],

***Abstract*—Foreground segmentation/background subtraction is a vital step in many high-level video analysis applications. While many methods have been proposed for foreground segmentation, most assume the cameras to be stationary. With this assumption, they are unable to handle the movements caused by camera rotation. In this paper, we propose a robust hybrid-sample-based foreground segmentation method for moving cameras, and especially for pan–tilt–zoom cameras. First, we propose the use of motion clustering registration to reduce the impact of registration errors. Next, we propose a frame-level reinitialization scheme to solve the problem of sudden large movement between consecutive frames. Third, we adopt a hybrid-sample-based background modeling technique to easily detect camouflaged foreground objects. Lastly, in order to deal with dynamic backgrounds, we propose moving scene pixel-level feedback schemes to dynamically and locally control the sensitivity and adaptation speed of the background model. We evaluate the proposed method using the ChangeDetection.NET 2014 dataset. Experimental results show that our proposed motion clustering registration can eliminate most of the noise caused by registration errors. The proposed reinitialization scheme can handle the noises caused by sudden large movements. The proposed method performs at least 8% better than other state-of-the-art algorithms in terms of the *F-score* in the pan–tilt–zoom camera scenario, and it also achieves the highest *F-score* in camera jitter scenarios.**

***Index Terms*—Foreground segmentation, Moving camera, Optical flow, Motion Clustering, Adaptive feedback**

I. Introduction

In recent years, the pan–tilt–zoom (PTZ) camera has gained in popularity because of its rotation flexibility, which provides a broader view. Although many researchers have addressed background subtraction, most have assumed the camera to be fixed. In such an approach, each pixel represents only one fixed position of the scene. Therefore, variations of each pixel can be modeled independently. However, in PTZ cameras which can rotate, each pixel can represent more than one scene position. This means that fixed-camera approaches can not directly apply to PTZ cameras. Recently, several approaches specific to PTZ cameras have been proposed. These methods can be roughly divided into two categories: frame-to-global and frame-to-frame. Frame-to-global methods [1], [2], [3] build a panoramic background model by stitching together several frames that cover the whole monitored scene. When the current frame aligns with the panoramic background model, the segmentation step can be easily processed in a similar way to that for a fixed camera. However, these approaches may be sensitive to internal camera parameters and are often characterized by heavy distortion. Frame-to-frame methods [4], [5], [6] focus on the reuse of overlapping regions in consecutive frames. The transformation between consecutive frames is estimated by using pairs of feature points. Although this approach results in less distortion, registration error may still be a problem. Kim *et al.* [5] proposed a single spatiotemporally distributed Gaussian model that uses the spatial information around the pixel to reduce the registration error. However, it is vulnerable to corruption of the background model. In addition, the alignment between consecutive frames can be heavily distorted from the occurrence of any sudden large movement. This is because the regions between consecutive frames are not sufficiently overlapped to accurately estimate the transformation.

Several methods have been developed that dynamically adjust the background model parameters [7], [8]. However, most of these are designed for a stationary camera. With a PTZ camera rotation, the observed frame comprises the newly covered region and the overlapped region. The newly covered region may include foreground objects that cause an improper initialization of the background model. Then, the false foreground region detected by this flawed background model can cause a suspension of the adaptation process. Ultimately, the flawed background model will not be able to be quickly recovered by the adaptive feedback mechanism without considering pixel stability. The adaptation speed of a pixel in an unstable region should be higher than that in a stable region.

This paper is organized as follows: In section II, we present a flowchart of our proposed hybrid-sample-based foreground segmentation for the moving camera. We also introduce details of the four main aspects of our proposed method, including hybrid-sample-based pixel-level modeling, motion clustering registration, an adaptive feedback scheme, and a reinitialization scheme. Next, in section III, we present the experimental results of our proposed method with respect to both a PTZ camera and a jittery camera for the ChangDetection 2014 dataset [9], and compare our results with those from state-of-the-art foreground segmentation methods. Finally, we draw our conclusions in section IV.

II. Proposed Motion Clustering with Hybrid-sample-based Foreground Segmentation for Moving Cameras

In this section, we present our proposed motion clustering with hybrid-sample-based foreground segmentation method for moving cameras which focuses especially on PTZ cameras. Figure 1 shows a flowchart of our proposed method. In the

subsections below, we describe the operation flow of the components in each module. First, to detect subtle changes and camouflaged moving objects, we include extra gradient texture information, as the authors do in [8]. We present the detail of our hybrid-sample-based pixel-level modeling in section II-A, and present our motion clustering technique in section II-B. This technique solves the problem of the corruption of the background model associated with the Kim *et al.* [5] method. In section II-C, we propose our reinitialization scheme to handle sudden large movements. Finally, in section II-D, we outline our classification steps in detail and explain how the feedback scheme controls the sensitivity and adaptation speed of the model.

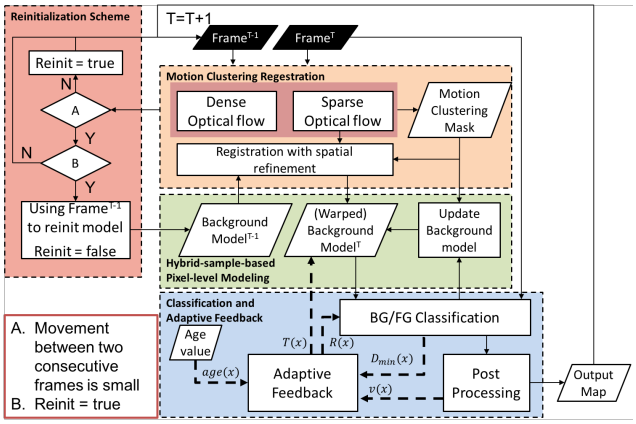


Fig. 1. Detailed flowchart of motion clustering with hybrid-sample-based foreground segmentation

A. Hybrid-sample-based Pixel-level Modeling

Each pixel-level background model is directly characterized by a collection of K background samples:

$$B^t(p) = \left\{B_1^t(p), B_2^t(p), \cdots, \cdots, B_K^t(p)\right\}, \quad (1)$$

where $B^t(p)$ represents the background model of a pixel p at time t. Each background sample B_i^t at pixel p is composed of three factors:

$$B_i^t(p) = \left\{B_i^{rgb}(p), B_i^{gradient}(p), B_i^{lbsp}(p)\right\}, 1 \le i \le K. \quad (2)$$

We add extra gradient texture information to the approach as [8]. A background sample is composed of an rgb color intensity value B_i^{rgb}, gradient magnitude $B_i^{gradient}$, and modified local binary similarity patterns (LBSP) B_i^{lbsp}.

We use the same assumption used in [10], that the neighboring pixels share a similar temporal distribution. Therefore, we can initialize our background model with a single frame.

B. Motion Clustering Registration (MCR)

To detect moving objects with non-stationary cameras, alignment is necessary between the observed frame and a background model. In many studies, researchers have built a panoramic background model and aligned the observed frame with that background model. However, these frame-to-global methods typically have several drawbacks. First, the initialization of the panoramic background model is time-intensive because all the angles of the monitor scene must be captured by rotating the PTZ camera. Secondly, the results may be heavily distorted if no camera calibration parameters are provided.

Therefore, here, we adopt a frame-to-frame approach similar to that in [5] and solve the problem of the corruption of the background model by the proposed motion clustering registration (MCR) technique. As shown in Fig. 1, the MCR module uses two types of optical flow (dense and sparse) to produce a motion clustering mask (MCM) and aligns two consecutive frames in the spatial refinement step in the registration process. Spatial refinement can reduce the impact of registration errors and the MCM can prevent background model corruption.

First, we extract the Harris corner feature points [11] from the $frame^{t-1}$. Next, we track the feature points using the Lucas and Kanade algorithm [12], and use the pairs of feature points between $frame^{t-1}$ and $frame^t$ to calculate the homography matrix H. We then use the random sample consensus (RANSAC) algorithm [13] to eliminate the outliers that would affect the accuracy of the transformation. We denote the sparse optical flow at pixel p as $OF_{sparse}(p)$. To calculate the dense optical flow, we use an efficient algorithm [14], which adopts the conjugate gradient for solving large linear systems. The dense optical flow of the pixel p is denoted as $OF_{dense}(p)$. The difference between the sparse and dense optical flows at pixel p is denoted as $OF_{diff}(p)$.

The sparse optical flow estimates the global transformation between consecutive frames by using feature points that mostly belong to the background region. The feature points that belong to moving objects are not considered because they are mostly eliminated by the RANSAC algorithm [13]. The dense optical flow estimates the movement between consecutive frames at every pixel. Therefore, in a static background pixel, the difference between the sparse and dense optical flows should be close to zero. On the other hand, if a pixel belongs to a moving object, the difference would increase. We use this difference to create the MCM:

$$MCM(p) = \begin{cases} 1 & ||OF_{diff}(p)|| > T_{mc} \\ 0, & otherwise \end{cases}, \quad (3)$$

$$||OF_{diff}(p)|| = \sqrt{u_{diff}^2 + v_{diff}^2}, \quad (4)$$

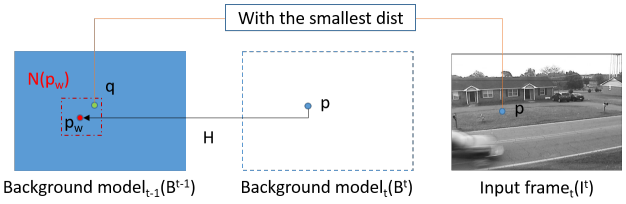
where T_{mc} is the motion clustering threshold and is a constant. If the difference between the two optical flows at pixel p is larger than the threshold T_{mc}, then this pixel may be located on a moving object.

Next, we improve the registration method proposed by the authors in [5] by adding the MCM to prevent background model corruption. The following equations show the process of MCR:

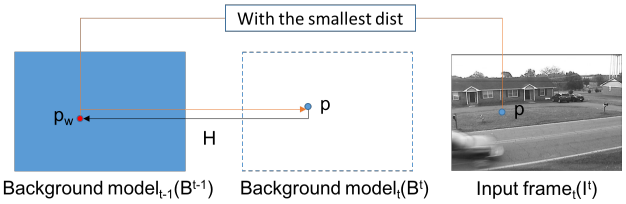
$$B^t(p) = \begin{cases} B_{init}(B_I^t(p_n)) & p\ in\ NCR \\ B^{t-1}(p_w) & p\ in\ OR, MCM(p) = 1 \\ B^{t-1}(q) & p\ in\ OR, MCM(p) = 0 \end{cases}, \quad (5)$$

$$q = argmin_{p_{wn} \in N(p_w)} \sum_{i=0}^{K} dist(B_I^t(p), B_i^{t-1}(p_{wn})), \quad (6)$$

where $B^t(p)$ denotes the background model of pixel p at time t, $B_I^t(p)$ denotes the observed sample at pixel p, $dist(B_I^t(p), B_i^{t-1}(p_{wn}))$ is the distance between the current observed sample $B_I^t(p)$ and a given background sample $B_i^{t-1}(p_{wn})$, and q denotes the position of the most similar background model $B^{t-1}(q)$ to the current observation. As mentioned above, the observed frame is composed of the newly covered region (NCR) and the overlapped region (OR). For a pixel p in the newly covered region, we directly initialize the background model at time t by the current neighboring observed samples $B_I^t(p_n)$, since there is no useful background model at time $t-1$. On the other hand, for a pixel p in the overlapped region, we use MCR to warp the background model. Figure 2 illustrates our proposed MCR process. If the value of the pixel p in the MCM is zero and pixel p is in the overlapped region, we select the background model $B^t(p)$ from the neighborhood of the previous background model $B^{t-1}(p_w)$. As shown in Fig. 2(a), we select the neighboring background model $B^{t-1}(q)$ that is most similar to the current observation. This process is called spatial refinement which can reduce the number of false detections caused by registration error. On the other hand, in Fig. 2(b), if pixel p is in the overlapped region and the value of pixel p in the MCM is one, which means this pixel may belong to a moving object, to prevent background model corruption we simply use the background model $B^{t-1}(p_w)$ calculated by the homography transform without spatial refinement.



(a) Spatial refinement



(b) Direct warping

Fig. 2. Illustration of proposed motion clustering registration process. (a) Spatial refinement and (b) direct warping

C. Reinitialization Scheme

Homography transformation cannot handle a sudden large movement between consecutive frames because the overlapped region is insufficient to correctly estimate the transformation. Therefore, when there is a sudden large movement, it is impossible to compare a currently observed frame to the background model since the alignment between frames is not accurate. As shown in Fig. 1, we propose a reinitialization scheme (left red module). The detail flow of our reinitialization scheme is shown in Fig. 3.

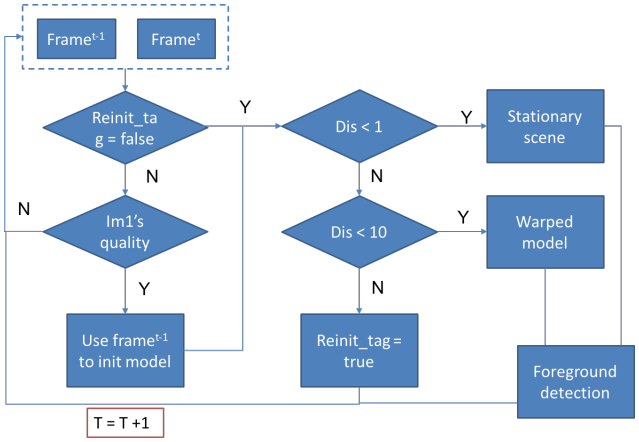


Fig. 3. Reinitialization scheme

We use the number of feature points to check the frame's quality by following equation:

$$|fp^{t-1}| \geq \frac{2 \cdot |fp^t|}{3}, \quad (7)$$

where $|fp^{t-1}|$ and $|fp^t|$ denote the number of feature points in frame at time $t-1$ and t, respectively. Once the equation is satisfied, we use $frame^{t-1}$ to reinitialize the background model.

D. Classification and Adaptive Feedback

Based on its similarity to background samples, an observed pixel is classified as either foreground or background. For this purpose, we adopt a color-LBSP classification process similar to that used in [8]:

$$S^t(p) = \begin{cases} 0 & if\ p\ in\ newly\ covered\ region \\ 0 & \sharp\{dist(B_I^t(p), B_i^t(p)) < R^t(p), \forall i\} \geq \sharp_{min} \\ 1 & otherwise \end{cases}, \quad (8)$$

where $S^t(p)$ is the output segmentation result of pixel p at time t, $R^t(p)$ is the pixel-level threshold of the pixel p which indicates the model sensitivity at time t, $\sharp_{min}$ is the minimum number of matches required for a background classification, $B_I^t(p)$ is the current observed sample, and $dist(B_I^t(p), B_i^t(p))$ returns the distance between the current observed sample and a given ith background sample at pixel p. If the pixel is located in the newly covered regions, it is classified as background because there is no useful background model in previous frames. However, if the pixel is located in the overlapped regions, the classification is made based on the color-LBSP distance. Once the background pixel p is detected, the current observed sample $B_I^t(p)$ has $1/T^t(p)$ probability of replacing a randomly selected sample of the background model $B^t(p)$, where $T^t(p)$ is the adaptation rate or a "time

subsampling factor,"" as in [10]. A randomly selected sample of one randomly selected neighboring background model of $B^t(p)$ also has $1/T^t(p)$ probability of being replaced by the current observed sample $B_I^t(p)$.

There are two parameters in our hybrid-sample-based background model—the decision threshold $R^t(p)$ and the adaptation rate $T^t(p)$. To control these two parameters in moving camera scenarios, we modified the feedback scheme from that presented in [8]. We propose the age value $age^t(p)$ for controlling the adaptation rate in newly covered regions. When the PTZ camera is moving from left to right, the age values in the overlapped regions increase and the age values in newly covered regions are initialized to 1. The age values are limited to [1, 5]. We weight the age values that the smallest age value 1 has the greatest weight 5 and the biggest age value 5 has the lowest weight 1.

Then, we adjust the decision threshold $R^t(p)$ and the adaptation rate $T^t(p)$ by the following equations:

$$R^t(p) = \begin{cases} R^t(p) + v^t(p) & R^t(p) < (1 + D^t_{min}(p) \cdot 2)^2 \\ R^t(p) - \frac{1}{v^t(p)} & otherwise \end{cases}, \tag{9}$$

$$T^t(p) = \begin{cases} T^t(p) + \frac{1}{v^t(p) \cdot D^t_{min}(p)} & S^t(p) = 1 \\ T^t(p) - \frac{v^t(p) \cdot ages^t_{weight}(p)}{D^t_{min}(p)} & S^t(p) = 0 \end{cases}, \tag{10}$$

We use age^t_{weight} to increase the adaptation speed in newly covered regions (unstable regions). More unstable regions have a higher adaptation speed, which can help us to quickly recover the corrupted background model caused by its improper initialization. Please refer to [8] for more details about background dynamic $D^t_{min}(p)$ and blinking pixel monitoring value $v^t(p)$.

III. Experimental Results

To properly evaluate the performance of our method, we used the ChangeDetecion.net (CDnet) 2014 dataset [9] as a benchmark, and focused on non-stationary camera events. In these scenarios, we tested two dataset categories including the *pan-tilt-zoom camera* and *camera jitter* categories. We examined five video sequences for a total of 4630 frames. We compared the proposed method with other methods that are also performing well in these two categories, and used the *F-score* to compare the performance of the different methods. Table I shows the average performance of the *Pan-tilt-zoom Camera* category, for which our method has the highest *Precision* and *F-score*. Table II shows the average performance in the *Camera Jitter* category with a total of 2380 frames. Our method yielded the highest *F-score* and *Recall*. Table III shows the average performance of the five sequences with 4630 frames in which our method produced the best performance and our detection results were more robust especially in camouflaged foreground regions. The noise caused by camera movement and jitter can be reduced by our proposed MCR.

TABLE I
AVERAGE PERFORMANCE COMPARISON OF DIFFERENT METHODS FOR *Pan-tilt-zoom Camera* CATEGORY (2250 FRAMES)

	Precision	Recall	F-score
C-EFIC[15]	*0.687893*	0.830654	*0.752562*
EFIC[16]	0.434560	*0.899214*	0.585949
IUTIS[17]	0.247574	0.614338	0.352922
UBSS[18]	0.408079	0.679889	0.510030
SuBSENSE[8]	0.174165	0.805984	0.286434
Proposed	*0.782657*	*0.896596*	*0.835761*

TABLE II
AVERAGE PERFORMANCE COMPARISON OF DIFFERENT METHODS FOR *Camera Jitter* CATEGORY (2380 FRAMES)

	Precision	Recall	F-score
C-EFIC[15]	*0.804747*	0.855831	*0.829503*
EFIC[16]	0.747409	*0.861320*	0.800332
IUTIS[17]	0.796539	0.806826	0.801649
UBSS[18]	*0.800874*	0.749354	0.774258
SuBSENSE[8]	0.787556	0.769626	0.778488
Proposed	0.771843	*0.946456*	*0.850277*

TABLE III
AVERAGE PERFORMANCE COMPARISON OF DIFFERENT METHODS FOR *Pan-tilt-zoom Camera* CATEGORY AND *Camera Jitter* CATEGORY (4630 FRAMES)

	Precision	Recall	F-score
C-EFIC[15]	*0.734634*	0.840724	*0.784107*
EFIC[16]	0.559699	*0.884056*	0.685442
IUTIS[17]	0.467160	0.691333	0.557557
UBSS[18]	0.565197	0.707675	0.628462
SuBSENSE[8]	0.419521	0.791441	0.548368
Proposed	*0.778331*	*0.916540*	*0.841800*

IV. Conclusion

In this paper, we presented a robust foreground segmentation method for moving cameras. The main contribution of our method is as follows. First, the use of a motion clustering mask can reduce the effect of registration errors without corrupting the background model. Second, the reinitialization scheme allows us to control both moving and stationary scenarios. Third, the age value is useful for quickly adapting unstable background regions. Finally, the newly added gradient magnitude can improve *Recall*. The results reveal that our proposed method significantly improves PTZ camera scenarios. Our *F-score* was at least eight percent higher than other existing methods in the *Pan-tilt-zoom Camera* category. Furthermore, our method is also effective in camera jitter scenarios.

References

[1] S. N. Sinha and M. Pollefeys, "Pan-tilt-zoom Camera Calibration and High-resolution Mosaic Generation," *Comput. Vis. Image Underst.*, vol. 103, no. 3, pp. 170–183, Sep. 2006.

[2] A. Mittal and D. Huttenlocher, "Scene modeling for wide area surveillance and image synthesis," *Computer Vision and Pattern Recognition, 2000. Proceedings. IEEE Conference on*, vol. 2, pp. 160–167 vol.2, 2000.

[3] K. Xue, Y. Liu, G. Ogunmakin, J. Chen, and J. Zhang, "Panoramic Gaussian Mixture Model and large-scale

range background substraction method for PTZ camera-based surveillance systems," *Machine Vision and Applications*, vol. 24, no. 3, pp. 477–492, 2013.

[4] S. Kang, J. Paik, A. Koschan, B. R. Abidi, and M. A. Abidi, "Real-time video tracking using PTZ cameras," *Proceedings of the International Conference on Quality Control by Arficial Vision*, vol. 5132, no. May, pp. 103–111, 2003.

[5] S. W. Kim, K. Yun, K. M. Yi, S. J. Kim, and J. Y. Choi, "Detection of moving objects with a moving camera using non-panoramic background model," *Machine Vision and Applications*, vol. 24, no. 5, pp. 1015–1028, 2013.

[6] D. Zamalieva, A. Yilmaz, and J. W. Davis, "A Multi-transformational Model for Background Subtraction with Moving Cameras," *Computer Vision – ECCV 2014: 13th European Conference, Zurich, Switzerland, September 6-12, 2014, Proceedings, Part I*, pp. 803–817, 2014.

[7] M. Hofmann, P. Tiefenbacher, and G. Rigoll, "Background segmentation with feedback: The Pixel-Based Adaptive Segmenter," *2012 IEEE Computer Society Conference on Computer Vision and Pattern Recognition Workshops*, pp. 38–43, June 2012.

[8] P. L. St-Charles, G. A. Bilodeau, and R. Bergevin, "SuBSENSE: A Universal Change Detection Method With Local Adaptive Sensitivity," *IEEE Transactions on Image Processing*, vol. 24, no. 1, pp. 359–373, Jan 2015.

[9] Y. Wang, P. M. Jodoin, F. Porikli, J. Konrad, Y. Benezeth, and P. Ishwar, "CDnet 2014: An Expanded Change Detection Benchmark Dataset," *2014 IEEE Conference on Computer Vision and Pattern Recognition Workshops*, pp. 393–400, June 2014.

[10] O. Barnich and M. V. Droogenbroeck, "ViBe: A Universal Background Subtraction Algorithm for Video Sequences," *IEEE Transactions on Image Processing*, vol. 20, no. 6, pp. 1709–1724, June 2011.

[11] C. Harris and M. Stephens, "A combined corner and edge detector," *In Proc. of Fourth Alvey Vision Conference*, pp. 147–151, 1988.

[12] J. yves Bouguet, "Pyramidal implementation of the Lucas Kanade feature tracker," *Intel Corporation, Microprocessor Research Labs*, 2000.

[13] M. A. Fischler and R. C. Bolles, "Random Sample Consensus: A Paradigm for Model Fitting with Applications to Image Analysis and Automated Cartography," *Commun. ACM*, vol. 24, no. 6, pp. 381–395, Jun. 1981.

[14] C. Liu, "Beyond pixels: exploring new representations and applications for motion analysis," *Proceedings of the 10th European Conference on Computer Vision: Part III*, pp. 28–42, 2009.

[15] G. Allebosch, D. Van Hamme, F. Deboeverie, P. Veelaert, and W. Philips, "C-EFIC: Color and Edge Based Foreground Background Segmentation with Interior Classification," *Computer Vision, Imaging and Computer Graphics Theory and Applications: 10th International Joint Conference, VISIGRAPP 2015, Berlin, Germany, March 11-14, 2015, Revised Selected Papers*, pp. 433–454, 2016.

[16] G. Allebosch, F. Deboeverie, P. Veelaert, and W. Philips, "EFIC: Edge Based Foreground Background Segmentation and Interior Classification for Dynamic Camera Viewpoints," *Advanced Concepts for Intelligent Vision Systems: 16th International Conference, ACIVS 2015, Catania, Italy, October 26-29, 2015. Proceedings*, pp. 130–141, 2015.

[17] S. Bianco, G. Ciocca, and R. Schettini, "How Far Can You Get By Combining Change Detection Algorithms?" *CoRR*, vol. abs/1505.02921, 2015.

[18] H. Sajid and S. C. S. Cheung, "Background subtraction for static and moving camera," *Image Processing (ICIP), 2015 IEEE International Conference on*, pp. 4530–4534, Sept 2015.