A Robust Human Detection System

Amit Kumar Gautam
Department of Electronics and Communication Engineering, Delhi Technological University Delhi, India
amitgautam.cicdu@gmail.com

Ajay Kaushik
Department of Computer Science and Engineering, Kurukshetra university, Haryana, India
ajaykaushik777@gmail.com

Subradeb Choudhary
Department of Electronics and Communication Engineering, Adamas University, West Bengal
subradeb.2010@gmail.com

Abstract— **In the present scenario, Video Surveillance is receiving tremendous attention. It has a wide range of applications such as it can be used in Border areas of a country, in market areas and also in restricted areas for monitoring objects. Human Detection is a field of Video Surveillance where monitoring of humans take place i.e. the human is detected first and its trajectory is estimated for the purpose of monitoring. In this paper, a robust human detection system is proposed. The Human Detection System consists of two stages. The first stage involves Image Pre-processing where the Motion region is extracted and Image Segmentation is applied to this motion region. The second stage classifies the segmented image as a human or a non-human based on Aspect Ratio of Human. So, we can say that the Motion region is incorporated with the Aspect Ratio feature to propose a Robust Human Detection Method. A Dataset is made where the background colour matches with the Human Skin Colour. In this situation, it is very difficult to track the human. We propose a system where we can track human under such conditions. The system is tested in PETs Database also and an overall Detection Rate of 85% is reported. However, the Detection rate gets reduced drastically when the human is occluded in the scene.**

Keywords— Frame Differencing, Aspect Ratio, Human Detection, HoG.

I. Introduction

The main aim of computer vision is to visualize an object in a frame and track that object in a video sequence. Recent Researches in the field of Computer Vision has increased its focus on observing humans, understanding their appearance and activities that are kept under continuous surveillance, intelligent control and human computer interaction. The challenge in Human Detection System is to detect people with number of poses with variable clothing and appearance in complex backgrounds as well as occluded objects. The problem is further compounded by ambiguous background clutter, changing of lighting conditions, shadowing and self-shadowing. The primary step for monitoring people activities is by classifying the movement of various objects. We can summarize a Human Detection and Tracking System as follows. The first and the foremost requirement is an appearance model which is needed to be constructed correctly. Then, detection is performed based on this model to get the correct position of the person. Finally, tracking is done by continuous detection of person in consecutive video frames to obtain the correct trajectory of person. Researches on Human Detection is not only based on detection of single person but also based on detection of multiple persons. Many detectors are proposed to detect humans from still images. Characteristic information of human is extracted (Gradient Information which has a powerful discriminative property and can represent the shape of the human body well) to distinguish human from another object. If human detection from video is required, motion information can be used. Motion caused by human is different from motion caused by other objects. Robustness is an important factor for detection. Many systems cannot read motions accurately or otherwise cannot function optimally due to factors such as background or lighting changes. Depending on different applications the requirements for detection and tracking may differ significantly, e.g. detection of single object is much easier as compared to detection of multiple objects.

Human detection in Computer Vision indicates a category of methods that estimate locations of human bodies in images or video frames. Objects to be located can be full or part of (usually upper half) human bodies. Methods of human detection can be divided into two categories: Signal-processing-based methods, which classify objects by matching specific spatial information, and machine-learning-based methods, which statistically learn models and classify objects. Human tracking methods are applied to estimate movement of human bodies in frame sequences. Movement of human bodies can be achieved by analysing differences between frames, or finding out connections between located human bodies in consecutive frames

The requirements for detection and tracking may differ significantly depending on different applications. Detection of single object is much easier as compared to detection of multiple objects. Accuracy is a significant issue in Multiple Human Tracking. The trajectory of humans is important for surveillance. A multiple people tracking algorithm should make sure that the system tracks the same person under different situations such as temporary people occlusion. There is always a trade-off between precision and computing time because if we want to increase the precision the computing time increases and if we want to decrease the computing time, the precision also decreases. In almost all the circumstances, robustness and accuracy needs to be satisfactory then algorithm optimization and speed of the system are taken into consideration for improvement of performance of the system.

II. Related works

The purpose is to detect human in a video sequence or an image. The techniques used to detect humans can be broadly classified into two categories. One is by Background Subtraction or Segmentation and the other is to detect humans directly without any pre-processing.We do a pre-processing of the image (the scene is captured first) and then subtraction is done frame by frame to detect the object. This method is known as Background Subtraction. In direct method of detection, frame differencing is performed. The difference of present frame with the previous frame is found and if the object is moving, then the object can be detected.

The challenges and need behind the design and implementation of Content based video retrieval (CBVR) system has captured the attention of researchers. A. Sambath and A. Nirmala have presented work on CBVR [1]. It includes various steps: input video, frames extraction, Features extraction. After that extracted video is matched with featured database to get the desirable output. A. Eweiwi et.al [2] have presented work on temporal key poses for human action recognition. This human action is recognized by using Motion energy images and motion history images temporal templates. Classification of query video is determined by nearest neighbour classifier. On the other hand, Experimental result for MuHAVi and Weizmann dataset is shown demonstrated using leave-one-out cross validation. The figure 98.5% accuracy is obtained from MuHAVi dataset which consist of 8 actions and on MuHAVi dataset with 14 actions having similar setup as previous one gives 91.9% accuracy. The figure of 100% accuracy is obtained on Weizmann dataset consists of 10 actions. The recognition rate in this approach decreases with uncompromising changes in camera view. S. Mukherjee, S. Biswas and D. Mukherjee [3] have presented work on recognition of interactions between human performers by ‘Dominating Pose Doublet’. Histogram of gradient (HOG) is used to detect people. The pose descriptors of the detected performers are obtained by using optical flow of sequence of video frames. Separate codebook is created for both performer that consisted of pose descriptors. Bipartite graph is generated for both performers by using dominating poses. There are nodes which represent poses from different codebook and edges consisted of weights depending on the frequency of occurrence of the poses in the videos. The least no. of poses required to cover all the variation of poses are dominating poses. UT interaction dataset is used in experiment and a figure of 86.67% accuracy is obtained from the experiment. This approach is limited to two persons only but can be extended for more than two people. M. Jimenez, E. Yeguas and N. Blanca [4] have proposed a system for exploring STIP-based models for the recognition of human interactions in TV videos. Spatio-temporal interest points (STIP) are selected from videos on the basis of Harris3D or by dense sampling of STIP. Descriptor HOG and HOF are calculated for STIP. BOW model is used for encoding Support vector machine (SVM) is used for the classification. In this approach TV Human Interaction, Dataset (TVHID), UT Interaction Dataset (UTID) and Hollywood-2 dataset are used. Performance of 0.3661 on TVHI dataset, 0.6077 on Hollywood 2 dataset and 0.86 for set1 and 0.88 for set2 of UTID is obtained. In [5], human interaction recognition based on the cooccurrence of visual words is proposed. 3D-XYT volume is extracted from two interacting persons. Two-dimensional co-occurrence matrix is constructed for two interacted persons. UT-interaction dataset is used in the experiment. Euclidean distance and k-nearest neighbour classifier are used as a distance function for set1 SVM classifier is used for set2. 40.63% and 66.67% accuracy is obtained for set 1 and set 2 respectively. This approach is limited to two persons with constant background.

Lo and A. Tsoi [6] have implemented work on motion boundary trajectory for human action recognition. Optical flow is used to find motion boundaries and to track the featured points. These Feature points are selected on the basis of Harris corner condition. To track these points precisely, they are sampled on 8 frames. The resultant trajectory is used as local descriptor based on displacement of trajectory. The histogram of gradient, histogram of optical flow and motion boundary histogram are computed around trajectories and used as features. To convert features into fixed dimensional vectors Bag of features technique is used. Support vector machine (SVM) is used for classification. The figure of 64.4% accuracy is measured on Hollywood2 dataset. E. Victor and J. Niebles [7] have introduced a work on spatio-temporal human-object interactions for action recognition in videos. Spatial and temporal evolution of relationship between human and object is described. Frames are obtained from videos are used to find human location, these frames are further used to extract features and to encode information about relative motion. For classification SVM technique is used. By analysing daily Activity dataset, the figure 96.3% accuracy is obtained on Gupta dataset and 98% is obtained on Rochester.

Nacim and C. Djeraba [8] have proposed work on real time crowd motion analysis. An Approach is presented to detect irregular activity of human body. Motion heat map is computed of the image. The motion heat map computes hot and cold region. High motion and low motion intensities is represented in hot region and cold region respectively. This approach can detect all collapsing situation videos. This analysis can be used in autonomous video surveillance system. N. Nguyen and A. Yoshitaka have proposed work on Human interaction recognition using independent subspace analysis algorithm. This framework [9] uses three-layer convolution Independent Subspace Analysis and PCA for human interaction recognition from segmented and unsegmented videos. UT-interaction database is used for the experiment. Two sets are created for segmented as well as for unsegmented videos for segmented videos the figure of 93% accuracy is obtained on set1 and 93% accuracy is obtained on set2. Similarly, for unsegmented videos the accuracy is 90% for set 1 and 85% for set 2. The experiment is carried out for two person or person-object interaction. The limitation of this approach is that Spatial and temporal localization of activities are not possible.

H. Wang, A. Kläser, C. Schmid and C. Liu have proposed work on Action recognition by dense trajectories. In the proposed approach [10] dense trajectories are used to carried human action recognition. Optical flow is used to track the feature points. A separate codebook is created using Bag of features. Trajectory, histogram of oriented gradient, histogram of optical flow and motion boundary histogram features are computed. Hollywood-2 database is used for the experiment. The experiment is carried out for Two person or person-object interaction. A figure of 58.3% accuracy is obtained from this approach. The limitation of this approach is accuracy which is only 58.3%.

M. Jiménez and N. Blanca have proposed work on Human interaction recognition by motion decoupling", Pattern Recognition and Image Analysis. The proposed method [11] takes input video and form this video it detects shots, then region of interest that consisted of upper body of both actors are detected. The optical flow between interest regions of successive frame is computed and pyramid of accumulated histogram of optical flow descriptor is computed. The video is represented as bags of descriptors and bag-support vector machine is used for the classification. TV Human Interaction database is used for the experiment. The experiment is carried out for Two person or person-object interaction. A figure of 0.463% accuracy is obtained from this approach. The limitation of this approach is accuracy is very low. K. Yun, J. Honorio, D. Chattopadhyay, T. Berg and D. Samaras have proposed work on Two-person interaction detection using body-pose features and multiple instance learning. In this paper [12] body pose features are used to detect human interaction. For this skeleton of person are created, Joint features of skeleton are extracted and MIL is used. RGBD Videos captured using Microsoft Kinect Sensors is used as a database for the experiment. The experiment is carried out for Two person or person-object interaction. A figure of 87.3% accuracy is obtained from this approach. The limitation of this approach is that the experiment is carried out for the videos from specific view point.

Features plays a key role to separate the background object to foreground object. Most of the Feature extraction technique has been done so far for initial stage of detection process. In [13], edge feature is used for feature descriptor which is more invariant to illumination than other descriptor such as color [14] etc. further other descriptor eg. texture feature [15], optical flow and biological features may be added to enhance the computational efficiency. In [16], color-based method is used for classification of objects, gives high accuracy and computational time. The local binary pattern (LBP) texture is applied for pattern analysis in grey-scale images [20]. In [21], a real time human motion detection and tracking security system was introduced. Video-based human detection methods can be classified into three categories namely, appearance-based, motion-based, and hybrid methods. first type of methods is based on the static object. [17] proposed a detector trained on large pedestrian databases which scan every frame and search the model patterns that can be used for non-static cameras. In [26], Histogram of oriented gradients provides the distribution of edge. Motion-based methods are used for mobile vehicles or objects which are in transient phase such as moving gestures. In [19] Blob motion statistics based pedestrian detection is applied for moving object which consist cyclic pattern in blob trajectory and relationship between changing position and blob size. Hybrid methods are the combination of previous methods.

In this paper, the proposed technique has successfully detected humans in variable lighting conditions with complex background. We made our own dataset where we tried to make different lighting conditions and choose the background colour similar to the skin colour. We constructed the dataset such that a car is moving along with a human and our algorithm detected the human successfully.

III. PROPOSED WORK

This paper presents a 2D human motion analysis system which is used to detect human bodies. If the input is a video, it designs to process and estimate the position of human in each frame separately, remove the unwanted objects or noise in the background, detect if human is present in the scene and post-process the estimated results to achieve a final detection percentage. The Video used in the experiment is a RGB Video. A piece-wise analysis of each frame of the video is done. Each frame which is a RGB Image is converted into a Grayscale Image. Frame Differencing is performed in this Grayscale Image. It only extracts the objects in motion. The ith frame is subtracted with i-nth frame to get the objects in motion. If there is any motion in these n frames, then it can be extracted by this method. Frame differencing results shows better accuracy than the original Background Subtraction Method. Generally, Regions of Interest in an image are objects (humans, cars, text etc.) in its foreground. After image pre-processing (which may include image de-noising etc.) object localisation is required which may make use of this technique. Frame Differencing is widely used for detecting moving objects in videos from static cameras. The motive of this approach is detecting the moving objects from the difference between the current frame and a reference frame, often called “background image”, or “background model”. Edges reflect strong intensity change therefore we tried to find out the edges by converting the grayscaled image to a binary image by some thresholding operation. Binary images may contain numerous imperfections. In particular, the binary regions produced by simple thresholding are distorted by noise and texture. Morphological image processing removes these imperfections by accounting for the form and structure of the image. Morphological image processing is a collection of non-linear operations related to the shape or morphology of features in an image. However, if scene is changed, the value of the filters used is to be changed to a certain extent so that we can get the de-noised image. After the Morphological operations are performed, the algorithm checks whether there is an object or not. If there are objects, then we calculate the width/length (w/l) ratio i.e. the Aspect Ratio of the object and if the w/l ratio comes in the range of 0.3 to 0.75 then it can conclude that the object is a Pedestrian. So, we can detect a pedestrian and a stationary human. If we change the aspect ratio to a value about 2 to 3.5 then we can detect vehicles. So, with a small modification we can detect the vehicles as well. However, in case of occlusion this method doesn’t work well. The system flowchart of the algorithm is shown below:

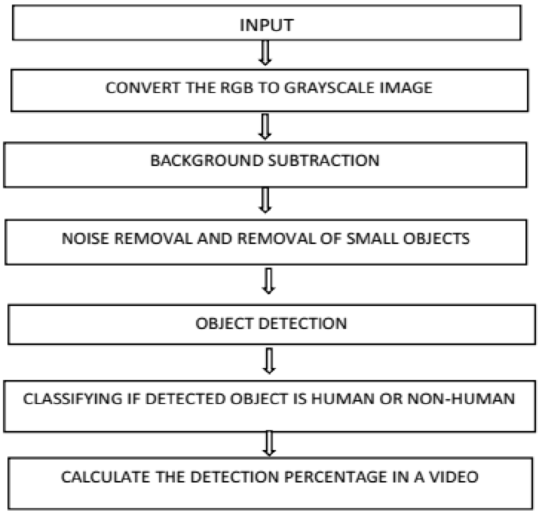
INPUT
CONVERT THE RGB TO GRAYSCALE IMAGE
BACKGROUND SUBTRACTION
NOISE REMOVAL AND REMOVAL OF SMALL OBJECTS
OBJECT DETECTION
CLASSIFYING IF DETECTED OBJECT IS HUMAN OR NON-HUMAN
CALCULATE THE DETECTION PERCENTAGE IN A VIDEO

Figure 1. Flow Chart of the Proposed Human Detection System.

Fig.1 shows the different steps of the algorithm of the proposed system. The process flow of the proposed system is shown in Fig. 2. It shows the various steps in the algorithm with the outputs of each step. It describes the whole process of the system and shows the way we are classifying humans from other objects.

The two images at the first step are the Frame No. 15 and Frame No. 30 of the video. Each of the frames are converted to Grayscale Images. Frame differencing operation of these grayscale frames are performed in the whole video with a difference of 15 frames, to extract the moving objects. Image Segmentation is used to separate all the moving objects. This is done in step 2. First, the unwanted noise is removed by morphological operations and then Connected Component Analysis is done to obtain the number of moving objects present in the frame and separate them. A set of pixels that form a connected group is known as a Connected Component in a Binary Image. This connected component detects connected regions in a binary image. They are mainly used for blob extraction on a binary image from a thresholding step. After finding the number of objects in the frame, we find the w/l ratio of each object. The w/l ratio is calculated in (1) :

$$Wl_{\text{ratio}} = (\text{ext}_{1_{\max}} - \text{ext}_{1_{\min}})/(\text{ext}_{2_{\max}} - \text{ext}_{2_{\min}}) \qquad (1)$$

where $\text{ext}_{1_{\max}}$ = right-most pixel of the object

$\text{ext}_{1_{\min}}$ = left-most pixel of the object

$\text{ext}_{2_{\max}}$ = bottom-most pixels of the object

$\text{ext}_{2_{\min}}$ = top-most pixels of the object

The w/l ratio of each object is stored in a matrix. The next step is to classify whether the object is a pedestrian or a non-pedestrian. The w/l ratio or the aspect ratio of a pedestrian varies from 0.3 to 0.75. So, in the matrix if any value of a w/l ratio is in the above range then that object is classified as a pedestrian.

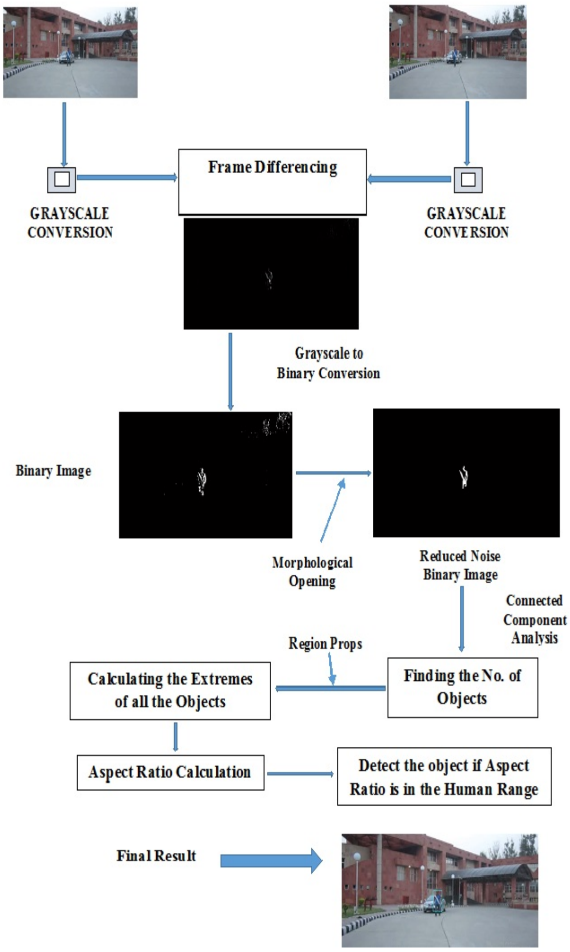
Frame Differencing
GRAYSCALE CONVERSION
GRAYSCALE CONVERSION
Grayscale to Binary Conversion
Binary Image
Morphological Opening
Reduced Noise Binary Image
Connected Component Analysis
Region Props
Calculating the Extremes of all the Objects
Finding the No. of Objects
Aspect Ratio Calculation
Detect the object if Aspect Ratio is in the Human Range
Final Result

Figure 2. Process Flow of the Human Detection System.

IV. EXPEERIMENTAL RESULTS

The Algorithm is performed in MATLAB 2016. The frames are in 640*360 resolutions and they are sampled at a rate of 15 samples per second. In order to verify and evaluate the performance of our human detection system, it is tested in different challenging environments. The experiment is performed in different videos. It is also performed in PET 2000, PET 2009 and Visor Dataset. We have created our own Dataset where a Human and a Car is present. At first, we have made a dataset where a human is moving in any random direction and the car is stationary while turning off the street lights. On second instant, human motion is in a fixed direction and keeping the car stationary while turning off the street lights. On third instant, the car and human both are in motion in random directions. Our algorithm is tested in this dataset and the algorithm successfully detected humans in all these situations. Our algorithm is tested on various dataset like in PET 2000, PET 2009 and VISOR Dataset. The algorithm worked well for multiple pedestrian detection as well and a comparative analysis of our algorithm in various dataset is given in table 1. The output of our algorithm in various dataset is given below:

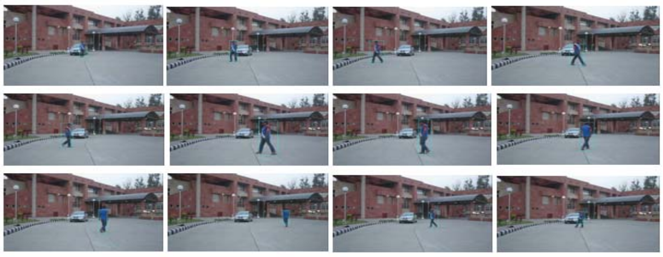

Figure 3. Output of Dataset A.

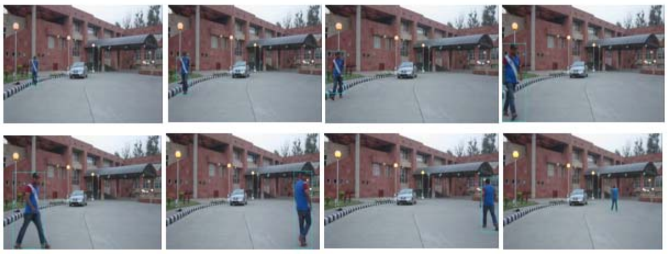

Figure 4. Output of Dataset B.

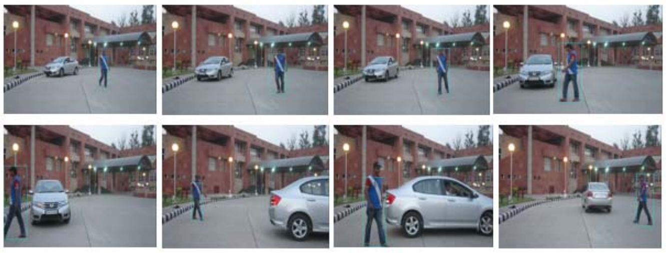

Figure 5. Output of Dataset C.

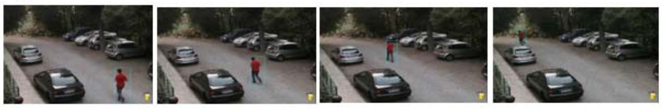

Figure 6.Output of Dataset D.

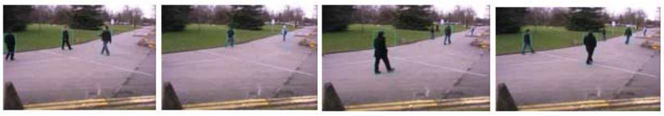

Figure 7.Output of Dataset E.

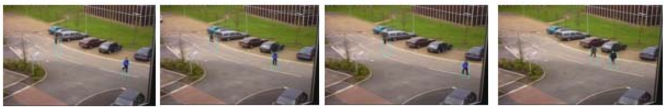

Figure 8.Output of Dataset F.

The output of our algorithm is shown in the Figure 3 to 8. It shows that across the pedestrian, a bounding blue box is generated to detect the humans. The output of Dataset E and Dataset F shows the output of the multiple pedestrian dataset. The table given below shows a comparative analysis of the output of the algorithm working with various datasets. The execution time and detection efficiency is taken as the parameters of comparison in various datasets.

TABLE 1. COMPARATIVE ANALYSIS OF OUR ALGORITHM WITH VARIOUS DATASETS

Datasets	Number Of Testing Samples	Resolution of the Dataset	Execution Time	Detection Rate
Dataset A	680	640*360	177 sec	87.06%
Dataset B	502	640*360	125 sec	85.66%
Dataset C	1073	640*360	342 sec	74.66%
Dataset D	78	640*480	15 sec	98.72%
Dataset E	437	640*480	152 sec	85.67%
Dataset F	776	600*480	257 sec	93.135%

V. CONCLUSIONS

In this paper, a robust human detection system based on video processing is proposed. This shape based method is divided in two stages. The first stage involves Image Pre-processing where the Motion region is extracted and Image Segmentation is applied to this motion region. The second stage classifies the segmented image as a human or a non-human based on Aspect Ratio of Human. So, we can say that the Motion region is incorporated with the Aspect Ratio feature to propose a Robust Human Detection Method. A Dataset is made where the background colour matches with the Human Skin Colour. In this situation, it is very difficult to track the human. We propose a system where we can track human under such conditions. The system is tested in PETs Database also and an overall Detection Rate of 85% is reported. However, the Detection rate gets reduced drastically when the human is occluded in the scene. However, in case of occlusion this method doesn't work well. This can be improved by adding other feature e.g. texture, Fourier descriptor, statistical moments, moments of 2D functions etc.

REFERENCES

[1] A. Sambath and A. Nirmala, “A survey on multimodal techniques in visual content-based video retrieval”, International Journal of Advanced Research in Computer Science and Software Engineering, vol. 5, 2015.

[2] E. Abdalrahman, S. Cheema, C. Thurau and C. Bauckhage, "Temporal key poses for human action recognition", International Conference on Computer Vision Workshops, IEEE, pp. 1310-1317, 2011.

[3] S. Mukherjee, S. Biswas and D. Mukherjee, "Recognizing interactions between human performers by ‘Dominating Pose Doublet’ ", Machine Vision and Applications, Springer Berlin Heidelberg, vol. 25, pp. 1033-1052, 2014.

[4] M. Jiménez, E. Yeguas and N. Blanca, "Exploring STIP-based models for recognizing human interactions in TV videos", Pattern Recognition Letters, Elsevier, vol. 34, pp. 1819-1828, 2013.

[5] K. Slimani, Y. Benezeth, and F. Souami, "Human interaction recognition based on the co-occurrence of visual words", Computer Vision and Pattern Recognition Workshops, IEEE, pp. 461-466, 2014.

[6] S. Lo and A. Tsoi, "Motion boundary trajectory for human action recognition", Computer Vision-ACCV 2014 Workshops, Springer International Publishing, pp. 85-98, 2014. S.

[7] E. Victor and J. Niebles, "Spatio-temporal human-object interactions for action recognition in videos", Computer VisionWorkshops (ICCVW), IEEE, pp. 508-514, 2013.

[8] I. Nacim and C. Djeraba, "Real-time crowd motion analysis", 19th International Conference Pattern Recognition, IEEE, pp. 1-4, 2008.

[9] N. Nguyen and A. Yoshitaka, "Human interaction recognition using independent subspace analysis algorithm", International Symposium on Multimedia (ISM), IEEE, pp. 40-46, 2014.

[10] H. Wang, A. Kläser, C. Schmid and C. Liu. "Action recognition by dense trajectories", Computer Vision and Pattern Recognition, IEEE, pp. 3169-3176, 2011.

[11] M. Jiménez and N. Blanca, "Human interaction recognition by motion decoupling", Pattern Recognition and Image Analysis, Springer Berlin Heidelberg, pp. 374-381, 2013.

[12] K. Yun, J. Honorio, D. Chattopadhyay, T. Berg and D. Samaras, "Two-person interaction detection using body-pose features and multiple instance learning", Computer Society Conference on Computer Vision and Pattern Recognition Workshops, IEEE, pp. 28-35, 2012.

[13] K. A. Joshi and D.G. Thakore, “A survey on moving object detection and tracking in video surveillance system,” International Journal of Soft Computing and Engineering (IJSCE), ISSN: 2231-2307, Vol. 2, Issue 3, July 2012.

[14] B. Pu, F. Zhou, X. Bai, “Particle filter based on color feature with contour information adaptively integrated for object tracking,” Fourth International Symposium on Computational Intelligence and Design, Vol 2, pp. 359-362, 2011.

[15] B. Deori and D. M. Thounaojam, “A survey on moving object tracking in video,” International Journal on Information Theory (IJIT), Vol. 3, No. 3, July 2014.

[16] H. S. Parekh, D. G. Thakore and U. K. Jaliya, “A survey on object detection and tracking methods,” International Journal of Innovative Research in Computer and Communication Engineering, Vol. 2, Issue 2, Feb 2014.

[17] M. Enzweiler and D. M. Gavrila, “Monocular pedestrian detecion: Survey and experiments,” IEEE Trans. Pattern Anal. Mach. Intell., vol. 31, no. 12, pp. 2179–2195, Dec. 2009.

[18] P. Felzenszwalb, R. Girshick, D. McAllester, and D. Ramanan, “Object detection with discriminatively trained part-based models,” IEEE Trans. Pattern Anal. Mach. Intell., vol. 32, no. 9, pp. 1627–1645, Sep. 2009.

[19] P. Borges, “Pedestrian detection based on blob motion statistics,” IEEE Trans. Circuits Syst. Video Technol., vol. 23, no. 2, pp. 224–235, Feb. 2013.

[20] B. Deori and D. M. Thounaojam, “A survey on moving object tracking in video,” International Journal on Information Theory (IJIT), Vol. 3, No. 3, July 2014.

[21] R. Shaalini, C. Shanmugam, R. N. Kumar, G. Myilsamy, and N. Manikandaprabu, “Human motion detection and tracking for real-time security system,” International Journal of Advanced Research in Computer Science and Software Engineering, Vol. 3, Issue 12, 2013.