U.Porto Journal of Engineering, 8:6 (2022) 178-188
ISSN2183-6493
DOI: 10.24840/2183-6493_008.006_0013

Received: 21 September, 2021
Accepted: 20 August, 2022
Published: 28 November, 2022

Advanced Perspective on Human Detection system with Hybrid Feature Set

Manikanda Prabu Nallasivam[1], Vijayachitra Senniappan[2]

[1]Associate Professor, Department of Electronics and Communication Engineering, Nandha Engineering College, Erode, Tamil Nadu, India - 638052 (manikandaprabube@gmail.com) ORCID 0000-0002-7117-1865; [2]Professor, Department of Electronics and Instrumentation Engineering, Kongu Engineering College, Erode, Tamil Nadu, India - 638052 (dr.svijayachitra@gmail.com) ORCID 0000-0002-0339-4303

Abstract

Detecting and discriminating humans in video frames for surveillance applications is a demanding task. Identifying and highlighting humans by eliminating shadows from the video frames is vital for prudence motive. In this paper, a three-step procedure is proposed, which includes motion detection by background subtraction in live video frames, morphological gradient-based shadow removal, and human detection by Hybrid Feature Set (HFS), which comprises Histogram Oriented Gradient (HOG) and Local Binary Pattern (LBP) with adaptive Neuro-Fuzzy inference system. The first step incorporates static background subtraction and dynamic background subtraction. The second step is to remove shadows by using a morphological gradient with the horizontal directional mask. The third step includes near-field, mid-field, and far-field human detection by using an adaptive Neuro-Fuzzy inference system. The results obtained from the various performed experimental analysis demonstrates diverse parametrical measures, which outperforms comparatively when benchmark databases and real-time surveillance video frames were used.

Author Keywords. Feature Extraction. Image Motion Analysis. Subtraction Techniques. Morphological Operations. Machine Vision.

Type: Research Article
Open Access Peer Reviewed CC BY

1. Introduction

Surveillance cameras are most important in crowd-migrating areas like airports, railway stations, banks, malls, etc. Many surveillance systems with Artificial Intelligence (AI) aims to detect accidents, human gait recognition, gender classification, and crowd analysis. The optimization of an existing algorithm leads to an increase in the rate at which humans can be detected. In a complex environment, the initial task of extracting the required information is to discriminate foreground and background details of the continuous images acquired by the surveillance system. This discriminated image is further taken for extracting the foreground information by suppressing the shadow and background details. An adaptive method for static and dynamic background modeling is proposed and the consequential results are considered for human detection. The shadow of the moving object leads to false detection in identifying humans. The shadow information is available only in grayscale and can be eliminated by considering the RGB color space. This grayscale shadow information can also be removed by wavelet transform. The shadow-less image is engendered by the horizontal decomposition of the wavelet transform. In the proposed approach, time complexity is considered an important parametrical measure for removing shadows using the horizontal gradient of video frames. Even when the shadow features are not matched, the Neuro-Fuzzy classifier does not have any constraint in identifying humans. Complete removal of the shadow is not required in this

case; an average of 60% of the shadow of each individual moving target has to be removed to proceed with the feature extraction process. The proposed shadow removal procedure achieves removing the shadow of 95% by morphological gradient with the horizontal mask. Further, a hybrid feature set which is a combination of the Histogram Oriented Gradient (HOG) and Local Binary Pattern (LBP), is used for extracting the required human features from the foreground image. Proceeding with the detection of humans among various moving objects in the video frame is achieved with the Neuro-Fuzzy classifier.

2. Related Works

An intelligent surveillance system for identifying and discriminating humans from other moving objects with a prudent outcome has too much societal demand. Few related works carried out by researchers are stated; Chen et al. (2014) proposed a Bayesian framework for background modeling by using the most significant principle features at each pixel. The various stages of the algorithm, such as change detection, change classification, foreground segmentation, and background maintenance, were tested in different environmental situations. Zhu and Zeng (2016) proposed a non-parametric model for background subtraction, in which segmenting of moving objects from the continuous images acquired by a static camera was incorporated. The system uses pixel intensity-based background estimation for detecting the most sensitive moving targets, which suits both gray and color images and reduces the false alarm rate. The human shadows detected in the foreground image lead to distraction in identifying the human and are degraded by using colored information. The chromaticity coordinates of primary colors are insensible for changes that occur due to shadow. Chen et al. (2012) proposed a hierarchical background model which primarily uses a mean-shift algorithm for segmenting background images into several regions. For background subtraction, the Gaussian mixture model and pixel models were used. Gaussian mixture model extracts the histogram of Region Of Interest (ROI) and the pixel model depends on the co-occurrence of image variation described by HOG. This system helps in detecting and tracking moving objects from the continuous images acquired by static and dynamic cameras.

Tian et al. (2011) proposed an abandoned object detection and human detection approach in complex surveillance situations. The human detection framework detects humans in the near-field, midfield, and far-field approaches. In the near-field approach, the resolution was more than enough to extract the ROI. Nearly 4000 faces with size 24x24 were used to train the learning classification algorithm. The midfield human detection approach uses the head and shoulder for detecting humans in low-resolution images. In the far-field person detection approach, blob analysis was used to detect the human.

Ko, Son, and Nam (2015) focused on the Bayesian-based face part detector by using a random forest classifier. Edgelet features were used for detecting and tracking humans in both static and dynamic backgrounds. The part detectors were trained by boosting the weak classifiers with edgelet features. The head, shoulder, torso, and legs are the human parts used for part detection and a full-body detector had also been trained. This system tested with INRIA human dataset and produced a 3% increased detection rate.

Liu et al. (2013) proposed a local Support Vector Machine (SVM)-based approach for human action recognition. The local space-time features with SVM for motion recognition outperform the major benchmark database. Improvement in reliability based on local features had been observed in the static background and remained as an exemption for dynamic background. Cheng, Huang, and Ruan (2011) proposed a two-stage background estimation module that

recognizes abrupt changes in illumination. To obtain high-quality background, rough training was followed by the precise training procedure. The motion detection was computed by taking the absolute difference between the current frame and background frame. The threshold-based binary mask generated for background modeling and usage of the effective threshold training procedure supports an automatic threshold computation. Quantitative analysis performed using recall, precision, and similarity metrics was compared with other state-of-art methods by Nallasivam and Senniappan (2021).

Jeon et al. (2015) proposed the detection of humans based on the background image generation by FIR (Far-Infra Red) camera. The background difference detector produces a monochromatic difference image using the current frame and background frame. The highly sensitive, accurate image was obtained by considering the individual color component difference between the current frame and background frame. The background image was updated for every 25th frame in the live video sequence.

Park et al. (2012) proposed a set of Related Histogram Oriented Gradient (RHOG) features and a human detection framework composed of AdaBoost and SVM. The concatenation of elementary HOG feature discriminates video frames into variable-sized blocks and further divides these variable-sized blocks into four blocks known as cells. Each cell consists of nine bins ranging from 0 to 180-degree orientation. The integral imaging technology and convoluted trilinear interpolation with the AdaBoost algorithm were used for selecting the faster elementary HOG feature from the feature pool. The human detection rate was improved using a cascade rejecter and SVM classifier.

3. Human Detection System

Closed-Circuit Television (CCTV) cameras, such as static or dynamic type cameras, were used for condition monitoring in surveillance applications. Static cameras are immobile and cameras that rotate along 360 deg or cameras placed on moving objects/vehicles are considered a dynamic type of camera. Video frames were extracted from live video sequences with a frame rate of 30fps and every 15th frame of the video sequence was considered for motion analysis. The overall block diagram for Human Detection System is shown in Figure 1.

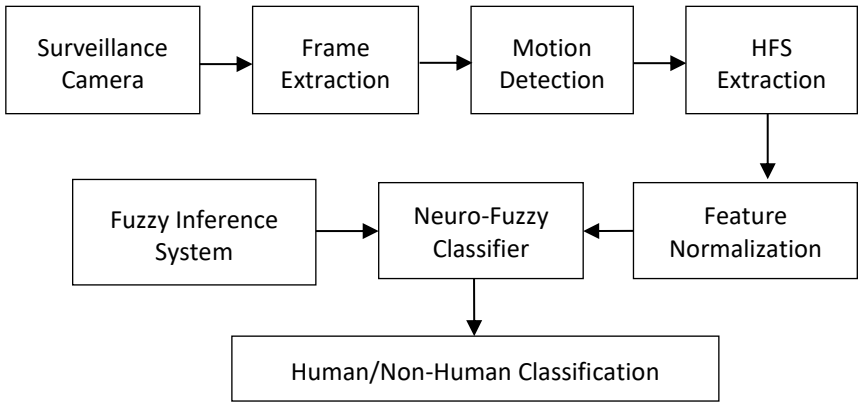


Figure 1: Overall block diagram for Human Detection System

Each 15th frame extracted was considered in the process of background subtraction from which the foreground information was acquired by the running average background subtraction algorithm.

The background image $B_t(x, y)$ was computed by considering the running average between the current frame $I_t(x, y)$ and the previous frame $B_{t-1}(x, y)$ of the video sequence. The equation for running average computation is given below:

$$B_t(x,y) = (1-\alpha).B_{t-1}(x,y) + \alpha.I_t(x,y) \quad (1)$$

Where α denotes an adjustable parameter for achieving fast background adaption and an improved background adaption could be achieved by increasing the α parameter. The flow of the motion detection process using running average background subtraction algorithm is shown in Figure 2.

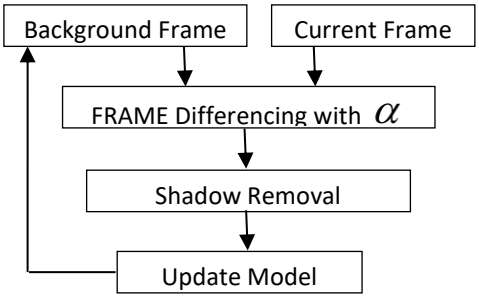
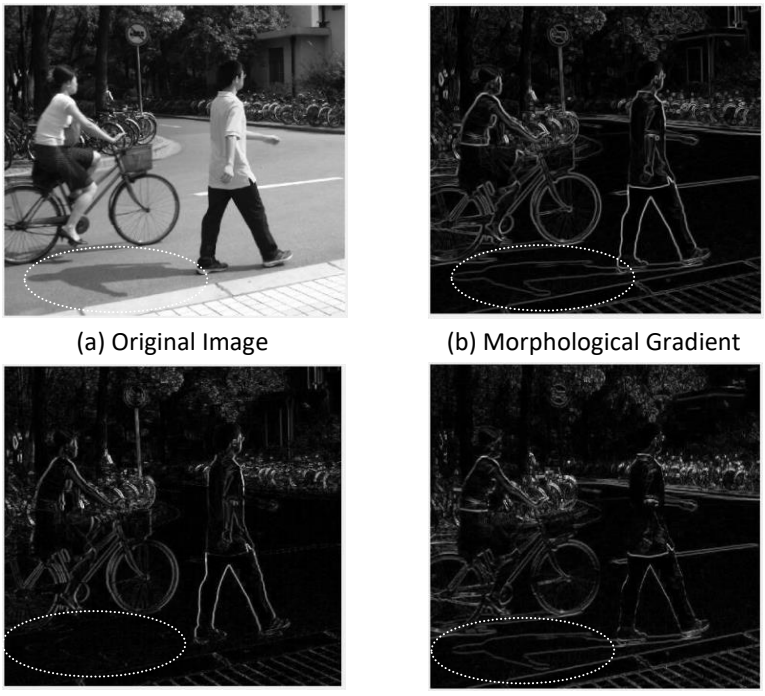


Figure 2: Flow of motion detection using running average background subtraction algorithm

The perfect identification of the region boundary of moving objects is essential to discriminate the foreground and background objects. The individual computation of morphological gradient, horizontal gradient, and vertical gradient is shown in Figure 3 for the shadow removal process.



(a) Original Image (b) Morphological Gradient

(c) Horizontal Gradient (d) Vertical Gradient

Figure 3: Morphological gradient based shadow removal

The shadows and noises were removed by using a mathematical morphological process which includes morphological gradient computation by considering the difference between morphological dilation and morphological erosion, as shown in Equation (2) and Equation (3). The complete removal of shadows and noises was achieved by incorporating horizontal and vertical masks of the directional gradients.

$$HG = Dilation\,(Img, HMask) - Erosion\,(Img, HMask) \quad (2)$$

$$VG = Dilation\,(Img, VMask) - Erosion\,(Img, VMask) \quad (3)$$

Where HG: Horizontal Gradient & VG=Vertical Gradient.

Figure 3(b) shows the morphological gradient of an image with shadows and a similar pattern of the image was also obtained by considering the vertical gradient of the image, as shown in Figure 3(d). The images with shadows in the morphological gradient and vertical gradient are encircled and the shadow removed in the horizontal gradient is also encircled in Figure 3(c), highlighting the differences between the mathematical morphological gradient computational processes of shadow removal. The horizontal gradient in which the shadow removal process was achieved is considered for further feature extraction process.

4. HFS Extraction

The limitation of the feature-based human detection by Histogram Oriented Gradient does not support achieving the required outcome based on texture process. This limitation could be overcome by combining the process of Local Binary Pattern along with Histogram Oriented Gradient, as Local Binary Pattern had set a drawback of low detection rate when used separately. Combining Histogram Oriented Gradient and Local Binary Pattern initiates Hybrid Feature Set formation, which renders major support in identifying humans along with Adaptive Neuro-Fuzzy Inference System.

Computation of magnitude $|\Delta f(x,y)|$ and orientation $\theta(x,y)$ of the gradient $\Delta f(x,y)$ is the two steps included in the Histogram Oriented Gradient. The human detection window size is 64x128, and it is divided into 16 x 16-pixel image patches (sub-images 7X15). Each image patch consists of four cells and individual cell has 8 x 8 pixels. In each cell, the orientation from 0 to 180 degrees consists of 9 bins, i.e., $i \times \frac{pi}{9}$ where i=0, 1,…, 8.

Figure 4 shows the 3X3 neighborhood Local Binary Pattern descriptor.

45	43	42	→	1	1	1
40	39	31		1		0
10	16	30		0	0	0

Figure 4: A 3X3 neighborhood LBP texture descriptor

Each block consists of 4 cells x 9 bins=36 features and each sub-image consists of 7x15x36=3780 features in total. Ojala, Pietikäinen, and Mäenpää (2002) introduced the Local Binary Pattern, which is a powerful descriptor of texture. While considering a 3 x 3 neighborhood, each cell value p_b is a threshold with center pixel p_c and is described in binary form.

$$LBP_{x,y} = \sum_{b=0}^{b-1} s(p_b - p_c)2^b \quad (4)$$

In the above Equation (4) $LBP_{x,y}$ is the Local Binary Pattern value of the center pixel p_c. $s(p_b - p_c)$=1 if $(p_b - p_c) > 0$ and otherwise 0. The Binary Code is 11100001 and its decimal equivalent is 225.

5. Feature Normalization

It was observed that the pixel value varied when the same view was captured using different imaging sensors. These captured pixel values were classified with a learning classification algorithm. Despite variations in the pixel values, the acquired details remain the same. The scaling and normalization of the obtained features are required to be restricted within a single scale. For feature extraction, Hybrid Feature Set was used, while feature normalization is not

as essential in Adaptive Neuro-Fuzzy Inference System for classification. On considering the actual feature values, we observe that there is an increase in time complexity. The min-max algorithm was deliberated for normalizing the features on a single scale.

The time complexity was reduced by 25% by the process of optimization. The different ranges of feature values obtained are rescaled between 0 and 1 with Modified Min-max algorithm. The preferred scale limits are 0 at the minimum and 1 at the maximum. Let us consider the feature values as F1, F2, F3...FN. The algorithm for feature normalization is given as follows,

Step1: Extracting feature values (F1,F2,F3…Fn) of 10000 frames.

Step2: Finding the maximum value of each feature (max(F1), max(F2),max(F3)….max(Fn)).

$$F_{i\ \max} = \max\left[F_{iI1}, F_{iI2}, F_{iI3}, \ldots F_{iIn}\right] \tag{5}$$

Where i= number of Features

$$F_{i\ \max} = \max\left[F_{ik}\right]_{k=I_1}^{k=I_n} \tag{6}$$

Where k is the images ranging from I1 to In.

Step3: Dividing each feature value by the maximum value (Eg: Each value of F1/max (F1))

$$FeatureNormalization = F_{Norm} = \frac{F_i I_n}{\max\left[F_{ik}\right]_{k=I_1}^{k=I_n}} \tag{7}$$

Where Equation (5) and Equation (6) support finding maximum of each feature. Each feature value (F1) is divided by its maximum value (max(F1)). The maximum feature value obtained and each feature divided for feature normalization is stated in Equation (7). As a result, all the feature values come under a single scale (from 0 to 1). When the system has implemented the feature, values are assigned in the floating-point format to achieve good precision.

6. Neuro-Fuzzy Classifier

ANFIS, a type of adaptive neural network with a combination of fuzzy and artificial networks based on the Takagi-Sugeno fuzzy inference system, comprises five layers. The fuzzification layer generates the membership function based on the input given in the first layer and the second layer frames the fuzzy rules, followed by normalization and defuzzification as the third and fourth layers. The output is taken from the final fifth layer. The Sugeno model generates Root Mean Square Error (RMSE) by comparing the actual target with the output obtained. On getting adapted, it also assists the system in learning from the training data set. Grid partitioning and subtractive clustering are the two possible algorithms for generating the fuzzy inference system (Lalli et al. 2014). We consider a subtractive clustering algorithm for generating fuzzy inference system and were also trained by a hybrid learning algorithm. The values for the parameters acquired for clustering GENFIS are the range of influence: 0.05, squash factor: 1.25, accept ratio: 0.5 and reject ratio: 0.15. Averages of 10000 real-time images were considered to train the system with a processing time of 960 seconds. Table 1 shows the comparison of training time with different real-time video frame sizes.

Video Frame Size	Training Time in Seconds
854 x 480 (480p)	945
1280 x 720 (720p)	960
1920 x 1080 (1080p)	1063

Table 1: Comparison of training time with different real-time video frame sizes

The Mean Square Error for the training process is 0.0001 and zero error tolerance was fixed in the training process in Adaptive Neuro-Fuzzy Inference System. The time taken for training

different-sized video frames in real-time with Adaptive Neuro-Fuzzy Inference System is shown in Table 1. In the proposed human detection system video frame size of 1280 x 720 (720P) with a training time of 960 seconds is considered. The ROC curve for Adaptive Neuro Fuzzy Inference System for Human Detection is shown in Figure 5.

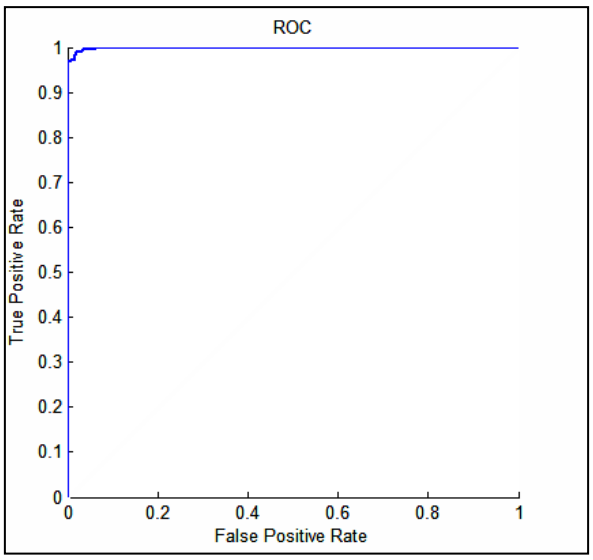


Figure 5: Receiver Operating Characteristics

The classifier performance was evaluated by measuring the Receiver Operating Characteristics (ROC) curve. True Positive rate and False Positive rate are the parameters considered for plotting the ROC curve where the False Positive Rate is defined as (1-Specificity) and the range lies between 0 & 1.

7. Performance Evaluation

Precision, Recall, and Errors on Shadow (T_{Shadow}) are the parameters used to evaluate the performance of the proposed system. The Gaussian Mixture Model (GMM), Visual Background Extraction (VBE), and Normalized Cross-Correlation (NCC) are the three related works considered for comparative analysis. The Precision, Recall, and Errors on Shadow (T_{Shadow}) were computed using Equations (8), (9) and (10). Figure 6 shows the Recall, Precision and Accuracy values versus various approaches.

$$\mathrm{Pr}ecision = \frac{TP}{(TP + FP)} \tag{8}$$

$$\mathrm{Re}call = \frac{TP}{(TP + FN)} \tag{9}$$

$$T_{Shadow} = \frac{nse}{FP} \tag{10}$$

Methods	Recall in %	Precision in %	Accuracy in %
GMM	98	93	96
VBE	93	96	95
VBE+NCC	83	99	81
Proposed HFS+ANFIS	95	99	96

Table 2: Comparison of Recall, Precision, and Accuracy with three related works

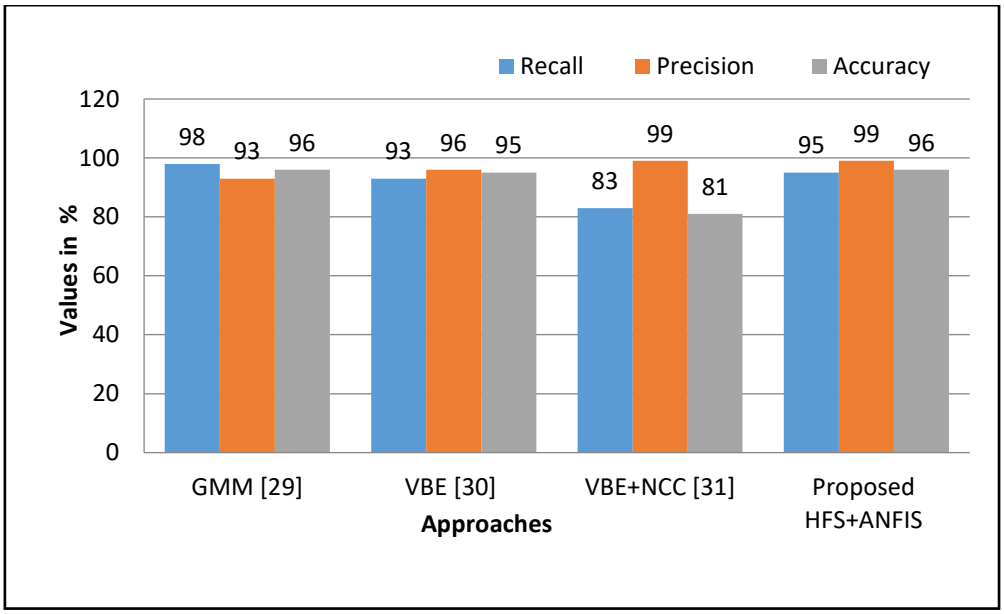


Figure 6: Recall value for various approaches

Where TP (True Positive) denotes the number of positive information that is correctly hit by a classifier. TN (True Negative) denotes the number of negative information that is correctly hit by a classifier. FP (False Positive) denotes the number of positive information that is incorrectly hit by a classifier. FN (False Negative) denotes the number of negative information that is incorrectly hit by a classifier. *nse* is the total number of false positives formed in the shadow area according to ground truth.

As shown in Figure 6, the precision value of the proposed approach is around 99% and recall is 95%. Table 2 shows the values of Precision, Recall, and accuracy. The recall value is nearly equal to the superior Gaussian Mixture Model. The maximum value is achieved in Precision which is equal to the combination of Visual Background Extraction and Normalized Cross-Correlation. The accuracy for proposed human detection using Adaptive Neuro-Fuzzy Inference System with Hybrid Feature Set is 96%.

Methods	Error on Shadow in %	Time complexity in seconds
GMM	22	0.62
VBE]	50	0.32
VBE+NCC	11	0.21
Proposed HFS+ANFIS	03	0.22

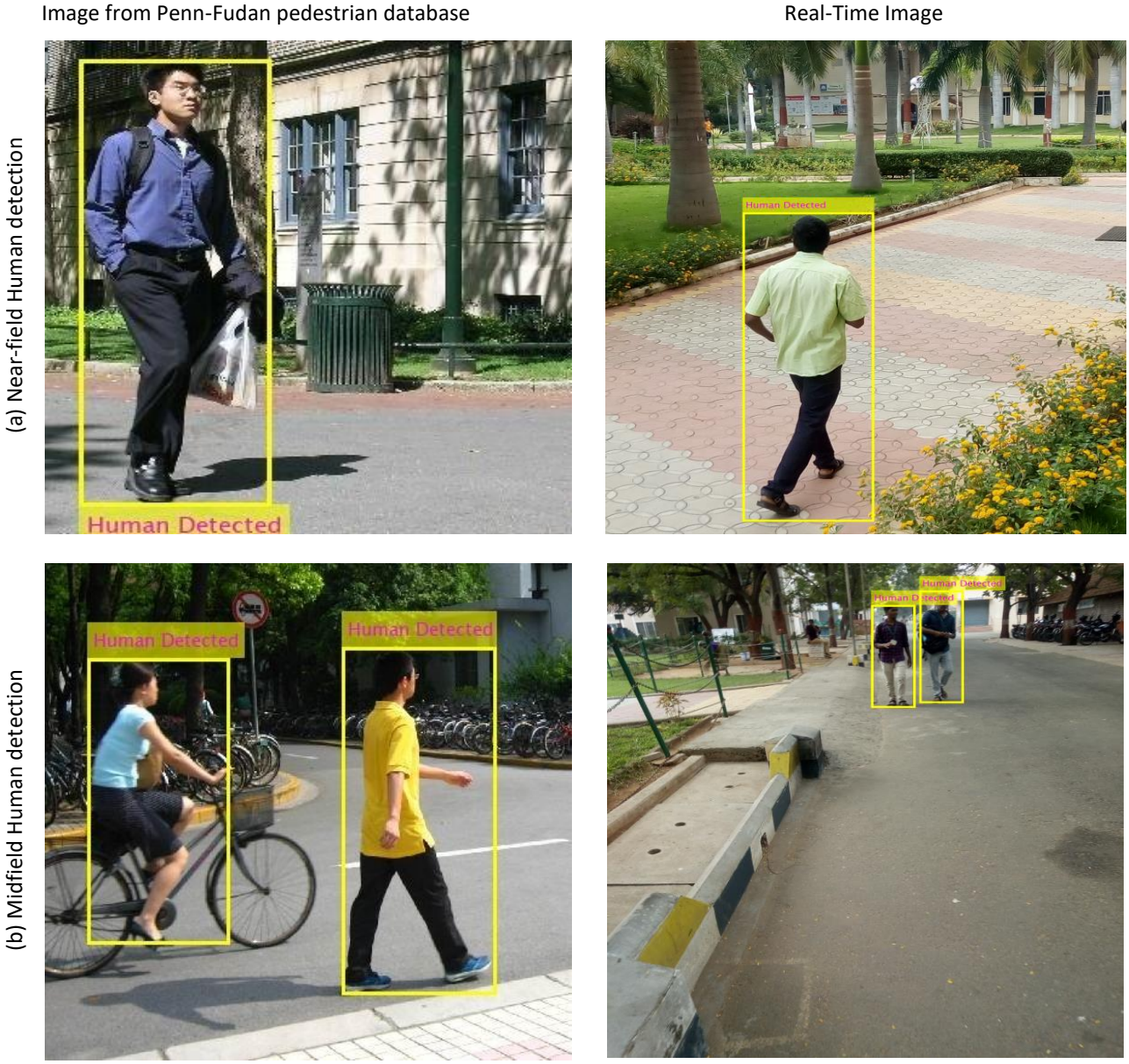
Table 3: Comparison of Error on shadow and time complexity with three related works

Table 3 shows the error on shadow and time complexity for GMM, VBE, VBE+NCC and the proposed system HFS+ANFIS. The error that occurred due to shadow is 3%, which is the least value achieved. Achieving the lowest value in Errors on Shadow would be considered the best outcome in the shadow removal process.

Morphological gradient based shadow removal technique for Human Detection achieves a percentage of 3 in error on the shadow. The time complexity is calculated by computing the average CPU time by testing 100 samples from Caltech Pedestrian detection benchmark database. The average computing time calculated for processing a single frame was around 0.22 seconds in Intel core i3-2365M CPU @ 2.8GHz. The time complexity is 0.21 for VBE+NCC and it is the lowest time complexity, but the accuracy for the same is too low when compared with the other approaches; hence does not give a satisfactory outcome. By considering both

time complexity and accuracy parameters, the proposed HFS+ANFIS shows significant outperforming results.

Human detection results from the Penn-Fudan pedestrian database and the results of real-time surveillance video are shown in Figure 7. The human migrants are identified and indicated using a bounding box. Figure 7(a) shows a properly estimated human detection in the near field, Figure 7(b) shows the human detection in midfield with the human structure estimated with some tolerance, Figure 7(c) shows the human detection in far-field and the human structure was not estimated properly. This limitation is overcome by framing the training image set with variant scaling. To avoid increased time complexity in a crowded environment, the individuals migrating together are remapped again onto a single bounding box.



(c) Far-Field Human detection

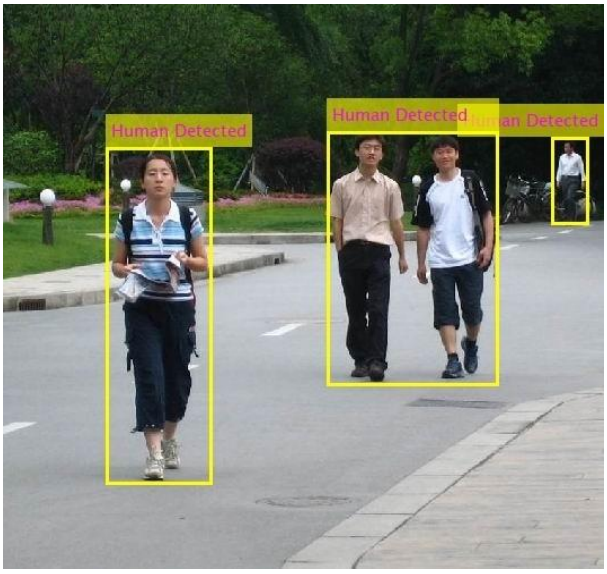


Figure 7: Human detection Results of Images from Penn-Fudan pedestrian database and Real-time images

8. Conclusion

This paper discussed the morphological gradient-based shadow removal technique for human detection using ANFIS with HFS algorithm. The false detection because of shadows in the video frame is rectified by using the horizontal gradient in the background subtraction process. The result shows that the proposed system outperforms in detecting humans in the near field, midfield and far-field scenarios. The system acquires every 15^{th} frame in the video sequence for processing in order to reduce the time complexity. The time consumption for detecting multiple numbers of humans at an instant in the video frame is getting reduced and outperforming results were achieved by using Neuro-Fuzzy classifier. The performance of the system was evaluated by using the parameter precision, recall, and shadow detection. It shows 95% recall and 99% precision and error due to shadow is 3%. The proposed methodology works in color for day vision video frames and in grayscale for night vision video frames.

Reference

Chen, S., J. Zhang, Y. Li, and J. Zhang. 2012. "A hierarchical model incorporating segmented regions and pixel descriptors for video background subtraction". *IEEE Transactions on Industrial Informatics* 8, no. 1: 118-27. https://doi.org/10.1109/TII.2011.2173202.

Chen, P. Y., C. C. Huang, C. Y. Lien, and Y. H. Tsai. 2014. "An efficient hardware implementation of HOG feature extraction for human detection". *IEEE Transactions on Intelligent Transportation Systems* 15, no. 2: 656-62. https://doi.org/10.1109/TITS.2013.2284666.

Cheng, F. C., S. C. Huang, and S. J. Ruan. 2011. "Scene analysis for object detection in advanced surveillance systems using laplacian distribution model". *IEEE Transactions on Systems, Man and Cybernetics Part C: Applications and Reviews* 41, no. 5: 589-98. https://doi.org/10.1109/TSMCC.2010.2092425.

Jeon, E. S., J. S. Choi, J. H. Lee, K. Y. Shin, Y. G. Kim, T. T. Le, and K. R. Park. 2015. "Human detection based on the generation of a background image by using a far-infrared light camera". *Sensors* 15, no. 3: 6763-88. https://doi.org/10.3390/s150306763.

Ko, B. C., J. E. Son, and J. Y. Nam. 2015. "View-invariant, partially occluded human detection in still images using part bases and random forest". *Optical Engineering* 54, no. 5: Article number 053113. https://doi.org/10.1117/1.OE.54.5.053113.

Lalli, G., N. Manikandaprabu, D. Kalamani, and C. N. Marimuthu. 2014. "A development of knowledge-based inferences system for detection of breast cancer on thermogram images". In *2014 International Conference on Computer Communication and Informatics: Ushering in Technologies of Tomorrow, Today, ICCCI 2014*, Article number 6921743. https://doi.org/10.1109/ICCCI.2014.6921743.

Liu, H., T. Xu, X. Wang, and Y. Qian. 2013. "Related HOG features for human detection using cascaded adaboost and SVM classifiers". In *Advances in Multimedia Modeling*, 345-55. Lecture Notes in Computer Science, vol. 7733. https://doi.org/10.1007/978-3-642-35728-2_33.

Nallasivam, M., and V. Senniappan. 2021. "Moving human target detection and tracking in video frames". *Studies in Informatics and Control* 30, no. 1: 119-29. https://doi.org/10.24846/v30i1y202111.

Ojala, T., M. Pietikäinen, and T. Mäenpää. 2002. "Multiresolution grayscale and rotation invariant texture classification with local binary patterns". *IEEE Transactions on Pattern Analysis and Machine Intelligence* 24, no. 7: 971-87. https://doi.org/10.1109/TPAMI.2002.1017623.

Park, W. J., D. H. Kim, Suryanto, C. G. Lyuh, T. M. Roh, and S. J. Ko. 2012. "Fast human detection using selective block-based HOG-LBP". In *Proceedings - International Conference on Image Processing, ICIP*, 601-04. https://doi.org/10.1109/ICIP.2012.6466931.

Tian, Y., R. S. Feris, H. Liu, A. Hampapur, and M. T. Sun. 2011. "Robust detection of abandoned and removed objects in complex surveillance videos". *IEEE Transactions on Systems, Man and Cybernetics Part C: Applications and Reviews* 41, no. 5: 565-76. https://doi.org/10.1109/TSMCC.2010.2065803.

Zhu, T., and P. Zeng. 2016. "Background subtraction based on non-parametric model". In *Proceedings of 2015 4th International Conference on Computer Science and Network Technology, ICCSNT 2015*, 1379-82. https://doi.org/10.1109/ICCSNT.2015.7490985.