1-1-2020

Selective subtraction for handheld cameras

Adeel A. Bhutta
Luddy School of Informatics, Computing, and Engineering

Imran Nazir Junejo
Zayed University

Hassan Foroosh
University of Central Florida

Recommended Citation

Bhutta, Adeel A.; Junejo, Imran Nazir; and Foroosh, Hassan, "Selective subtraction for handheld cameras" (2020). *All Works*. 3060.
https://zuscholars.zu.ac.ae/works/3060

IEEE Access
Multidisciplinary ⋮ Rapid Review ⋮ Open Access Journal

Received January 20, 2020, accepted February 5, 2020, date of publication February 13, 2020, date of current version March 2, 2020.

Digital Object Identifier 10.1109/ACCESS.2020.2973655

Selective Subtraction for Handheld Cameras

ADEEL A. BHUTTA[1], (Member, IEEE), IMRAN NAZIR JUNEJO[2], AND HASSAN FOROOSH[3], (Senior Member, IEEE)

[1]School of Informatics, Computing, and Engineering, Indiana University, Bloomington, IN 47408, USA
[2]College of Technological Innovation, Zayed University, Dubai 19282, UAE
[3]School of Electrical Engineering and Computer Science, University of Central Florida, Orlando, FL 32816, USA

Corresponding author: Adeel A. Bhutta (aabhutta@indiana.edu)

ABSTRACT Background subtraction techniques model the background of the scene using the stationarity property and classify the scene into two classes namely foreground and background. In doing so, most moving objects become foreground indiscriminately, except in dynamic scenes (such as those with some waving tree leaves, water ripples, or a water fountain), which are typically ''learned'' as part of the background using a large training set of video data. We introduce a novel concept of background as the objects *other than* the foreground, which may include moving objects in the scene that cannot be learned from a training set because they occur only irregularly and sporadically, e.g. a walking person. We propose a ''selective subtraction'' method as an alternative to standard background subtraction, and show that a reference plane in a scene viewed by two cameras can be used as the decision boundary between foreground and background. In our definition, the foreground may actually occur behind a moving object. Furthermore, the reference plane can be selected in a very flexible manner, using for example the actual moving objects in the scene, if needed. We extend this idea to allow multiple reference planes resulting in multiple foregrounds or backgrounds. We present diverse set of examples to show that: 1) the technique performs better than standard background subtraction techniques without the need for training, camera calibration, disparity map estimation, or special camera configurations; 2) it is potentially more powerful than standard methods because of its flexibility of making it possible to select in real-time what to filter out as background, regardless of whether the object is moving or not, or whether it is a rare event or a frequent one. Furthermore, we show that this technique is relatively immune to camera motion and performs well for hand-held cameras.

INDEX TERMS Background subtraction, object detection, image understanding.

I. INTRODUCTION

Background subtraction is the fundamental step used in many applications including object detection, tracking, action recognition, and activity recognition. Background subtraction techniques traditionally use one or more views to classify the objects (or image pixels) as either foreground or background. However, standard methods have a rigid definition of what constitutes a background, which often leads to classifying almost all moving objects as foreground, except for small persisting motions that can be learned from a training set. This strict binary classification and loss of 'intra-class separability' results in inability to model partial background or partial foreground and thus the notion of a background object being *in front* of a foreground object or a moving object belonging to the background. If scene modeling is to be made more effective, the background subtraction techniques need to ensure that the statistical models can learn partial backgrounds and thus an intra-class taxonomy is preserved; which can prove very useful in many real world applications such as video surveillance and detection and tracking in crowds.

Existing background subtraction techniques can be classified into two main categories: techniques using monocular sequences and those using stereo sequences. Our method relies on two views but does not require configuring cameras rigidly as a stereo pair. Most of the existing literature focuses on different aspects such as the statistical approach used to model the background, type of scene used (dynamic or static), the learning method applied to the training set, and the model used for the background or foreground. The background of a scene is generally defined as being *motionless* for static scenes (e.g., video conference) and *almost-motionless* for dynamic scenes (e.g., scenes which include changes such as illumination, shadows, waving tree leaves, water ripples, or fountains). Most single-view background subtraction techniques try to model the background (and the

The associate editor coordinating the review of this manuscript and approving it for publication was Yongqiang Zhao.

dynamic changes) either by modeling each pixel or different regions statistically, and then use those statistical models to detect the moving objects, known as foreground. This type of modeling requires large amount of training data for learning the statistical properties of the background. Another limitation of traditional techniques is that an object (or pixels) can either be classified as foreground or background, not both. Alternatively, stereo-based techniques rely on estimating disparity maps by rectifying the views and using similarity measures in order to estimate the background. Such disparity maps are in practice difficult to estimate in real-time, susceptible to noise and highly error prone. Also, these techniques require special camera setup and are computationally expensive. Recent algorithms have used sensors that can measure the depth of the objects accurately and are typically part of custom-designed hardware or device which is often very expensive. Furthermore, all background subtraction techniques (whether based on single view or two views) classify moving objects as foreground *indiscriminately*. Consider a case when a camera is looking at a street with multiple objects moving across the camera in both directions. The object closer to the camera occludes the object crossing behind it which, in turn, is occluding another object crossing behind and so on. Any standard background subtraction technique will consider all of the moving objects as foreground thus will not be able to selectively distinguish which moving object should be kept as foreground and which ones should be discarded. What if you are only interested in the first two objects closest to the camera, or only one object at the back, and all other objects are irrelevant. Thus, the *foreground-of-interest* is now the partial foreground while *background-of-interest* is a combination of traditional background and partial foreground. In this context, the standard definition of background is insufficient. Current background subtraction techniques fail to model such backgrounds or foregrounds.

Our technique has five novel contributions. Firstly, most background subtraction techniques require extensive training or learning of the background model using data consisting of different examples of background alone. Even when such data is available, these techniques cannot learn the partial background as defined above. We challenge the requirement of training and propose the use of a *reference plane* inducing a *base homography*, estimated using only two frames. This base homography can be used in the background subtraction of the scene when traditional technique fail, because they cannot classify a frequently occurring moving object as background. Secondly, we propose to use the actual moving objects in the scene to estimate the base homography and show how a simple *walk* (or an object in motion) can be used to define a reference plane. Thirdly, most background subtraction techniques need large amount of data to model the background (which usually ranges from several hundreds to several thousands of examples). We propose and show that the base homography can be estimated using an object in motion viewed only in two frames. Thus the presence of large amount of training data is no longer required in our method.

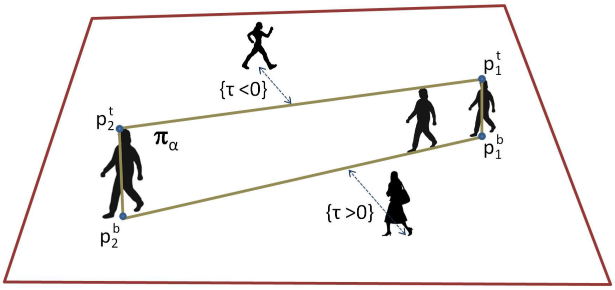


FIGURE 1. Reference plane: The reference plane is defined by a moving object or human *walk* and the projective depth (τ) is defined as the distance between the *reference plane* and the objects in the scene.

Fourthly, standard background subtraction techniques fail to change the background model once it is learned. Only some minor dynamic changes are incorporated in the updating of the background model. In our proposed technique, the base homography can be modified using a different moving object or a plane in the scene in real time, and can be replaced altogether with a new base homography, thus providing flexibility in the background subtraction. This enables us to select multiple reference planes where one object can be classified as foreground with reference to one base homography while the same object can be classified as background with reference to a different base homography. Lastly, we avoid the explicit use of depth map and the requirement of rectifying two views for calculating depth as in stereo-based methods or use of special sensors to measure depth, and propose a solution based entirely on projective depth calculated from traditional cameras. Our technique also does not require rigid camera setup and works for handheld cameras.

The rest of the paper is organized as follows. A summary of related work is presented in Section II with the theoretical formulation and the description of the proposed approach in Section III. Experimental results performed on real world sequences and brief discussion of results is presented in Section IV, followed by conclusion in Section V.

II. RELATED WORK

Background subtraction has been an active area of research over the many past decades. It is beyond the scope of this work to review all the methods and techniques, hence we refer the reader to [1], [2] for a good review of the related work in this area. Background subtraction techniques have generally used one or more views to model the pixels or regions. The idea of defining foreground as non-moving object in a static scene has been used in background subtraction and object tracking for a very long time [3], [4]. In order to improve the results in real-world scenarios, dynamic background subtraction techniques use a single three-dimensional Gaussian distribution to model each pixel in the scene [5] or a Mixture of Gaussian (MoG) [6] or a non-parametric kernel density estimation (KDE) [7]. Region based techniques have also been proposed to improve background subtraction which try to use a covariance matrix from a region around a pixel [8] or auto-regression models [9] or propose the use of temporal persistence with single probability density in a Maximum A Posteriori in the Markov Random Field (MAP-MRF) selection framework [10] to model the spatial and appearance attributes [8]. See [11]–[20] for review of other single view methods.

Deep learning, and convolutional neural network (CNN) in general, has made its impact on background subtraction as well. Based on deep learning networks, [21] performs semantic segmentation on the images. This pixel level information is leveraged for motion detection in the video sequence. Pixels with low semantic probability are deemed as background. In order to reduce any false negatives, a semantic background model is maintained at each pixel as well. In case of ambiguity, any background subtraction method can be used in their method as the final step. Dividing an input image into patch [22], SuBSENSE algorithm [23], combined with Flux Tensor algorithm [24], is applied first to create a background image. CNNs are fed with matching pairs of patches from background and the input image. For application to the field of agriculture, [25] combines a standard background subtraction method with features learned from CNNs. It is hoped that these features would be robust to camera motion and view changes, and sensitive to any new elements in the area. Pixel-wise segmentation map is computed by [26] and they proposed an encode-decoder framework where the input image is temporally aligned to the reference image. An atrous convolution is introduced by [27] to expand the receptive field of the network and, Mimicking res-net, shortcut connections are added to reduce training complexity. Conditional Random Fields (CRF) are added in the last layer for refinement. A triplet convolutional neural network is proposed by [28] which used an encoder-decoder type network while utilizing pre-trained VGG-16 Net. Each branch of the triplet network operates on different scale to perform feature encoding. Decoding is performed by the transposed convolutional network. Their method works on an image at a time, not utilizing any temporal information. In order to utilize the temporal information, [29] proposes a deep end-to-end framework where pixel-wise semantic features are extracted using an encoder-decoder network. Long Short-Term Memory method (LSTM), is then used to model pixel-wise changes over time. In order to reduce sensitivity to camera motion, conditional random fields (CRFs) are used in the last layer. In order to fully capture the temporal information of a scene. a 3D CNN is proposed by [30]. Their specific 3D-CNN consists of 6 convolutional layers and the input is a window of 10 consecutive frames. These 10 frames are divided into a group of 4 frames and fed to 4 convolutional layers. Up-sampling is performed using kernels of various strides to retain the fine details from the input images, these layers are then concatenated to produce the final predication layer.

An alternate approach and the one most related to the technique presented in this paper, is based on stereo, which attempts to recover dense disparity maps in real time for segmenting the scene. References [31] used stereo cameras and their disparity maps to perform background subtraction by checking the color intensity values of corresponding pixels. Each pixel was warped to the corresponding pixel in the reference image and the color and luminance values were used to decide if the pixel belongs to the foreground or background. This method suffers from false and missed detections. [32] proposed the use of a stereo configuration, in which the cameras are vertically aligned, to improve the background subtraction. A multi-view approach is proposed by [33] to remove static background. They propose two methods, one with rough camera localization and other with accurate camera localization. For the first method, they use scene-specific pre-trained background model (using SVMs) to perform foreground extraction. For their second approach, mutli-view stereo approach is employed to perform a dense matching (using Structure from Motion technique) of the scene with dataset of existing images to remove static background. However, scene-specific labeled trained dataset is very expensive to acquire and SfM is known to be noise prone. Out of plane object are detected by [34] and a stereo image pair is used initially to compute the planar homography between them which is done off-line. During the test phase, one image is super-imposed on the other using the pre-computed homography and then a similarity map is created. A similarity map is created to detect out of plane objects, as pixels corresponding to a background have specific values (close to 1). The background pixels, on the other hand, have low values in the similarity map. A two-view based hierarchical algorithm is proposed by [35] where stereo images are decomposed using the Discrete Wavelet Transform (DWT). Adaptive models are build over sub-bands at each level. A depth based model is also created, which is applied to pixels that do not conform to the adaptive model. However, DWT is an expensive process and is known to be effected by noise. There are two major limitations of these techniques: color and luminance is not sufficient to decide if the pixels belong to foreground or background especially when objects are roughly similar in color. Furthermore, the cameras need to be in strict configurations to have sufficient accuracy. Recently, advances in 3D depth cameras such as Microsoft Kinect have improved the accurately of depth information but these devices are typically designed and configured specifically with multiple sensors to measure depth and are often expensive to purchase and difficult to setup. Our technique does not use such devices and can work with traditional cameras including handheld cameras or cellphones.

III. SELECTIVE SUBTRACTION APPROACH

In this section, we first define selective subtraction and provide the theoretical formulation for implementing it.

A. REFERENCE PLANE π AND BASE HOMOGRAPHY

Consider a sequence of images $\{\mathbf{I_t}\}^{t=1\ldots n}$, where multiple objects are moving across the scene as shown in Figure 3. A simple change detection algorithm can be used to detect the moving objects (or blobs) and their head and feet positions

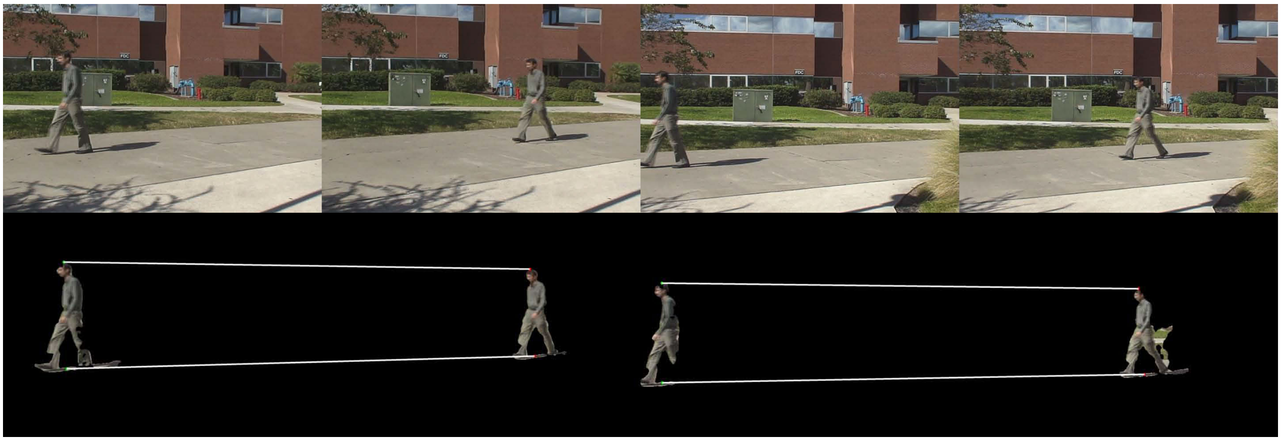


FIGURE 2. **Base homography: First row shows the images used to estimate the base homography from reference plane. The pair of images on the left shows the first and the last image of the walk (from first view) and the pair of images on the right shows the first and the last images of the walk (from second view). The second row shows the head and feet positions of the object used to estimate the base homography. It is clear that the correspondences of head and feet positions from first and the last frames only are sufficient to estimate the base homography.**

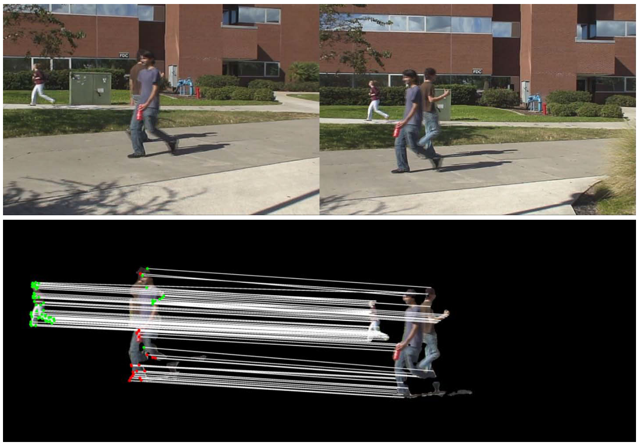


FIGURE 3. **Occlusion handling by selective subtraction method: First row shows the input images from two views where two objects are occluding each other. The *reference plane* used in these results lies in the middle of both occluding objects as seen in Figure 2 and thus both objects must fall on the opposite sides of the *reference plane*. The correspondences between feature points are shown in the second row. The projective depth of each point was calculated using the proposed technique and the points belonging to front-side are shown in red while the points lying on the other side of the *reference plane* are shown in green. The results show that the proposed technique was correctly able to estimate the projective depth even when the objects are occluded especially near the head and leg positions. For the sake of simplicity we have shown the point correspondences on the first view only.**

can be obtained by using the approach described in [36]. Let $\mathbf{P}_1$ and $\mathbf{P}_2$ be the two 3 × 4 camera projection matrices of two arbitrary cameras observing the scene. Since we do not require any calibration or a specific configuration, without loss of generality, we will model the two cameras as canonic cameras, i.e. $\mathbf{P}_1 = [\mathbf{I}, \mathbf{0}]$ and $\mathbf{P}_2 = [[\mathbf{e}']_\times \mathbf{F}, \mathbf{e}']$, where $\mathbf{F}$ is the fundamental matrix, $\mathbf{e}'$ is the epipole in the second camera view, and for any vector $v = (a, b, c)$ the notation $[v]_\times$ denotes the skew symmetric matrix defined as:

$$[v]_\times = \begin{pmatrix} 0 & -c & b \\ c & 0 & -a \\ -b & a & 0 \end{pmatrix} \tag{1}$$

Next, define the head and feet positions of a person viewed by these two cameras at a given instant in time as $\mathbf{p}_1^\mathbf{t}$ (top), $\mathbf{p}_1^\mathbf{b}$ (bottom) and $\mathbf{p}_2^\mathbf{t}$ (top), $\mathbf{p}_2^\mathbf{b}$ (bottom) points, respectively. These corresponding pair of points define a one parameter family of planes given by

$$\begin{aligned} \pi_\alpha &= \alpha \mathbf{P}_1^T [\mathbf{p}_1^\mathbf{t}]_\times \mathbf{p}_1^\mathbf{b} + \mathbf{P}_2^T [\mathbf{p}_2^\mathbf{t}]_\times \mathbf{p}_2^\mathbf{b} &(2)\\ &= \alpha \begin{pmatrix} [\mathbf{p}_1^\mathbf{t}]_\times \mathbf{p}_1^\mathbf{b} \\ 0 \end{pmatrix} + \begin{pmatrix} \mathbf{F}^T [\mathbf{e}']_\times^T [\mathbf{p}_2^\mathbf{t}]_\times \mathbf{p}_2^\mathbf{b} \\ \mathbf{e}'^T [\mathbf{p}_2^\mathbf{t}]_\times \mathbf{p}_2^\mathbf{b} \end{pmatrix} &(3) \end{aligned}$$

where α is a scalar parameter.

The homography induced by this family of planes is then given by

$$\begin{aligned} \mathbf{H}_\alpha &= \left(\mathbf{e}'^T [\mathbf{p}_2^\mathbf{t}]_\times \mathbf{p}_2^\mathbf{b} I - \mathbf{e}' \mathbf{p}_2^{\mathbf{b}^T} [\mathbf{p}_2^\mathbf{t}]_\times^T \right) [\mathbf{e}']_\times \mathbf{F} \\ &\quad - \alpha \mathbf{e}' \mathbf{p}_2^{\mathbf{b}^T} [\mathbf{p}_2^\mathbf{t}]_\times^T &(4)\\ &= [[\mathbf{p}_2^\mathbf{t}]_\times \mathbf{p}_2^\mathbf{b}]_\times [\mathbf{e}']_\times [\mathbf{e}']_\times \mathbf{F} + \alpha \mathbf{e}' \mathbf{p}_2^{\mathbf{b}^T} [\mathbf{p}_2^\mathbf{t}]_\times^T &(5)\\ &= [[\mathbf{p}_2^\mathbf{t}]_\times \mathbf{p}_2^\mathbf{b}]_\times \mathbf{F} + \alpha \mathbf{e}' \mathbf{p}_2^{\mathbf{b}^T} [\mathbf{p}_2^\mathbf{t}]_\times^T &(6) \end{aligned}$$

Now, let m and m' be two corresponding points of a 3D point M viewed by the two cameras. The homography $\mathbf{H}_\alpha$ would map m from the left image to the right image as

$$\begin{aligned} \mathbf{H}_\alpha m &= [[\mathbf{p}_2^\mathbf{t}]_\times \mathbf{p}_2^\mathbf{b}]_\times \mathbf{F} m + \alpha \mathbf{e}' \mathbf{p}_2^{\mathbf{b}^T} [\mathbf{p}_2^\mathbf{t}]_\times^T m &(7)\\ &= [[\mathbf{p}_2^\mathbf{t}]_\times \mathbf{p}_2^\mathbf{b}]_\times [\mathbf{e}']_\times m' + \beta \mathbf{e}' &(8)\\ &= (1 - \gamma) m' + \gamma \mathbf{e}' + \beta \mathbf{e}' &(9) \end{aligned}$$

where $\beta = \frac{\alpha}{\mathbf{p}_2^{\mathbf{b}^T} [\mathbf{p}_2^\mathbf{t}]_\times^T m}$, γ is a scalar parameter, and the last equation follows from the fact that the point $[[\mathbf{p}_2^\mathbf{t}]_\times \mathbf{p}_2^\mathbf{b}]_\times [\mathbf{e}']_\times m'$ is on the epipolar line $[\mathbf{e}']_\times m'$ and hence can be written as a linear combination of $\mathbf{e}'$ and m'.

Therefore by proper scaling of the last equation we can get

$$\mathbf{H}_\alpha m = (1 - \tau) m' + \tau \mathbf{e}' \tag{10}$$

Here the scalar parameter τ may be interpreted as the projective depth of the point M from the plane π_α, because we can readily verify that if $M \in \pi_\alpha$, then $\tau = 0$. Otherwise, τ will

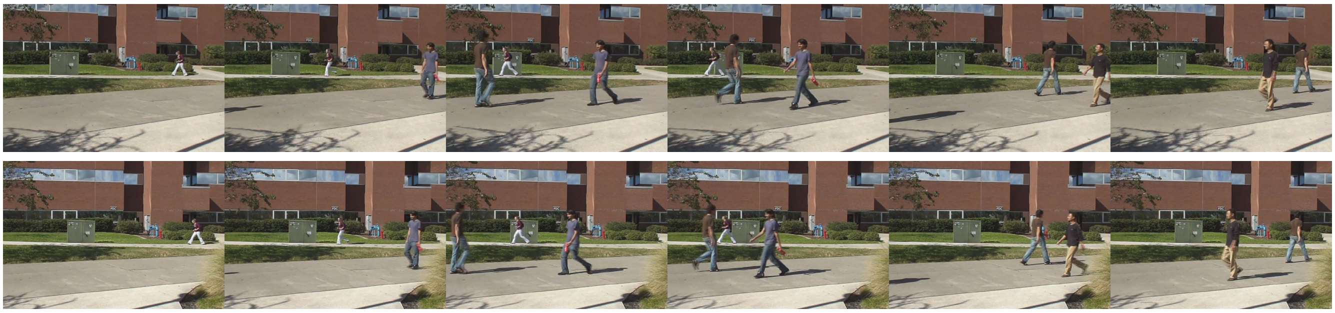


FIGURE 4. Input images: First row shows the selected images as seen from first view and second row shows the input images from second view. These images (from left to right) show multiple moving objects which (in the order of increasing distance from the far wall) include a girl walking from left to right, followed by a boy walking from left to right holding water bottle, another boy moving from right to left, and finally another boy moving from left to right.

be either positive or negative depending on which side of the plane, M lies.

Rearranging (10), we can determine τ from either x or y coordinates of the points m, m', and $\mathbf{e}'$. For instance using x coordinates we have:

$$\tau = \frac{(\mathbf{H}_\alpha m)_x - (m')_x}{(\mathbf{e}')_x - (m')_x} \tag{11}$$

where $(\cdot)_x$ denotes the x coordinate of the vector.

One last issue before we describe how we can use (11) for selective subtraction: The base homography $\mathbf{H}_\alpha$ as derived above is parameterized in terms of a scalar α. There are several ways we can determine α. One simple way is to use a pair of corresponding points between the two camera views to solve for α using (6). For instance, either the head or feet point correspondences of the person in the two cameras in a later frame can be used to determine α. In this way, a walking person would establish a reference plane as depicted in Figure 1.

B. SELECTIVE SUBTRACTION

We use the reference plane as the decision boundary between foreground and background objects. *Any plane in the scene can be chosen as the reference plane and thus it gives us the flexibility of selectively keeping or subtracting the objects on either side of the plane*. For instance, if the reference plane chosen is the farthest plane in the scene then all moving objects fall in front of the reference plane and thus the approach can be used as a traditional background subtraction technique. The projective depth (τ) for any moving object in the scene can be estimated and based on the sign of τ, the object can be classified as being on the foreground or the background. Moreover, The rate of change of τ over time may be interpreted as 'projective speed' of the object relative to the reference plane. For instance, in Figure 2 when an object moves, the rate of change of τ can be estimated and can be used in several applications including vehicle navigation or detecting anomalies in pedestrian paths. Furthermore, the idea of a single reference plane and the estimation of projective depth can be extended to the use of multiple reference planes which would allows us to classify a scene as layers of foreground or background [37] where different objects could belong to different foreground layers. The ability of having multiple reference planes enables us to define multiple foregrounds or multiple backgrounds and hence a notion of *in-between* two layers. Moreover, it is important to highlight that the proposed technique can also be used even when an object is fully or partially occluded (full occlusion can be detected as the object disappearing from the foreground).

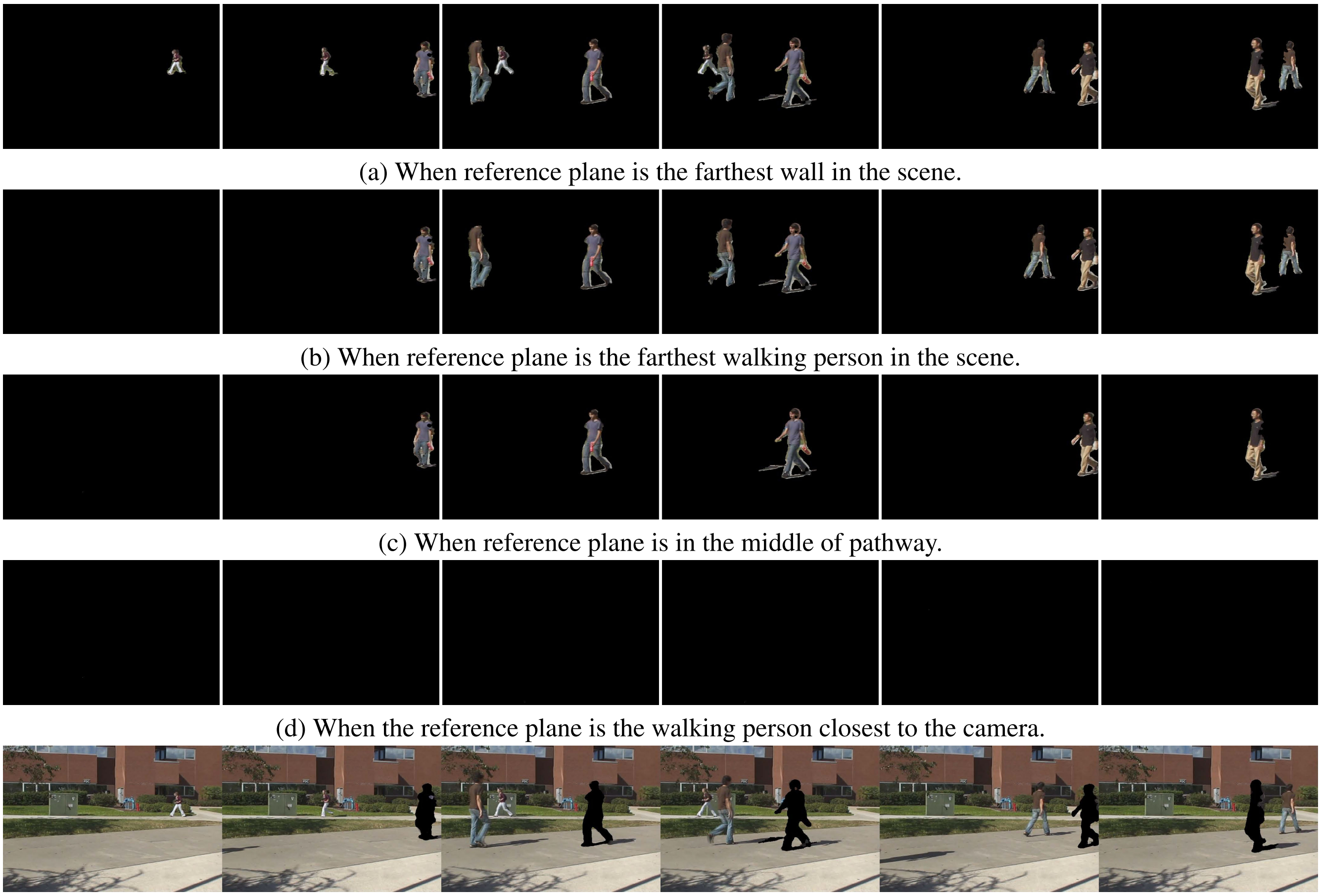
TABLE 1. Summary of the datasets.

Dataset	Total Number of Frames	Image Resolution
outdoor	1000	720 × 480
indoor	867	720 × 480
cellphone-A	140	1080 × 1920
cellphone-B	277	1080 × 1920
cellphone-C	224	1920 × 1080

IV. RESULTS AND DISCUSSION

The algorithm was tested on five set of challenging sequences with multiple moving objects with significant occlusions and illumination changes. These dataset are named as: **outdoor**, **indoor**, and three **cellphone** datasets. The comparative results with the Mixture of Gaussian method [6] have also been presented. A simple threshold based frame difference algorithm along with connected component analysis was used to detect the changes (or blobs) in the scene. We use state of the art feature matching algorithm, Scale Invariant Feature Transform (SIFT) [38], to find point correspondences. Table 1 summarizes the datasets used for testing the proposed method.

Outdoor Dataset: The first sequence contains an outdoor scene with several moving objects, possible casting shadows. It contains 1000 frames from each camera view with the resolution of 720 × 480. The scene also contains dynamic background motion, such as swaying tree leaves. The reference *walk* from a moving object was selected as *reference plane* and *base homography* was estimated using head and feet positions in the first and the last frames as shown in Figure 2. It should be highlighted that only four point correspondences are used to calculate the *base homography* and we do not require any additional training data. An alternate approach would be to track the head and feet positions throughout the



(a) When reference plane is the farthest wall in the scene.

(b) When reference plane is the farthest walking person in the scene.

(c) When reference plane is in the middle of pathway.

(d) When the reference plane is the walking person closest to the camera.

(e) Input images without the foreground blobs for (c).

FIGURE 5. Selective subtraction results for outdoor sequence with different *reference planes*: (a)First row shows the blobs found in foreground when the *reference plane* is the farthest wall in the scene. All moving objects are detected as foreground. (b)Second row shows the blobs detected as foreground if the farthest moving object (girl) is used as *reference plane*. All moving objects excluding the girl are now detected as foreground. (c)Third row shows the blobs detected as foreground when the *reference plane* used is in the middle of the pathway. Notice that the girl walking to the left and the boy walking to the right are both on the other side of the *reference plane* and are detected as background. Furthermore, two boys walking to the left are correctly detected as foreground. (d)Fourth row shows the results when the *reference plane* is the moving object closest to the camera and thus none of the moving objects are detected as foreground. (e)Last row shows the input images excluding the foreground blobs detected when the *reference plane* was in the middle of pathway shown in (c).

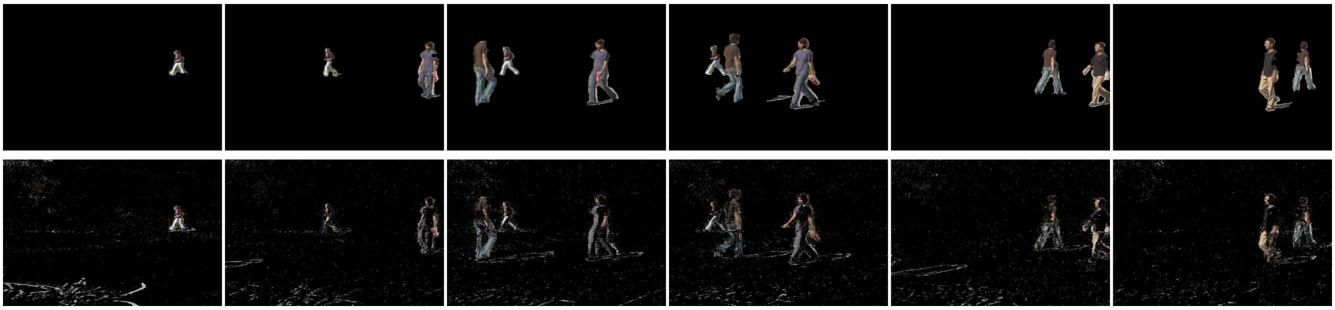


FIGURE 6. Selective subtraction as background subtraction: The results of the proposed selective subtraction method when the *reference plane* is the far wall and hence all moving objects are considered foreground as in traditional background subtraction techniques. First row shows the results obtained from the proposed method and the second row shows the results from state of the art mixture of Gaussian method [6]. The results indicate that the proposed technique can be used as background subtraction and gives better qualitative results.

reference *walk* and use curve fitting techniques to improve the precision of head and feet positions [40]. Moreover, numerous complex algorithms can be used to detect the blobs with varying degree of success. The discussion on these algorithms is outside the scope of this paper.

Once the blobs are detected, we use the proposed algorithm to estimate the projective depth (τ) as described in Section III-A. In our experiments, we first performed the blob detection followed by feature matching for point correspondences using SIFT. Notice that these two steps can be

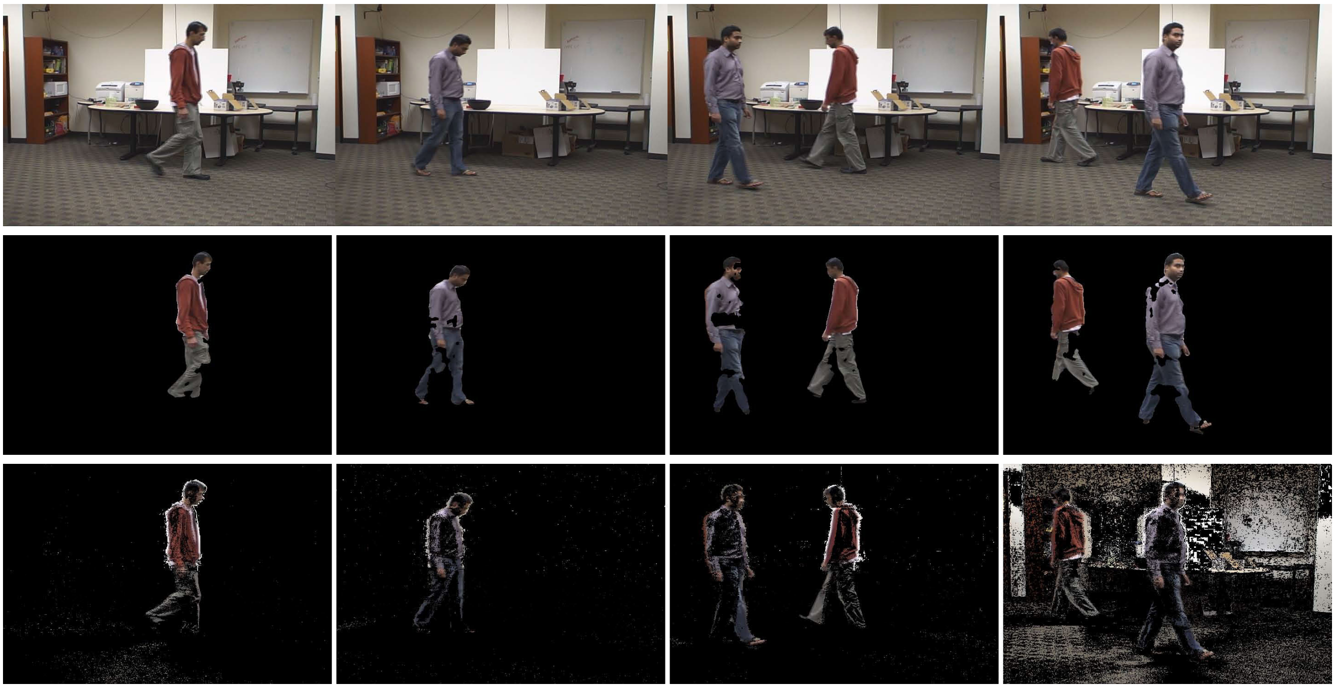


FIGURE 7. Selective subtraction results for indoor sequence: The results of the proposed selective subtraction method are shown here. First row shows the input images from the first view. The objects found in front of the *reference plane* using selective subtraction are shown in second row and the results of mixture of Gaussian method [6] are shown in the bottom row. The *reference plane* used in these results is the farthest wall in the scene. The results indicate that the proposed technique can effectively detect foreground objects in indoor environments.

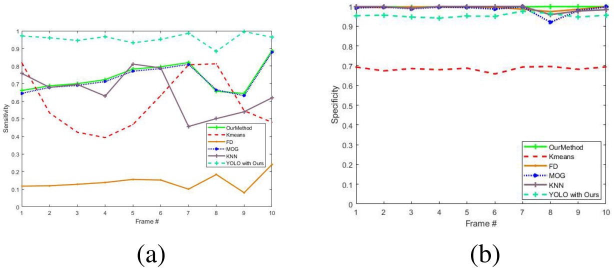


FIGURE 8. Quantitative analysis of detection accuracy on cellphone-C dataset: (a) shows the sensitivity of the proposed algorithm (Average values: Ours 73.6%, [6] 14%, [39] 72.7%), (b) shows the specificity (Average values: Ours 99.8%, [6] 99.3%, [39] 98.5%). The results show that the average detection sensitivity of the proposed is consistently better than [6] and [39] and specificity is comparable to these techniques.

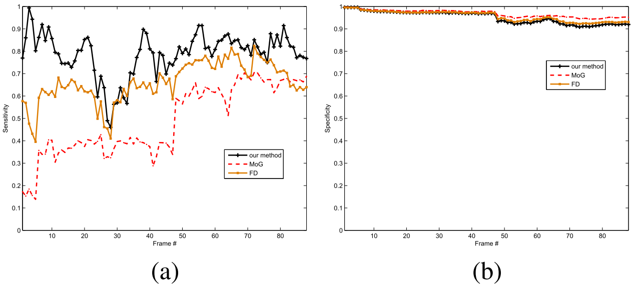


FIGURE 9. Quantitative analysis of detection accuracy: (a) shows the sensitivity of the proposed algorithm (Average values: Ours 79%, [6] 49%, [39] 64%), (b) shows the specificity (Average values: Ours 95%, [6] 96%, [39] 95%). The results show that the average detection sensitivity of the proposed is consistently better than [6] and [39] and specificity is comparable to these techniques.

reversed, as to finding point correspondences on the entire image followed by eliminating the ones outside the blobs. Figure 4 shows two views of the input images used for blob detection and feature matching for point correspondences. For each corresponding point, we calculate τ using (11) and use a majority voting scheme to classify the blob as foreground or background (i.e., as being on one side of the reference plane or the other). The results are depicted in Figure 5, showing that the proposed algorithm can correctly separate the foreground from background.

One of the most unique aspects of our proposed technique is the flexibility it provides in selecting the *reference plane* of choice. Figure 5 shows how the foreground detection changes when different *reference planes* are selected for selective subtraction. Figure 5(a) shows the results when the *reference plane* is the far wall and hence all moving objects are considered foreground as in traditional background subtraction technique. When the *reference plane* is changed to a moving object, the foreground changes accordingly as seen in Figure 5(b). Figure 5(c) shows the results when the selected *reference plane* is in the middle of pathway thus, detecting the objects in front as foreground. We also selected our *reference plane* as the object walking closest to the camera and found that all moving objects were detected as background. Figure 6 depicts the qualitative results, showing that the proposed technique performs better than mixture of Gaussian [6].

Indoor Dataset: Our second dataset contains an indoor scene with significant illumination changes and the results

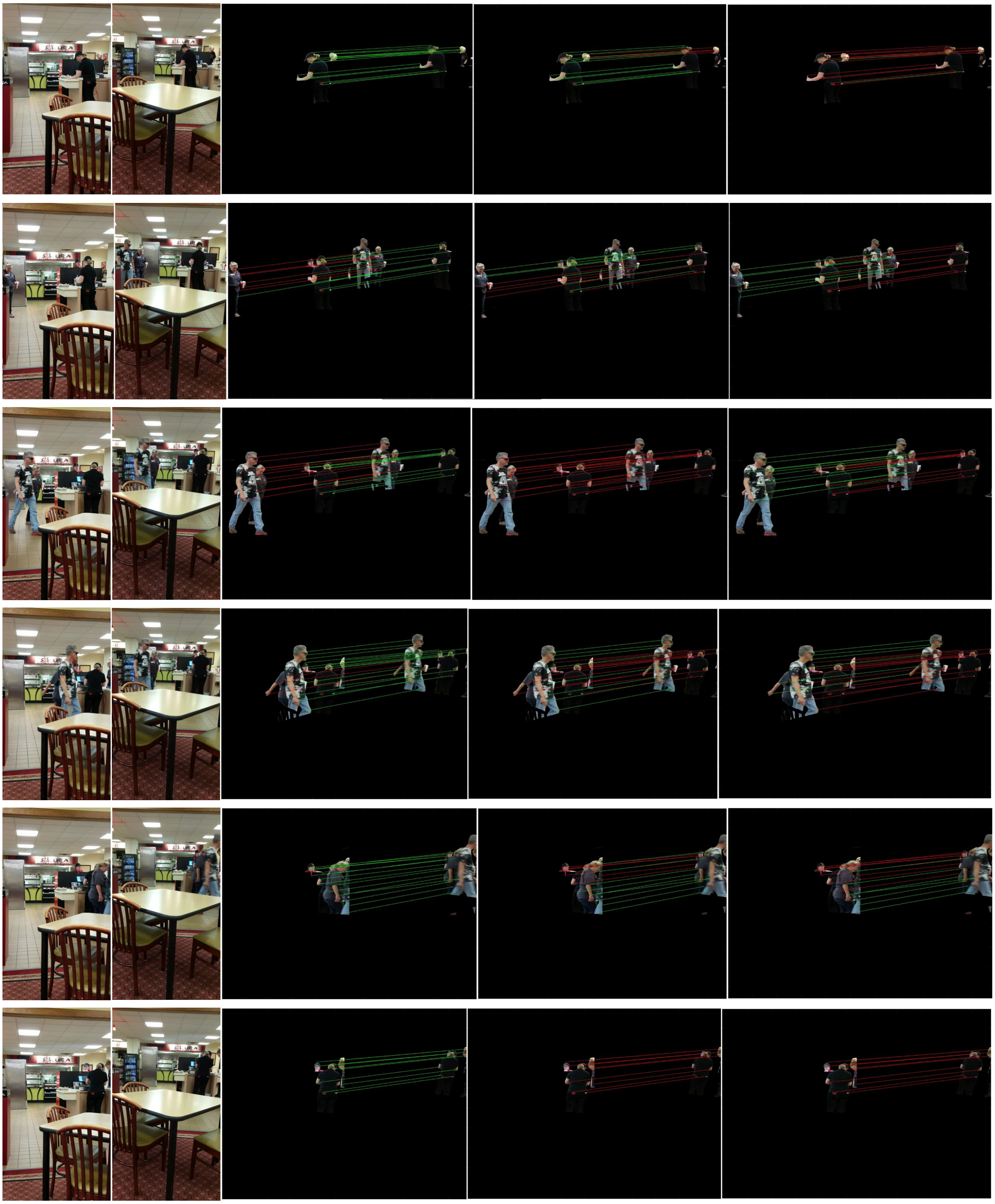


FIGURE 10. Selective subtraction results for **cellphone-A** sequence: Baristas are seen brewing the coffee and taking orders for the customers. We see some customers walking and pass in front of the staff from the left and move to the right of the scene. Each row shows the images captured from both cellphone cameras along with the foreground and background points detected by our algorithm when different reference planes are chosen. The first column of the figures shows the images captured from one cellphone camera and the second column shows images captured from the second camera. The remaining columns show the results obtained from our method when different *reference planes* are chosen. The third column shows the results when the farthest wall or plane is used as reference plane. The forth column shows the results when the middle plane is used as reference plane and the fifth column shows the results when foremost area is chosen as reference plan.

are shown in Figure 7. The dataset contains 867 frames from each camera view with the resolution of 720 × 480. The scene contains a table with some objects lying on the table and a book shelf to the back of the room. People walk in front of

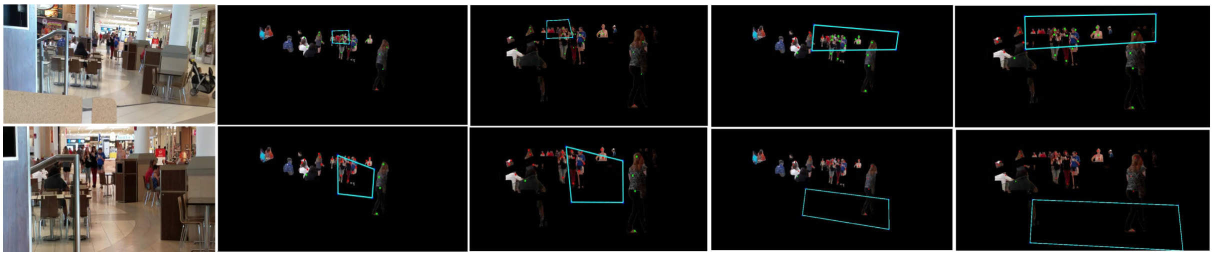


FIGURE 11. Reference planes: The reference plane used in Cellphone-B are shown here. First row shows images from left camera and second row shows the corresponding images from right camera. The first column shows the two frames used for sift matches. The remaining columns show selected reference planes as follows: (from left to right) when plane is farthest from camera, when plane is in the middle - one farther from camera and one closer, and when plane is closest to the camera.

TABLE 2. **Quantitative Analysis** This table shows results obtained from our methods and also compares to other methods, tested on the same data. Our result are consistently better than other approaches.

	Ours		**[7]**		**[46]**	
	Sensitivity	Specificity	Sensitivity	Specificity	Sensitivity	Specificity
outdoor	79	95	49	96	64	95
cellphone-C	74	99.8	14	99.3	73	98.5

the camera from the left of the room to the right and vice versa.

Cellphone Dataset: This dataset comprises of three separate recordings, which we denote at **cellphone-A**, **cellphone-B** and **cellphone-C**. These datasets were captured with two *handheld* SAMSUNG Galaxy S7 and Note 4 cellphones with an image resolution of 1080 × 1920. cellphone-A was captured inside a cafe, where baristas are seen brewing coffee and taking orders for the customers. We see some customers passing in front of the staff from the left and move to the right of the scene. This is shown in Figure 10. Each row shows the images captured from both cellphone cameras along with the foreground and background points detected by our algorithm when different reference planes are chosen. The first column of the figures shows the images captured from one cellphone camera and the second column shows images captured from the second camera. The remaining columns show the results obtained from our method when different *reference planes* are chosen. The third column shows the results when the farthest wall or plane is used as reference plane. In most results, objects are correctly classified as foreground objects. The forth column shows the results when the middle plane is used as reference plane and the fifth column shows the results when foremost area is chosen as reference plan. The average accuracy scores of 84%, 71.8% and 82.2% were observed for correct classification of each point shown in last three columns. Similarly, Figure 12 shows some of the frames in the **cellphone-B** dataset. This set of sequence captures a food court in a shopping mall. People are seen moving in the background and helping themselves with food. Each row of the figure shows results obtained from the proposed method. The first and second columns show two views of images captured from cellphone cameras. The remaining 4 plots in each row show the results obtained from our method when different *reference planes* are chosen. The top-left plot shows the results when the farthest wall or plane is used as reference plane. The bottom-right plot shows the results when the closest plane (i.e., closest to camera) is used as reference plane. The bottom-left and top-right plots show the results when different middle planes are used as reference planes. In most results, objects are correctly classified as foreground and background objects. The average accuracy scores of 94%, 94%, 85.3%, and 76% were observed for correct classification of each point shown in last two columns. Finally, Figure 13 shows some of the frames in the **cellphone-C** dataset. This sequence captures most challenging scene which includes dynamic moving objects (i.e., bushes, moving due to strong wind) as well as shadows. People are seen moving in both directions. The top two rows show some of the frames from two views of cellphone cameras. Third row show results from our proposed algorithm when the chosen reference planes is in the middle. The girl in green shirt walking from the left is selectively subtracted due to being in the background. Fourth row shows results from our proposed algorithm when the chosen reference plane is the farthest plane in the scene hence this approach becomes a traditional background subtraction approach. All moving objects are correctly classified as foreground. The remaining rows show results from other approaches. These results indicate that selective subtraction approach is effective and provides flexibility in selectively subtracting the objects of choice from the scene. The results are qualitatively demonstrated and compared to other methods, as shown in Figure 7. The qualitative analysis of these results clearly shows that our proposed technique performs very well in challenging environments even when used with datasets captured with handheld cameras.

A. QUANTITATIVE ANALYSIS

We also performed the quantitative analysis of the pixel-level detection accuracy. The per frame detection rates are calcu-

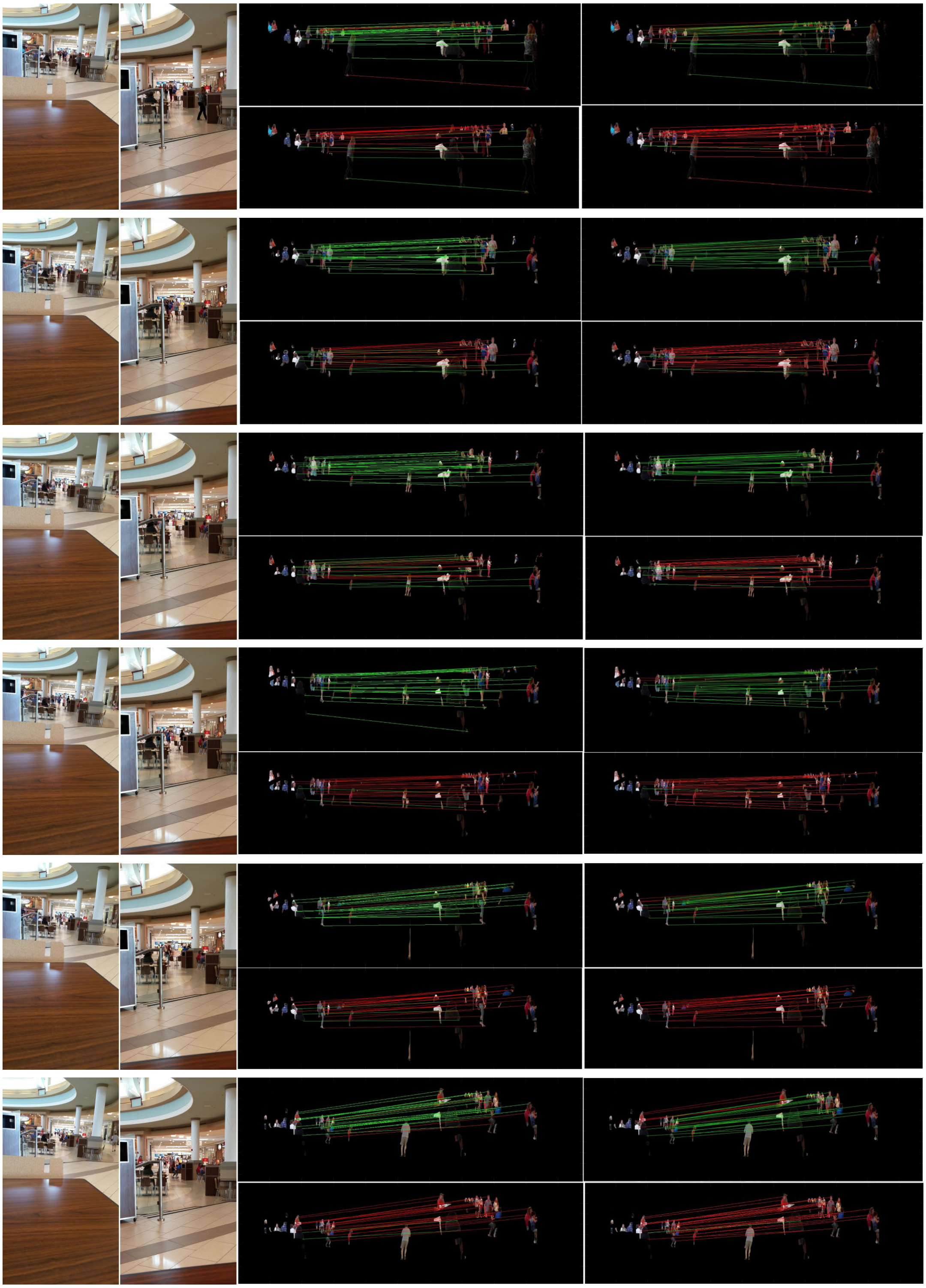
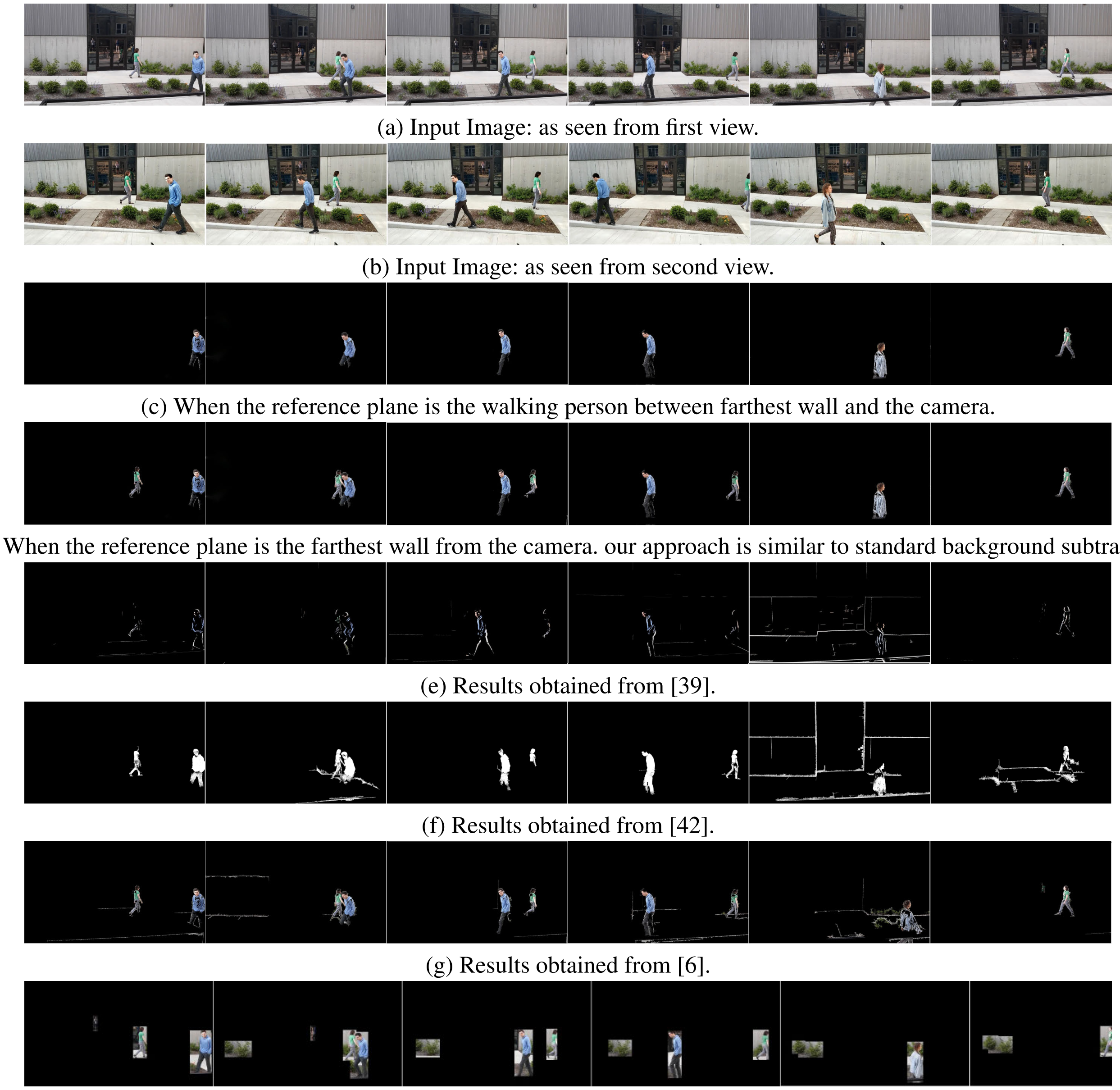


FIGURE 12. Selective subtraction results for **cellphone-B** sequence: This dataset captures a food court in a shopping mall. People are seen moving in the background and helping themselves with food. The third row of the figure shows results obtained from the proposed method. The first and second columns show two views of images captured from cellphone cameras. The remaining 4 plots in each row show the results obtained from our method when different *reference planes* are chosen. The top-left plot shows the results when the farthest wall or plane is used as reference plane. The bottom-right plot shows the results when the closest plane (i.e., closest to camera) is used as reference plane. The bottom-left and top-right plots show the results when different middle planes are used as reference planes.



(a) Input Image: as seen from first view.

(b) Input Image: as seen from second view.

(c) When the reference plane is the walking person between farthest wall and the camera.

(d) When the reference plane is the farthest wall from the camera. our approach is similar to standard background subtraction

(e) Results obtained from [39].

(f) Results obtained from [42].

(g) Results obtained from [6].

(h) Results obtained from [43].

FIGURE 13. Selective subtraction results for cellphone-C outdoor sequence with different *reference planes*: The top two rows (a) and (b) show some of the frames from two views of cellphone cameras. Third row (c) shows results from our proposed algorithm when the chosen reference plane is in the middle. The Fourth row (d) shows results from our proposed algorithm when the chosen reference planes is the farthest plane in the scene. In this case, our approach is similar to standard background subtraction. The remaining rows show results from other approaches.

lated in terms of sensitivity and specificity, where

$$Sensitivity = \frac{\#\ of\ true\ positives\ detected}{total\ \#\ of\ true\ positives}$$

$$Specificity = \frac{\#\ of\ true\ negatives\ detected}{total\ \#\ of\ true\ negatives}$$

Figure 9 shows the sensitivity and specificity of the proposed technique as compared to [6] and [39]. Clearly, the detection accuracy in terms of sensitivity is consistently higher than [6] and [39] while specificity is comparable to both techniques. One of the major advantages of the proposed technique is that it does not require any special camera setup or configuration or depth sensing device as needed in other two-view background subtraction techniques. We also do not use the disparity map and thus the proposed algorithm is fast and computationally efficient. The average computation time per frame (480 × 720 pixels) is 0.0029 seconds on Intel Core2 Extreme CPU with 4GB RAM (excluding the time needed for blob detection and the feature matching). It should be noted that we have not performed any shadow removal or other post-processing, such as graph cuts [10] to improve the boundaries of foreground objects.

Table 2 shows the results obtained from our method. We also compare our results to the standard methods of [6] and [39]. The first column shows different datasets that we have tested in this paper. The second column shows the specificity and sensitivity measurements obtained from the proposed method, where as the third and the fourth column mentions the obtained measure from [6] and [39], respectively. As can be seen from the table, results obtained from out approach are much higher and better. For the **outdoor** dataset, we obtained 79% and 95% for specificity and sensitivity, respectively. Similarly, for the **cellphone-C** dataset, we obtain 74% and 99.8%, where the best results obtained from the competition is that of 73% from [39] and 98.5% from [6] for specificity and sensitivity, respectively. These results show that the proposed method is robust and applicable. Moreover, the method is fast and computationally efficient. The above encouraging results demonstrate the practicality and viability of the proposed method.

B. IMPROVEMENTS

The proposed algorithm relies heavily on the change detection and feature matching algorithms for point correspondences and uses their results to estimate the projective depth. For improved results the following recommendations should be followed:

- An important constraint in the estimation of *base homography* is its consistency with the fundamental matrix and thus all point correspondences should satisfy this constraint.
- *Base homography* can be estimated using two instances of the *walk* (or two frames only). Any error in selecting appropriate instances can result in wrong *reference plane* and thus introduce errors in the estimation of *base homography*. A more robust approach can be adopted, by tracking the head and feet positions over a period of time and then using curve fitting techniques to select the best candidates for head and feet positions.
- The accuracy of point correspondences used is critical for selective subtraction approach. A reliable feature matching algorithm like SIFT or Triangle Constraint Measurements (TCM) [41] is recommended to minimize the probability of false matches. The consistency of these matches with fundamental matrix can also be used.
- The use of effective blob detection algorithm is also important for selective subtraction. Numerous complex change detection algorithms are available which can be used for blob detection such as those using statistical properties of the pixels or color features. Moreover, selective subtraction can be used within the framework of any object detection algorithm as a refinement step.

V. CONCLUSION

This work presents a number of fundamental innovations in the context of background subtraction. We present a novel concept of background as objects *other than* foreground which may include moving objects in the scene that cannot be learned from a training set because they occur only irregularly and sporadically. Our proposed method, ‘‘Selective Subtraction’’, is as alternative to standard background subtraction, and we show that a *reference plane* in a scene is sufficient as the decision boundary between foreground and background. Furthermore, the flexibility in selecting the *reference plane* using the actual moving object in the scene or an arbitrary plane in the scene, is truly unique to this method and is not available in existing background subtraction techniques. We also show that the proposed technique enables us to select multiple reference planes and thus relaxing the strict binary classification-based paradigm. We present promising results on a challenging set of image sequences to show that the selective subtraction approach performs effectively and has applications in background subtraction, vehicle navigation, path anomaly detection, and detecting objects in crowds. We also present results on images sequences from hand-held cameras to show that the proposed technique is relatively immune to camera motion and is robust. Furthermore, we provide recommendations to improve the results of selective subtraction approach.

REFERENCES

[1] A. Sobral and A. Vacavant, “A comprehensive review of background subtraction algorithms evaluated with synthetic and real videos,” *Comput. Vis. Image Understand.*, vol. 122, pp. 4–21, May 2014.

[2] T. Bouwmans, “Traditional and recent approaches in background modeling for foreground detection: An overview,” *Comput. Sci. Rev.*, vols. 11–12, pp. 31–66, May 2014.

[3] T. Zhao, M. Aggarwal, R. Kumar, and H. Sawhney, “Real-time wide area multi-camera stereo tracking,” in *Proc. CVPR*, Jun. 2005, pp. 976–983.

[4] O. Javed and S. M. , “Tracking and object classification for automated surveillance,” in *Proc. ECCV*, 2002, pp 343–357.

[5] C. Wren, A. Azarbayejani, T. Darell, and A. Pentland, “Pfinder: Real-time tracking of the human body,” *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 19, no. 7, pp. 780–785, Jul. 1997.

[6] C. Stauffer and W. E. L. Grimson, “Learning patterns of activity using real-time tracking,” *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 22, no. 8, pp. 747–757, Aug. 2000.

[7] A. Elgammal, R. Duraiswami, D. Harwood, and L. S. Davis, “Background and foreground modeling using nonparametric kernel density estimation for visual surveillance,” *Proc. IEEE*, vol. 90, no. 7, pp. 1151–1163, Jul. 2002.

[8] S. Zhang, H. Yao, S. Liu, X. Chen, and W. Gao, “A covariance-based method for dynamic background subtraction,” in *Proc. 19th Int. Conf. Pattern Recognit.*, Dec. 2008, pp. 1–4.

[9] A. Monnet, A. Mittal, N. Paragios, and V. Ramesh, “Background modeling and subtraction of dynamic scenes,” in *Proc. 9th IEEE Int. Conf. Comput. Vis.*, Oct. 2003, pp. 1305–1312.

[10] Y. Sheikh and M. Shah, “Bayesian modeling of dynamic scenes for object detection,” *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 27, no. 11, pp. 1778–1792, Nov. 2005.

[11] Y.-L. Tian, M. Lu, and A. Hampapur, “Robust and efficient foreground analysis for real-time video surveillance,” in *Proc. CVPR*, Jun. 2005, pp. 1182–1187.

[12] T. Ko, S. Soatto, and D. Estrin, “Background subtraction on distributions,” in *Proc. ECCV*, 2008, pp. 276–289.

[13] Y. Benezeth, P. M. Jodoin, B. Emile, H. Laurent, and C. Rosenberger, “Review and evaluation of commonly-implemented background subtraction algorithms,” in *Proc. 19th Int. Conf. Pattern Recognit.*, Dec. 2008, pp. 1–4.

[14] R. Pless, J. Larson, S. Siebers, and B. Westover, “Evaluation of local models of dynamic backgrounds,” in *Proc. CVPR*, Jun. 2003, pp. 2–73.

[15] Y. Ren, C.-S. Chua, and Y.-K. Ho, “Motion detection with nonstationary background,” *Mach. Vis. Appl.*, vol. 13, nos. 5–6, pp. 332–343, Mar. 2003.

[16] A. Mittel and N. Paragios, "Motion-based background subtraction using adaptive kernel density estimation," in *Proc. CVPR*, 2004, p. 2.

[17] J. Zhong and Sclaroff, "Segmenting foreground objects from a dynamic textured background via a robust Kalman filter," in *Proc. 9th IEEE Int. Conf. Comput. Vis.*, 2003, pp. 44–50.

[18] K. Karmann and A. V. Brandt, "Moving object recognition using an adaptive backgroound memory," in *Time-Varying Image Process*. Amsterdam, The Netherlands: Elsevier, 1990.

[19] N. Friedman and S. Russell, "Image segmentation in video sequences: A probabilistic approach," in *Proc. 13th Conf. Uncertainty Artif. Intell.* San Francisco, CA, USA: Morgan Kaufmann, 1997.

[20] W. Grimson and C. Stauffer, "Adaptive background mixture models for real-time tracking," in *Proc. CVPR*, 1999, pp. 246–252.

[21] M. Braham, S. Pierard, and M. Van Droogenbroeck, "Semantic background subtraction," in *Proc. IEEE Int. Conf. Image Process. (ICIP)*, Sep. 2017, pp. 4552–4556.

[22] M. Babaee, D. T. Dinh, and G. Rigoll, "A deep convolutional neural network for background subtraction," 2017, *arXiv:1702.01731*. [Online]. Available: https://arxiv.org/abs/1702.01731

[23] P.-L. St-Charles, G.-A. Bilodeau, and R. Bergevin, "SuBSENSE: A universal change detection method with local adaptive sensitivity," *IEEE Trans. Image Process.*, vol. 24, no. 1, pp. 359–373, Jan. 2015.

[24] R. Wang, F. Bunyak, G. Seetharaman, and K. Palaniappan, "Static and moving object detection using flux tensor with split Gaussian models," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. Workshops*, Jun. 2014, pp. 420–424.

[25] P. Christiansen, L. Nielsen, K. Steen, R. Jørgensen, and H. Karstoft, "DeepAnomaly: Combining background subtraction and deep learning for detecting obstacles and anomalies in an agricultural field," *Sensors*, vol. 16, no. 11, p. 1904, Nov. 2016.

[26] L. P. Cinelli, L. A. Thomaz, A. F. da Silva, E. A. B. da Silva, and S. L. Netto, "Foreground segmentation for anomaly detection in surveillance videos using deep residual networks," in *Proc.35th Simpósio Brasileiro De Telecomunicações E Processamento De Sinais*, 2017, pp. 3–6.

[27] L. Yang, J. Li, Y. Luo, Y. Zhao, H. Cheng, and J. Li, "Deep background modeling using fully convolutional network," *IEEE Trans. Intell. Transp. Syst.*, vol. 19, no. 1, pp. 254–262, Jan. 2018.

[28] L. A. Lim and H. Yalim Keles, "Foreground segmentation using convolutional neural networks for multiscale feature encoding," *Pattern Recognit. Lett.*, vol. 112, pp. 256–262, Sep. 2018.

[29] Y. Chen, J. Wang, B. Zhu, M. Tang, and H. Lu, "Pixelwise deep sequence learning for moving object detection," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 29, no. 9, pp. 2567–2579, Sep. 2019.

[30] D. Sakkos, H. Liu, J. Han, and L. Shao, "End-to-end video background subtraction with 3D convolutional neural networks," *Multimedia Tools Appl.*, vol. 77, no. 17, pp. 23023–23041, Dec. 2017.

[31] Y. Ivanov, A. Bobick, and J. liu, "Fast lighting independent background subtraction," in *Proc. Workshop Video Surveillence ICCV*, 1998, pp. 199–207.

[32] S. A. D. L. Lim Mittal and N. Paragios, "Fast illumination-invariant background subtraction using two views: Error. Analysis, sensor placement and applications," in *Proc. CVPR*, 2005, pp. 1071–1078.

[33] R. Diaz, S. Hallman, and C. C. Fowlkes, "Detecting dynamic objects with multi-view background subtraction," in *Proc. IEEE Int. Conf. Comput. Vis.*, Dec. 2013, pp. 273–280.

[34] W. Guan and P. Monger, "Real-time detection of out-of-plane objects in stereo vision," in *Proc. Adv. Vis. Comput.*, vol. 4291, Nov. 2006, pp. 102–111.

[35] T. Liu and G. Wang, "A hierarchical approach for robust background subtraction based on two views," in *Proc. WRI Global Congr. Intell. Syst.*, vol. 4, May 2009, pp. 325–329.

[36] T. Z. F. Lv and R. Nevatia, "Self-calibration of a camera from video of a walking human," in *Proc. ICIP*, 2002, pp. 562–567.

[37] A. A. Bhutta, I. N. Junejo, and H. Foroosh, "Selective subtraction when the scene cannot be learned," in *Proc. 18th IEEE Int. Conf. Image Process.*, Sep. 2011, pp. 3273–3276.

[38] D. G. Lowe, "Object recognition from local scale-invariant features," in *Proc. 7th IEEE Int. Conf. Comput. Vis.*, 1999, pp. 1150–1157.

[39] K. Toyama, J. Krumm, B. Brumitt, and B. Meyers, "Wallflower: Principles and practice of background maintenance," in *Proc. 7th IEEE Int. Conf. Comput. Vis.*, Sep. 1999, pp. 255–261.

[40] I. N. Junejo, "Using pedestrians walking on uneven terrains for camera calibration," *Mach. Vis. Appl.*, vol. 22, no. 1, pp. 137–144, Aug. 2009.

[41] X. Guo and X. Cao, "Triangle-constraint for finding more good features," in *Proc. 20th Int. Conf. Pattern Recognit.*, Aug. 2010, pp. 1393–1396.

[42] J. Malik, S. Belongie, T. Leung, and J. Shi, "Contour and texture analysis for image segmentation," *Int. J. Comput. Vis.*, vol. 43, no. 1, pp. 7–27, Jun. 2001, doi: 10.1023/A:1011174803800.

[43] J. Redmon, S. Divvala, R. Girshick, and A. Farhadi, "You only look once: Unified, real-time object detection," 2015, *arXiv:1506.02640*. [Online]. Available: http://arxiv.org/abs/1506.02640

ADEEL A. BHUTTA (Member, IEEE) received the B.S. degree in electronic engineering from the Ghulam Ishaq Khan Institute, Topi, Pakistan, in 1999, and the M.S. degrees in computer science and computer engineering from the University of Central Florida, in 2006 and 2012, respectively. He is currently pursuing the Ph.D. degree in computer engineering.

He is currently a Senior Lecturer of computer science with Indiana University, Bloomington, and his teaching expertise includes a wide range of subjects within computer science, electrical and computer engineering, and information technology. His primary areas of research include image processing and computer vision, and he is currently working on deep learning, selective subtraction, dynamic background subtraction, line tracking, and dense correspondences. He was a recipient of several scholarships, fellowships, and awards including the Champion of Inclusion Award in 2017, the TRESTLE Fellowship, and the IU Trustees Teaching Award in 2019.

IMRAN NAZIR JUNEJO received the Ph.D. degree from the University of Central Florida, in 2007. He is working as an Associate Professor with Zayed University, UAE. He has published in several top quality international conferences and journals. His current focus of research is human action recognition from arbitrary views. His other areas of research interests include camera calibration, crowd modeling and analysis, path modeling, video surveillance, scene understanding, and event detection.

HASSAN FOROOSH (Senior Member, IEEE) is currently a CAE Link Professor of computer science with the University of Central Florida (UCF), Orlando, FL, USA. He has authored or coauthored over 150 peer-reviewed scientific articles in the areas of computer vision, image processing, and machine learning. He received the Piero Zamperoni Award from the International Association of Pattern Recognition (IAPR), in 2004, the Best Scientific Paper Award from IAPR-International Conference on Pattern Recognition (IAPR-ICPR), in 2008, and the Best Paper Award from the IEEE International Conference on Image Processing (ICIP), in 2018. He is also the Principal Investigator and the Lead of the Science Data Center of the NASA GOLD Mission that launched a satellite into Earths geo-stationary orbit to study the space weather using ultraviolet imaging, in 2018. He has been serving on the Editorial Boards and Organizing Committees of various IEEE Transactions, conferences, and working groups.