Southeast-Asian J. of Sciences: Vol. 5, No 2 (2017) pp. 137-152

DETECTION AND TRACKING OF HUMANS BASED ON IMPROVED HISTOGRAM OF ORIENTED GRADIENTS AND KALMAN FILTER

Shayhan Ameen Chowdhury[1], Kaushik Deb[1*], Pranab Kumar Dhar[1] and Mirza A. F. M. Rashidul Hasan[2]

[1] *Dept. of Computer Science and Engineering*
Chittagong University of Engineering & Technology (CUET)
Chittagong-4349, Bangladesh
*email: shayhan@yahoo.com, * debkaushik99@cuet.ac.bd, pranabdhar81@gmail.com*

[2] *Dept. of Information and Communication Engineering*
University of Rajshahi
Rajshahi 6205, Bangladesh
email: mirza_iu@yahoo.com

Abstract

Human detection and tracking in a video surveillance system is critical for various application areas including suspicious event detection and human activity recognition. In the current environment of our society suspicious event detection is a burning issue. For that reason, this paper proposes a framework for detection and tracking of humans by generating a human feature vector. Initially, every pixel of a frame is represented as an incorporation of several Gaussians and use a probabilistic method to refurbish the representation. These Gaussian representations are then estimated to classify the background pixels from foreground pixels. Shadow regions are eliminated from foreground by utilizing a Hue-Intensity disparity value between background and current frame. Partial occlusion handling is utilized by color correlogram to label objects within a group. After that, the framework generates regions of interest (ROIs) by considering conditions related to human body. Afterward, features are extracted from ROI for classification. A feature descriptor, Improved

* Corresponding author.
Key words: Human detection, Shadow elimination, Partial occlusion handling, Color correlogram, Histogram of oriented gradients (HOG), Human tracking.

Histogram of Oriented Gradients (ImHOG) is proposed to alleviate the limitation of Histogram of Oriented Gradients (HOG). Finally, Kalman filter is utilized for human tracking to increase detection rate. Various videos containing moving humans are utilized to evaluate the proposed framework and presented outcomes demonstrate the adequacy.

1 Introduction

The escalation of computer vision usages impelled human detection and tracking as an active research field. Human detection and tracking in a video surveillance system has vast application areas including human locomotion characterization, fall detection for patients and intelligent gestural user interface (wiimote, kinect, smart TV). Video surveillance also plays a vital role in fighting crime and protecting public property. Video surveillance is a valuable aid to improve community safety by monitoring important crowded places such as town and city centers, industrial parks, hospitals and universities for early identification of crime and other disruptive incidents. However, with large scale implementation of video surveillance systems manually tracking each camera to identify suspicious events is not possible. For that reason, this paper proposes a framework for detection and tracking of humans in different appearances, poses, uneven illuminations and under occlusion.
Rest of the paper is summarized as follows. Section 2 gives a brief description of related research. The proposed framework for detection and tracking of humans is described in Section 3. Section 4 discusses the simulation results. Finally, Section 5 encloses the concluding remarks.

2 Related Work

In current human detection and tracking frameworks, ROI extraction and feature representation are two main problems being investigated. Over the past few years, a significant amount of work has been done to detect and track human in different appearances, poses, uneven illuminations and under occlusion. Human detection and tracking is a deep-seated and demanding issue because of two challenges: 1) Humans Intra-class divergences like appearance, clothing, skin color and pose; 2) External issues like uneven illumination and cluttered background.
Salient object features are captured by integrating intensity variation of every pixel with texture related features in [1]. These multidimensional features occupy large-scale knowledge about the object. However, the proposed method determined some key thresholds based on hypothesis. Which made the framework fragile when dealing with issues related to outdoor environment such as illumination changes, background clutter. Polar coordinate based shape feature

is generated and used SVM for classification in [2]. However, the system can only detect upper part of human body. In [3], a hybrid local transformation feature is proposed that integrates various regional features such as LGP, LBP and HOG. The proposed hybrid feature shows robustness to local illumination changes. However, the high dimensionality of the hybrid feature increases computational complexity.

Current human detection and tracking frameworks can be broken down into couple of processes. First process employs sliding window while other process employs a part-based detection. The sliding window based process can be improved in two areas: composing more discerning features to improve detection rate and use effective training methods to learn improved classifiers. Widely used discerning features involve Haar wavelet, HOG [4], shapelet, EOH, edgelet, region covariance [5] and LBP [6]. HOG is a very robust feature descriptor capable of detecting human in different appearances and poses.

Several classifiers have been approached for human detection and tracking. Most efficient human detection classifiers commonly employ different variations of boosting algorithms [7], different forms of SVMs or Neural networks. To improve the detection performance, a combination of these classifiers are used to develop a robust classifier structure [5].

Contrary to whole body human detection and tracking frameworks, human parts based detection [8, 9] is better suited to handle partial human occlusions. In these processes, a deformable part model of human body is generated and detection is accomplished if any or all parts are encountered. In [8], a part-based model similar to star structure is introduced. In this model human is represented by a combination of a base structure and various part structures. The presence of high false positives is the main disadvantage of parts-based methods. As a result, plenty of researches are committed to develop vigorous aggregation processes and also to reduce false positive rate. Moreover, for better efficiency high determination pictures or videos are required to capture acceptable and invariant information for every part of the human body. Usually, such kinds of images or videos are not always available. Some systems introduced hybrid features [9] to overcome this issue. Nonetheless, the enhanced detection efficiency also increases the computational cost. For instance, the multiple kernel learning (MKL) framework presented in [10] approximately takes one minute and seven seconds to process each frame. As a result, parts based human detection frameworks are not feasible to detect and track humans in video surveillance systems.

Foreground segmentation, or background subtraction has been another active research field for a long time. Cylindrical codebook structure is applied in [11] to captured background representation into codebook and adaptively renewed the representation. For a sequences of frames an abstract form of background is generated by sampling all background pixel values. In [12], pixel values are quantized with respect to time and grouped the results using mean shift

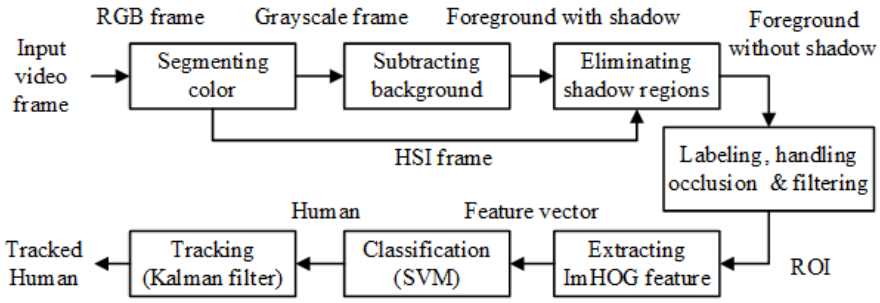

Figure 1: The proposed framework for detection and tracking of humans.

procedure. Every group is labeled with a weight based on the possibility of resulting from background. The efficiency of these frameworks significantly reduced when region of interest becomes motionless for extended periods.
This paper proposes a framework to detect and track humans by extracting foreground from background. The main emphasis of this paper is to eliminate shadow regions from foreground to find accurate region of interest (ROI). Shadows can be defined as portion of regions in a video frame that are not directly illuminated by lighting source. As a result, shadow regions contain same hue (pure color) as background with different intensity value. Based on these properties a hue-intensity disparity value is computed for every foreground pixel to detect and eliminate shadow regions from foreground. Another emphasis of this paper is to generate a feature vector for robust human detection. The limitation of Histogram of Oriented Gradients (HOG) is studied and identified that it cannot differentiate between some local patterns. A feature vector i.e., improved histogram of oriented gradients (ImHOG) is proposed to alleviate the limitation of HOG by concatenating gradient of opposite directions.

3 Proposed Framework

In this section the proposed framework has been described in details. The proposed framework consists of seven main stages: (1) Segmenting color, (2) Subtracting background, (3) Eliminating shadow regions, (4) Labeling, occlusion handling and filtering, (5) Extracting ImHOG feature, (6) Classification and (7) Tracking. Figure 1 shows the proposed framework for detection and tracking of humans.

3.1 Segmenting color

Initially, the input RGB frame is converted to grayscale and HSI frame. The grayscale frame is utilized to increase processing speed of the framework. And the framework uses HSI frame to eliminate shadow regions from foreground, since HSI color space is less sensitive to illumination variations compared to RGB color space. The grayscale and HSI frame is used for background subtraction and shadow elimination process respectively.

3.2 Subtracting background

Instead of representing the value of all the pixels by same dispersion, each pixel values are modeled as a mixture of Gaussians to describe numerous backgrounds. Based on the consistency and the variance of each Gaussian dispersion, the framework decides which Gaussian dispersions may correlate to background colors. Value of pixels that do not correlate to the background dispersions are treated as foreground until any Gaussian dispersion incorporates them with satisfactory confirmation. Usually, a single Gaussian would be enough to model the pixel value resulted from a specific plane under same illumination condition. To deal with uneven illumination condition a single adaptive Gaussian is required for each pixel. However, in real world multiple planes often appear in the field of view of a specific pixel under different illumination condition. Hence, multiple adaptive Gaussian dispersions are required to model numerous backgrounds. The proposed framework represents every pixel by a combination of Gaussian dispersions, as shown in (3.1).

$$P(V_t) = \sum_{(i=1)}^{J} w_{i,t} * N(V_t, \mu_{i,t}, \Sigma_{i,t}) \tag{3.1}$$

Where J denote the quantity of Gaussian dispersions. $\Sigma_{i,t}$, $\mu_{i,t}$ and $w_{i,t}$ denote co-variance matrix, mean and weight at time t of i^{th} Gaussian respectively. And N represent the probability density of Gaussian dispersion. Various Gaussians are supposed to model different intensity values. The weight parameter is introduced to incorporate the time proportion that a color stays in the field of view. The background pixels are separated from foreground pixels by estimating that the background is consist of B highest probable intensity values. The probable background intensity values are the ones which appear in the view frustum for a long time with low variance. Due to different reflecting surfaces a moving object likely to form large clusters in the color space than static single-color objects. As a result, a fitness value is computed for each Gaussians which can be defined as weight divided by standard deviation ($^{w}/_{\sigma}$). The J dispersions are ranked by fitness value and first B dispersions are considered as background representation. To deal with illumination changes these Gaussian dispersions

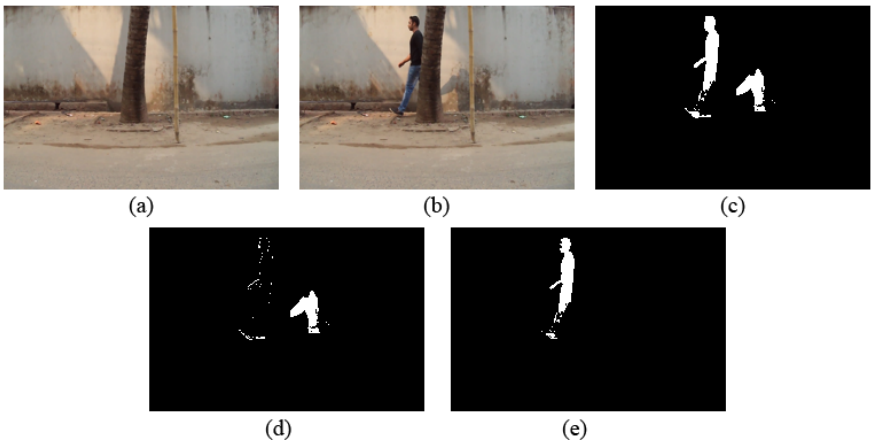

Figure 2: Processing example of shadow elimination process: (a) current frame, (b) background frame, (c) foreground with shadow image, (d) detecting shadow regions and (e) foreground without shadow image.

are selectively updated. Each new pixel value is compared with the J dispersions on the basis of fitness and then the dispersion with the highest ranking is updated. If no match is found, a new Gaussian dispersion is incorporated with mean equal to the unmatched pixel value and a small weighting parameter.

3.3 Eliminating shadow regions

The accuracy of ROI construction relies on generating accurate foreground extraction. As the shadows of an object continually follow the object, background subtraction process considers these shadows as foreground. Beside, these shadows also preserve the geometric properties of an object as a result; those shadows can be misclassified as human. For detecting shadow regions a Hue-Intensity disparity (D_{HI}) value between background and current frame for every pixel is calculated. For pixel X, D_{HI} value is defined as (3.2).

$$D_{HI}(X) = C * H_{Diff}(X) + |\log_e \left(\frac{I_{X,Bg}}{I_{X,Curr}} \right)| \tag{3.2}$$

Where $I_{X,Curr}$ and $I_{X,Bg}$ represent the Intensity of pixel X for current and background frame respectively, C is a constant and $H_{Diff}(X)$ denotes the hue difference between background and current frame for pixel X, which is

calculated by (3.3).

$$H_{Diff}(X) = \min(|hue_{Curr}(X) - hue_{Bg}(X)|, 360 - |hue_{Curr}(X) - hue_{Bg}(X)|) \tag{3.3}$$

Where $hue_{Curr}(X)$ and $hue_{Bg}(X)$ express hue of pixel X for current and background frame respectively. The Hue-Intensity disparity value is used to detect shadow regions by (3.4).

$$Shadow(X) = \begin{cases} 1 & \text{if } D_{HI}(X) < T \text{ and } FS(X) = 1 \\ 0 & \text{otherwise} \end{cases} \tag{3.4}$$

Here $FS(X)$ denotes the value of pixel X for FS image. FS image is the output of background subtraction process containing foreground(s) with shadows and T is a threshold value. For a shadow pixel Y, value of $D_{HI}(Y)$ will be zero as the hue of current and background frame for pixel Y will be same. And the proposed method can detect shadow region if the intensity ratio of background and current frame for pixel Y is at most 2.5. So, threshold T is defined as (3.5). Finally, the shadow regions are eliminated to construct foreground without shadow (FWS) image by using (3.6). Figure 2 illustrates the processing example of shadow elimination process.

$$T = C * 0 + |\log_e(2.5)| \approx 0.91 \tag{3.5}$$

$$FWS = FS * (1 - Shadow) \tag{3.6}$$

3.4 Labeling, handling occlusion and filtering

From the FWS image the framework detects occlusion events. An occlusion event is defined as, if binary large object (BLOB) number in the previous frame is greater than the BLOB number in the current frame and one of the BLOBs in current frame overlaps with more than one BLOBs in the previous frame. After detecting an occlusion event the framework labels individual BLOB in a group by computing likelihood of each pixel belonging to a particular BLOB with the utilization of back-projection histogram and color correlogram.
After correctly labeling grouped objects morphological closing operation is applied to remove holes in the foreground. Then connected component labeling and filtering is used to find ROIs and remove non-human regions. After that, the framework considers circumstances associated with human body which must be fulfilled by the labeled object in order to consider as ROI. The filtering conditions are aspect ratio and solidity of the labeled object.

3.5 Extracting ImHOG feature

The Histogram of Oriented Gradients (HOG) proposed by Dalal et al. [4] is a powerful feature descriptor that uses gradient magnitude and angle information

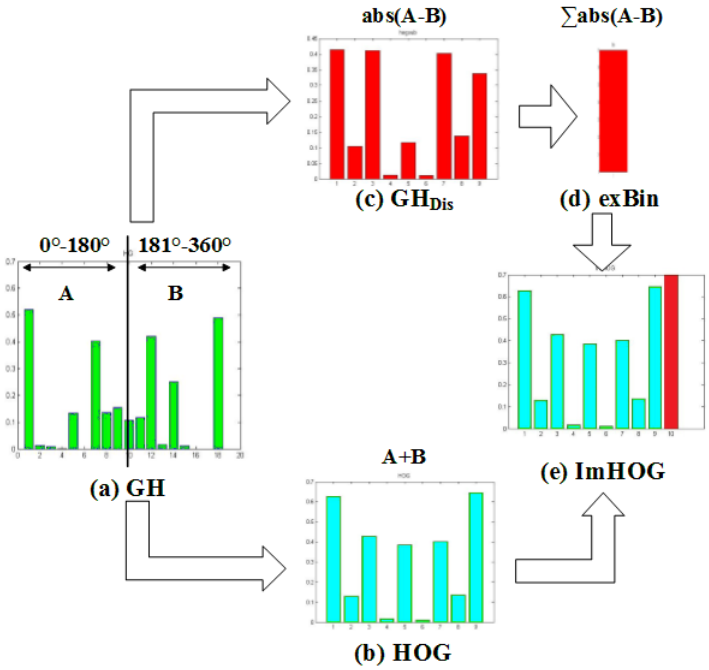

Figure 3: Steps in the construction of ImHoG in a cell.

for human detection. HOG is an improvement of the SIFT descriptor [13] that applied spatial normalization on Gradient Histogram (GH). Dalal et al. [4] experimented with both GH and HOG features for object detection and realized that GH discriminates the circumstances where a luminous human region is in front of a dim background and vice versa because the GH deals with gradient directions from 0° to 360°. For a human detection problem, this discrimination causes a vast intra-class variation. Dalal et al. [4] resolved the problem related to GH by calculating gradients of angle α and $\alpha + 180°$ (reversed orientation) to α only, where $0° \leq \alpha < 180°$. As HOG puts angles of reversed orientations to one histogram bin, some local patterns cannot be properly discriminated by HOG. Thus, it is possible for two distinct patterns to be represented by an identical HOG feature vector. Let $GH(x)$ denotes the value of bin x for sampled gradient angle α and M represents the amount of bins in GH. HOG is the sum of two corresponding bins of GH. As a result, HOG can be computed from GH by adding $GH(x)$ and $GH(x +^{M}/_{2})$, where $1 \leq x \leq^{M}/_{2}$. As HOG minimize GH some key features are lost. To resolve the previously stated issue related to HOG, a new histogram GH_{Dis}, called histogram of gradient disparity is generated by taking absolute difference between $GH(x)$ and $GH(x +^{M}/_{2})$,

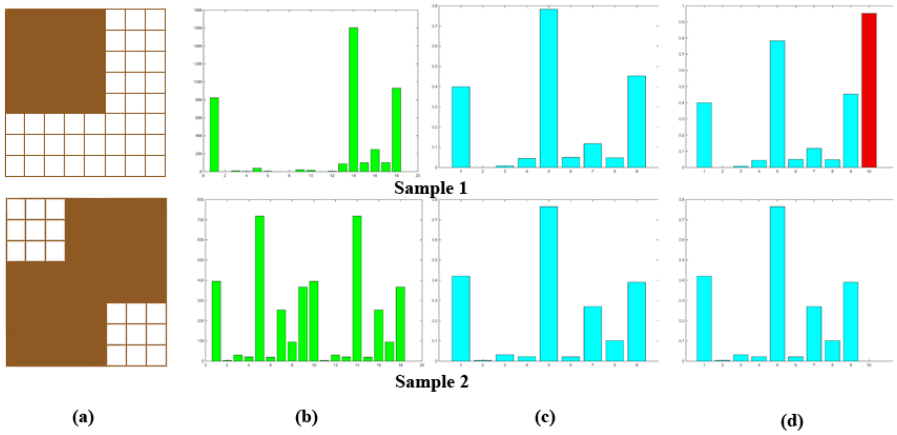

Figure 4: Problem of HOG and its solution by ImHOG : (a) pattern of 8×8 pixels, (b) GH, (c) HOG and (d)ImHOG.

where $1 \leq x \leq^{M} /_{2}$. Then the values of all the bins of GH_{Dis} are summarized into one bin called $exBin$. This new bin $exBin$ can properly discriminate the local patterns which are misclassified by HOG. After that, HOG and $exBin$ are concatenated for every cell in the ROI to generate improved histogram of oriented gradients (ImHOG). Figure 3 shows the construction of ImHOG in a cell.

The ImHOG features for cells are concatenated to generate a block feature. The feature for blocks are then further concatenated and normalized to form feature vector. ImHOG can properly discriminate local patterns misclassified by HOG. Figure 4 explains the situation where HOG represent two distinct patterns by an identical feature vector and its solution by ImHOG. In Figure 4, patterns Sample 1(a) and Sample 2(a) are represented similarly by HOG, which is shown in Sample 1(c) and Sample 2(c). However, the $exBin$ included in ImHOG provides different values for the patterns similarly represented by HOG. As a result, ImHOG can differentiate those patterns, which is shown in Sample 1(d) and Sample 2(d).

3.6 Classification

Finally, the ImHOG feature vector is sent to a linear SVM for human detection. SVM is a supervised margin classifier. For two grouped training dataset, linear SVM intends to find maximum-margin hyperplane, which leads to largest separation between the groups.

3.7 Tracking

Human tracking increases detection rate and framework reliability. Information obtained from prior frames can be utilized to explore formerly classified humans in the present frame. Another application of tracking is to find probable position of a human in the present frame if detection briefly fails due to human is not visible in the field of view.
A Kalman filter [15] has been employed in the proposed framework for tracking humans in sequential frames. The equations related with the Kalman filter are presented in [15]. This subsection presents the framework specific parameters for Kalman tracking. Humans are tracked in the framework by utilizing two specifications i.e. centroid x and centroid y coordinates associated with the detected human bounding box and these specifications are employed to model the measurement trajectory (z). The state vector ($\hat{x}$) and state error covariance matrix (p) are predicted for each detected human at frame k. The state transition matrix A is an identity matrix, as the speed and trajectory of a human does not alter much between frames. For the proposed framewok, A is

$$A = \begin{pmatrix} 1 & 0 & 1 & 0 \\ 0 & 1 & 0 & 1 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \end{pmatrix} \tag{3.7}$$

The resulting updated state vector ($\hat{x}$)

$$\hat{x_k} = \begin{pmatrix} x_{k-1} + \Delta x_{k-1} \\ y_{k-1} + \Delta y_{k-1} \\ \Delta x_{k-1} \\ \Delta y_{k-1} \end{pmatrix} \tag{3.8}$$

The measurement matrix (H) is

$$H = \begin{pmatrix} 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \end{pmatrix} \tag{3.9}$$

The sensitivity of the tracker to updates is determined by the measurement noise covariance matrix R. For human tracking an R matrix of $0.1I$ provides responsive results, where I is an identity matrix.
For each detected human an individual model is initialized. The framework preserves a set of objects S_k currently being tacked at fame k and a set of measurements T_k available on this fame. Let M_k and N_k denote the elements in S_k and T_k respectively. Let D_{M*M} be a distance matrix between $s \in$ S and $t \in T$ such that $D(i,j) = dist(s_i, t_j)$. The framework uses the Hungarian algorithm to find the optimal assignment for the distance matrix. If a human in the present frame is correlated to a human in the previous frame, the framework counts the number of times a human has been detected in a sequence of

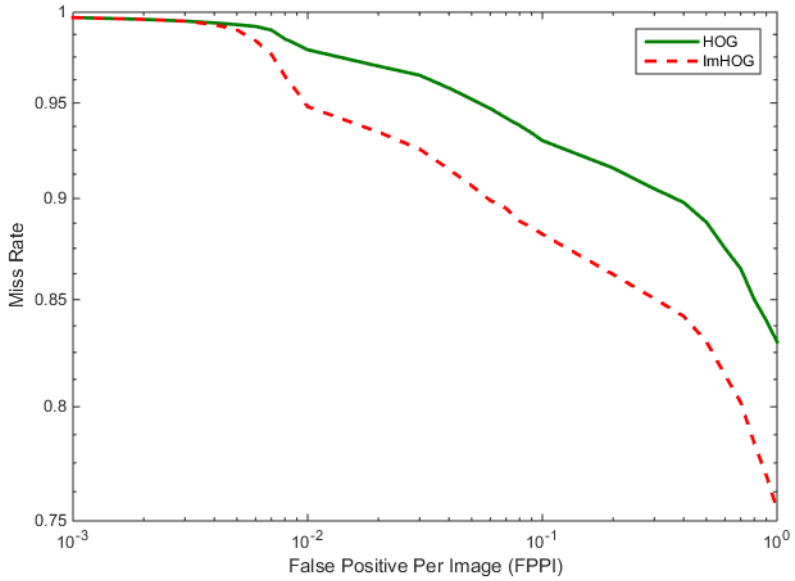

Figure 5: Performance comparison between ImHOG and HOG.

successive frames. If a human is detected in more than six successive frames, a tracker object for the human is created by the framework and a bounding box is drawn on the display around the human. Tracking has also been utilized to reduce false positives in the framework, as false positives likely to appear for a small period of time, (normally 3-4 successive frames). This process of human tracking increases detection rate by reducing false positives than a framework that only utilize human detection.

4 Experimental Results

In this section, experimental results of the proposed framework for detection and tracking of humans are explained. Experiments are performed on Intel Core i5 3.20 GHz CPU and 4 GB RAM memory using MATLAB environment. The performance of the framework is presented in three subsection i.e performance analysis of ImHOG, detection result and tracking result.

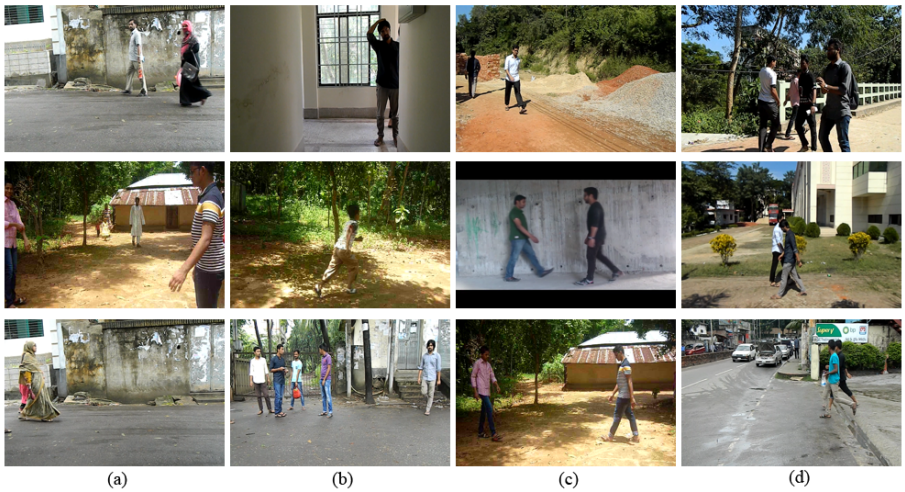

Figure 6: Various types of video frames: (a) different appearances, (b) different poses, (c) uneven illuminations and (d) under occlusion.

4.1 Performance analysis of ImHOG

Performance of the proposed feature vector ImHOG is compared with HOG proposed in [4] using INRIA human dataset. The training dateset consist of 2416 positive images and 121 negative images. And the testing set consist of 1126 positive images and 453 negative images. ImHOG and HOG are computed by considering the cells of 8×8 pixels. Each block consists of 2×2 cells. The features are calculated with a 50% block overlap. Total bins for each cell of ImHOG and HOG are 10 and 9 respectively. L-2 normalization is used for block normalization. Both features i.e. ImHOG and HOG are trained and tested using linear SVM classifier. Figure 5 post the miss rate against the false positive per image (FPPI) for both features as proposed in [14]. The lower the curve the better the performance. At 10^{-1} FPPI, ImHOG rank first followed by HOG. From Figure 5 it can be seen that ImHOG consistently outperforms over HOG.

4.2 Detection result

A robust data set has been collected to evaluate the detection and tracking performance of the proposed framework. Input videos have been captured with a static camera at a rate of 25 fps and a resolution of 320×240 pixels (QVGA) in urban, suburban and rural environments. Most of the humans in the data set are upright standing or walking. In some cases the partial oc-

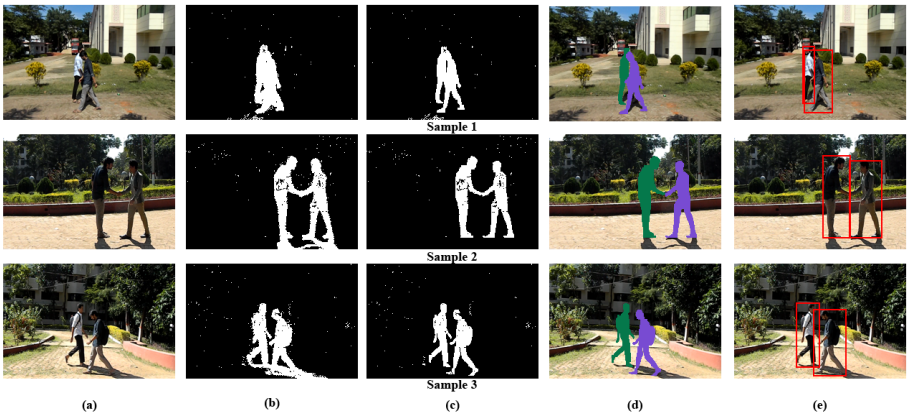

Figure 7: Processing example of human detection: (a) input RGB frame, (b) foregrounds with shadows, (c) foregrounds without shadows (d) labeled foreground(s) and (e) detected object(s).

clusions are occurred. These partial occlusions include humans walking across other objects or other humans. The proposed framework is trained with 140 frames and tested on 1685 frames. For testing, different types of indoor and outdoor frames are utilized, which contain humans in different appearances, poses, uneven illuminations and under occlusion as shown in Figure 6. Furthermore, these test sets are independent from the training sequences. Figure 7 illustrates processing example of human detection. In Figure 7, Sample 1 contains human with human occlusion in outdoor environment. Sample 2 contains humans in different appearance and poses. Finally, in Figure 7, Sample 3 contains humans with foreign objects (backpack). The processing example of these samples show that the framework is robust enough to detect humans in different appearances, poses, uneven illuminations and under occlusion.

The precision and recall value is computed from various types of video frames which are captured from different environment and illumination conditions. Table 1 shows the precision and recall value at various environmental conditions.

The proposed framework shows higher response to indoor and outdoor video frames and also provide satisfactory results for video frames containing complex background. Results extracted from the proposed framework are compared with [2] and [4]. The comparison is performed with respect to precision and recall value as shown in Table 2. According to results shown in Table 2 the proposed framework significantly outperforms over [2]. The framework presented in [2] is not robust enough to detect human in different appearances, poses, uneven illuminations and under occlusion. Furthermore, ImHOG pro-

vides marginally better detection result than [4]. The recall value presented in the Table 2 shows that the proposed framework is robust enough to handle video of normal condition as well as low contrast.

4.3 Tracking result

After a human has been detected successfully, the framework tracks the human in successive frames. If a human is detected in more than six successive frames, a tracker object for the human is created by the framework and a bounding box is drawn on the display around the human. This process of human tracking increases detection accuracy by reducing false positives than a framework that only utilize human detection as shown in Table 3. Some sample tracking results are shown in Figure 8.

5 Conclusion

This paper proposed a framework for detection and tracking of humans, with the goal to track humans from continuous frame sequences with higher adaptability. Initially, the RGB frame is converted to grayscale and HSI frame. Then background subtraction is performed to extract foreground regions. After that, shadow elimination process is used to remove shadow regions from foreground to find the accurate ROI. Then labeling is utilized by using color correlogram for occlusion handling and filtering is employed to remove noises. Afterward, ImHOG feature vector is extracted from ROI and sent to linear SVM for detecting human region. Finally, Kalman filter is utilized for human tracking to increase detection rates and framework robustness. The proposed framework is limited to detect and track humans from videos provided by a stationary

Table 1: Precision and recall value at different environment conditions

Frame type	**Environmental conditions**	**Total frame**	**Precision (%)**	**Recall (%)**	**Avg. time(s)**
Indoor	Normal condition, uneven illumination and low contrast	589	95.5	93.8	0.46
Outdoor		720	93.8	93.6	
Complex Back-ground	Normal illumination	376	92.8	89.7	
	Average	1685	**94.03**	**92.36**	

Table 2: Comparison of results among the proposed framework, [2] and [4]

Framework	**True Positive**	**False Positive**	**False Negative**	**Precision (%)**	**Recall (%)**
The proposed framework	1560	99	129	94.03	92.36
[2]	1156	197	463	85.4	71.4
[4]	1388	127	147	91.6	90.4

Table 3: Detection accuracy before and after tracking

Method	**Detection accuracy(%)**	**False Positive**
Before tracking	87.25	99
After tracking	95.5	34

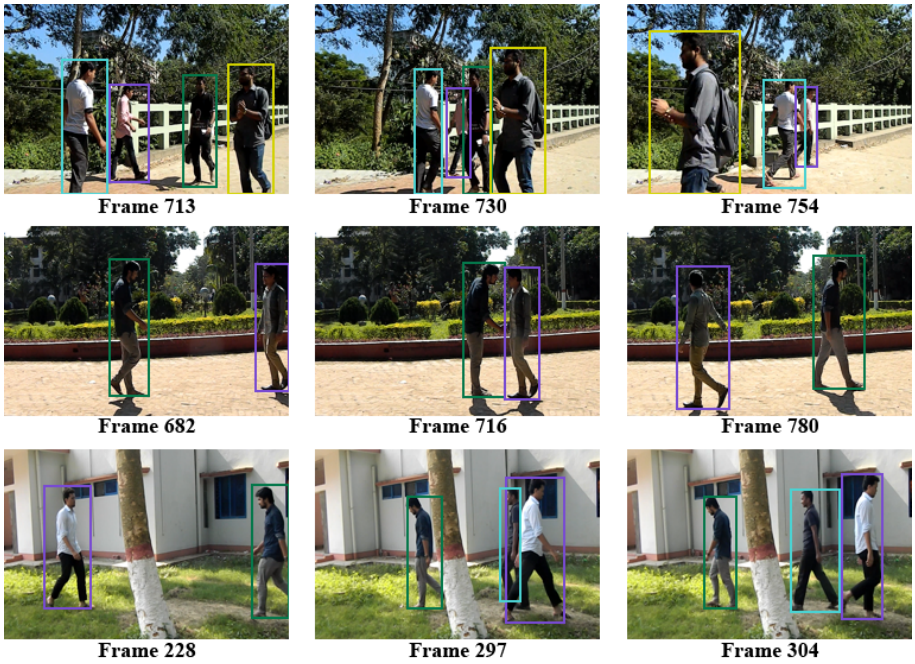

Figure 8: Sample tracking results.

camera. This framework may not provide better results if small portion of an occluded human is visible. This work will be extended to detected and track humans from moving background. And also focus will be given to implement human part-based tracking for better occlusion handling.

References

[1] Y. Ma, L. Deng , X. Chen and N. Guo, *Integrating Orientation Cue With EOH-OLBP-Based Multilevel Features for Human Detection*, IEEE Trans. Circuits Syst. Video Technol., vol. 23, no. 10, pp. 1755 - 1766, Oct. 2013.

[2] R. Tong , D. Xie and M. Tang, *Upper Body Human Detection and Segmentation in Low Contrast Video*, IEEE Trans. Circuits Syst. Video Technol., vol. 23, no. 9, pp. 1502 - 1509, Sept. 2013.

[3] B. Jun, I. Choi, D. Kim, *Local Transform Features and Hybridization for Accurate Face and Human Detection*, IEEE Trans. Pattern Recognit. Mach. Intell., vol. 35, no. 6, pp. 1423 - 1436, June 2013.

[4] N. Dalal and B. Triggs, *Histograms of oriented gradients for human detection* , in Proc. IEEE Int. Conf. Comput. Vision Pattern Recognit., pp. 886893, 2005.

[5] S. Paisitkriangkrai, C. Shen, and J. Zhang, *Fast pedestrian detection using a cascade of boosted covariance features*, IEEE Trans. Circuits Syst. Video Technol., vol. 18, no. 8, pp. 11401151, Aug. 2008.

[6] Y. Mu, S. Yan, Y. Liu, T. Huang, and B. Zhou, *Discriminative local binary patterns for human detection in personal album*, in Proc. IEEE Int. Conf. Comput. Vision Pattern Recognit., pp. 18, 2008.

[7] Y. Chen and C. Chen, *Fast human detection using a novel boosted cascading structure with meta stages*, IEEE Trans. Image Process., vol. 7, no. 8, pp. 14521464, Jul. 2008.

[8] P. F. Felzenszwalb, R. B. Girshick, D. McAllester, and D. Ramanan, *Object detection with discriminatively trained part based models*, IEEE Trans. Pattern Recognit. Mach. Intell., vol. 32, no. 9, pp. 16271645, Sep. 2010.

[9] B. Wu and R. Nevatia, *Detection and segmentation of multiple, partially occluded objects by grouping, merging, assigning part detection responses*, Int. J. Comput. Vis., vol. 82, no. 2, pp. 185204, Apr. 2009.

[10] A. Vedaldi, V. Gulshan, M. Varma, and A. Zisserman, *Multiple kernels for object detection*, in Proc. IEEE 12th Int. Conf. Comput. Vision, pp. 606613, 2009.

[11] K. Kim, T. Chalidabhongse, D. Harwood, and L. Davis, *Real-time foregroundbackground segmentation using codebook model*, J. RealTime Imag., vol. 11, pp. 172185, Jun. 2005.

[12] Y. Liu, H. Yao, W. Gao, X. Chen, and D. Zhao, *Nonparametric background generation*, in Proc. Int. Conf. Pattern Recongnit., vol. 4, pp. 916919, Sep. 2006.

[13] D. Lowe, *[Distinctive image features from scale-invariant keypoints*, International Journal of Computer Vision, vol. 60, no. 2, pp.91110, 2004.

[14] P. Dollar, C. Wojek, B. Schiele, and P. Perona, *Pedestrian detection: An evaluation of the state of the art*, IEEE Trans. Pattern Anal. Mach., vol. 34, no. 4, pp. 743 - 761, April 2012.

[15] G. Welch and G. Bishop, *An introduction to the Kalman filter*, Tech.Rep. TR 95041, University of North Carolina at Chapel Hill, Department of Computer Science, 2003.