J Inf Process Syst, Vol.15, No.1, pp.67~90, February 2019
https://doi.org/10.3745/JIPS.04.0100

ISSN 1976-913X (Print)
ISSN 2092-805X (Electronic)

JOURNAL OF INFORMATION PROCESSING SYSTEMS JIPS

Hierarchical Graph Based Segmentation and Consensus based Human Tracking Technique

Sunitha Madasi Ramachandra*, Haradagere Siddaramaiah Jayanna**, and Ramegowda***

Abstract

Accurate detection, tracking and analysis of human movement using robots and other visual surveillance systems is still a challenge. Efforts are on to make the system robust against constraints such as variation in shape, size, pose and occlusion. Traditional methods of detection used the sliding window approach which involved scanning of various sizes of windows across an image. This paper concentrates on employing a state-of-the-art, hierarchical graph based method for segmentation. It has two stages: part level segmentation for color-consistent segments and object level segmentation for category-consistent regions. The tracking phase is achieved by employing SIFT keypoint descriptor based technique in a combined matching and tracking scheme with validation phase. Localization of human region in each frame is performed by keypoints by casting votes for the center of the human detected region. As it is difficult to avoid incorrect keypoints, a consensus-based framework is used to detect voting behavior. The designed methodology is tested on the video sequences having 3 to 4 persons.

Keywords

Consensus Based Framework, Hierarchical Graph Based Segmentation, SIFT Keypoint Descriptor

1. Introduction

Computers are known to perform repetitive, computational and data-intensive task at a better scale and pace than human beings have become ubiquitous in our daily lives. Hence their capabilities can be used further to carry out many intellectual and higher level tasks like analysis of visual scenes, reasoning and logical analyses.

Image processing is a term which denotes the processing of an image, which is taken as an input and the result of processing may be a related parameter of an image. The major task of image processing is to observe objects and traits that are not visible to the naked eye. Moving object detection and tracking is a key task in the area of computer vision.

In recent years, human motion tracking and analysis are most popular aspects of research in image processing. For example, Dalal and Triggs [1] have described a variety of methods of human detection. They commented that the important part of human detection and motion analysis is human movement.

Manuscript received March 8, 2016; first revision January 13, 2017; second revision March 27, 2017; accepted May 16, 2017.
Corresponding Author: Sunitha Madasi Ramachandra (sunithamr2310@gmail.com)
* Dept. of Computer Science and Engineering, Adichunchanagiri Institute of Technology, Chikkamagaluru, Karnataka, India (sunithamr2310@gmail.com)
** Dept. of Information Science and Engineering, Siddaganga Institute of Technology, Tumakuru, Karnataka, India (jayannahs@gmail.com)
*** Bahubali College of Engineering, Shravanabelagola, Karnataka, India (gowdaait@gmail.com)

The main aim is to detect human movement from the background of the video sequences. The detection and analysis of human movements in a video surveillance scene has gained attraction by multiple applications like human gait characterization, abnormal event detection, person identification, etc. This is due to great popularity in the field of video surveillance, traffic monitoring, and has remarkable commercial significance.

However, there are a lot of limitations in human tracking like videos with low resolution, change in the illumination and the background changes to monitor events in surveillance displays. For an efficient classification of objects, intellectual methods that detect and capture movement information about moving objects are therefore required. The detected object is then monitored for further analysis.

This paper aims to present a robust method to track humans that involve two stages: (i) human detection using hierarchical graph-based segmentation and (ii) human tracking based on consensus-based matching and tracking technique. Hierarchical graph segmentation is used to divide an image into many disjoint subsets, where every subset has a significant part of an image [2]. Generated parts are grouped together to form human blobs. Each human blob undergoes novel human validation scheme. For each human blob, histogram of optical flow (HOF) features [3] are calculated and matched knowledge-based features to identify a human. Later, consensus-based matching tracking (CMT) approach, which uses a human representation based on a key point, is used. These estimations are also used in the identification of objects.

The remainder of the paper is organized so that the literature survey of recent work is discussed in Section 2. Section 3 proposes a methodology and Section 4 discusses different results obtained after the conduction of experiments. Finally, Section 5 presents conclusions.

2. Literature Review

2.1 Tracking

Santner et al. [4] provided a manner to deal with the visible tracking problems like drifting as they rely on self-updates of online finding out approaches. This work showed that the augmenting online studying strategy with complementary monitoring approaches can lead to extra steady outcomes. Additionally, they used an easy template mannequin as a non-adaptive mannequin and an optical-flow based mean shift tracker for adaptive detail. A web-based random forest is used as reasonably adaptive appearance headquartered learner. A combo of all the three trackers in a cascade is done. Since all these components execute on multicore techniques, the method works well for real-time performance. Experiments showed a precision rate of 0.9 for publicly available standard data set.

Andrilukaet al. [5] presented a novel process which combined the advantages of each detection and tracking in a single framework. In entire body for each human, an approximate articulation is detected utilizing nearby facets that mannequin the appearance of person physique components. Utilizing a hierarchical Gaussian approach latent variable model (HGPLVM), prior talents on viable articulations and temporal coherency within a running cycle is modeled in this work. Experimental results for detecting and monitoring multiple humans in cluttered scenes with reoccurring and occlusions on TUD dataset are established.

Andriyenko et al. [6] formulated multi-target tracking as a discrete continuous optimization which

handles each facet in its ordinary area and makes it possible for leveraging strong tactics for multi-model becoming. Information organization is carried out using discrete optimization with label charges, yielding close optimality. With an easy closed type resolution, which is utilized in turn, to update the label, trajectory estimation is posed as becoming a steady concerned. Accuracy and robustness of this method are validated on standard datasets and suffered with low precision rate.

A consensus based dispensed multi target monitoring algorithm was proposed to handle the problem of naivety in a digital camera network [7]. In this work they have jointly addressed the tracking estimation mistakes and data organization. The outcomes of naivety to the progress of an information weighted consensus technique, which is termed as the multi-target information consensus (MTIC). Additionally, it is stated that MTIC algorithm is made very powerful to false measurements clutter by the incorporation of the probabilistic technique for data association. But this method suffered with the problem of data association with an increase in targets.

Mixed discrete continuous conditional random field (CRF) system tracks multiple targets and handles inter-object exclusions at information association stage situated on non-sub modular constraints [8]. To derive suitable CRF potentials a statistical information evaluation was additionally used. Experiments had been carried out on PETS'09, TUD and ETH dataset benchmark videos with precision rate of 87.2% and when compared with modern-day approaches, outcome showed with a transparent growth from the simultaneous exclusion constraints.

In [9], a technique for quick monitoring of customary objects in a video sequence with the help of a detector was presented. This detector uses the Hough transform into pixel established descriptor. By means of a powerful model of segmentation, the system is equipped to monitor objects which bear flexibility deformations. It also takes cognizance of variations in appearance and shapes. This process gives better results than state-of-the-art methods for Babenko dataset video sequence and non-rigid dataset with less precision rate. The method was not able to change bounding box size and aspect ratio during tracking.

The N consensus algorithm was proposed to reduce consensus process rate to track objects of the target monitoring system with better efficiency [10]. To recognize all nodes from viewing nodes, inside twice the viewing range, the hop count is calculated based on viewing and communication stages. In contrast to normal consensus, the N consensus method no longer requires prior competencies of node connectivity. The reason is that, they use an elevated rapid covariance intersection algorithm during the consensus updating.

Babenko et al. [11] offered a multiple instances learning (MIL) algorithm that utilizes more than one circumstance to obtain a potential procedure with lesser parameter tweaks. This strategy is proposed to obtain superior results than state-of-the-art method. The drawback of the method is that it loses the track of the object that has left the scene. Experimental results show that the precision rate is about 90% of Sylvester video clippings.

2.2 Human Detection

Tian and Sclaroff [12] offered an algorithm for the recovery of the globally optimal 2D human figure detection with the usage of the loopy graph technique. This work has overcome the project of computation time due to the fact that the time complexity increases with the amount of the biggest clique in the graph. The awarded approach made use of BB (branch and bound) to recycle the DP (dynamic

programming) tables related to the tree model to search the tree situated lower bound instead of re-computing the lower bound from the beginning. The method fails while tracking multi-objects in a scene. The proposed work is evaluated on an Iterative parsing dataset with detection accuracy of 56.4% and shows that it runs speedy empirically.

Ta et al. [13] awarded new facets known as pairwise feature (PWF), for action awareness to encode both appearance and the spatio-temporal interest point (STIP) relations of the regional features. They proposed a combo of two codebooks for video representation. It fails to capture complex geometric structures among local feature. Experiments carried out on standard datasets KTH with a detection rate of 93.0% and 83% on Weizmann dataset.

Yang et al. [14] integrated a submission of two types of systems. It employs the brute force search approach to evaluate each and every space-time area within the video with the aid of a binary classifier on whether a detailed occasion occurs. The other method uses the capabilities of human detection and performs tracking to restrict the high priced brute drive search. It evaluates the candidate space-time cubes by means of combining 3D convolution neural networks (CNN) and support vector machine (SVM) classifiers founded on bag-of-words local features for the detection of event presence. To cut down the detection cost rates (DCR) by way of thorough cross-validation of the progress set, they have chosen suitable combining weights and thresholds. These systems also accomplished higher performance with a detection rate of 98% on a CAVIAR dataset with a single person like ObjectPut, CellToEar, and Pointing videos.

A method was presented in [15] which detects human by augmenting frequently used region, color and texture features. It leads to an enormously excessive dimensional characteristic area of larger than one hundred seventy thousand dimensions. Partial least squares (PLS) is an effective dimensionality reduction method and conserve tremendous discriminative knowledge for the projection of the info onto a much slash dimensional subspace. But works well only for low dimensional subspace. This process obtained 75% of recall rate on INRIA and ETHZ pedestrian dataset.

Peng and Zhang [16] proposed a structure to evaluate the first-class of specified segmentation quantitatively with a couple of ground truth segmentation. They have adaptively built a ground truth that is locally viable for segmentation. They also continue the structural consistency within the ground truths for a given segmentation. For quantitative segmentation analysis, this work effectively presents a structure to adaptively mix more than one ground truths. The analysis was completed on the benchmark Berkeley segmentation database.

Breitenstein et al. [17] addressed the dilemma in robotically monitoring individuals using an uncelebrated camera. A new technique for human monitoring via detecting in a particle filtering structure was proposed. As a graded statement model in addition to final excessive self-belief detection, proposed technique used the continuous self-assurance of pedestrian detectors and on-line expert, example specific classifiers. Detection and tracking of big quantity of dynamically moving persons in an elaborate environment without relying background modeling was done leading to impose less restriction. Demonstration of the algorithm presented compared to other current approaches is achieved on an ETHZ dataset with the precision rate of 70%. The result shows that robustness during partial occlusion is comparatively less.

A novel approach to conquer problems in tracking the formation of a swarm intelligence viewpoint was introduced in [18]. They presented a species-based particle swarm optimization process for a couple of objects monitoring via dividing global swarm into a couple of species matching the quantity of objects

within the scene. Every species search for its objects and continues to track it. The occlusion relationship between one of a kind objects is absolutely determined by the vigor of all species. In this method object is not closely wrap by the bounding box, hence it does not suit for a subspace based tracking method.

Existing algorithms considered in the literature are having many advantages and disadvantages. Few existing methods show that the track of the object is lost if the object disappears from the scene. In the proposed method, the human is tracked every time by making use of the Consensus-based tracking approach. These tracked objects are again validated using features stored in the database. If the tracked human has disappeared from the scene for a few frames and reappear again, human will be recognized and re-tracked again. Since the validation of the human is done in different levels, false positive rate is relatively low. In [9], it is seen that the bounding box does not vary with regard to the size of the object through the tracking and is hardly ever initialized to enclose the entire object giving the accuracy of about 87.41%. But in the proposed system once the keypoints are obtained using scale-invariant feature transform (SIFT), the optical flow of each of these keypoint is measured using Lucas-Kanade technique. The optical flow of the keypoints is compared with the predefined threshold, so that the keypoints which are not within this range are considered and centroids of these keypoints are calculated. By considering the centroid point and the corner of the object, the bounding box is drawn. Hence, the bounding box is varied every time in accordance with the size of the object.

We have implemented an efficient detection and tracking algorithm by making use of hierarchical graph based segmentation and consensus based human tracking algorithm and compared their detection and tracking performance. The combination of these algorithms along with the novel validation system produces the most hopeful results in terms of detection and tracking quality.

3. Methodology

Proposed system has two phase called training and testing phases as shown in Fig. 1. Each step is described in detail in the rest of the paper.

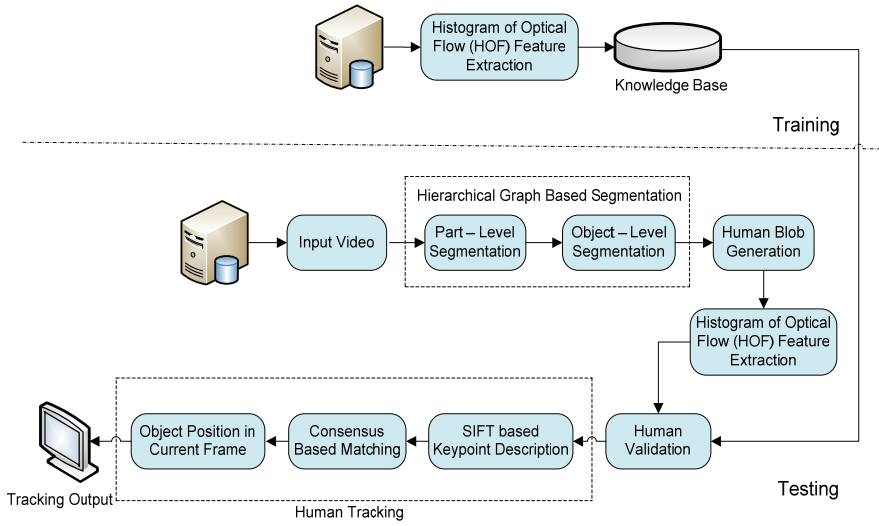


Fig. 1. A flow diagram of proposed human tracking system.

Training phase

In this phase, cropped human images are considered as input. All the images are resized in pre-processing stage. HOF features are extracted from these images and stored in a knowledge base. This knowledge base is further used for human validation.

Testing phase

The input for this phase is a video sequence from which frames are generated. Each frame is enhanced by applying a sharpening filter. The resultant enhanced image is passed to a human detection stage which is based on a hierarchical approach of graph-based segmentation. Part-level graph-based segmentation and object-level segmentation are two main steps of hierarchical graph-based segmentation algorithm as described in the below Section.

3.1 Human Detection

Humans in the input video sequence are detected by using hierarchical graph segmentation. First, color consistent clusters are obtained by using the segmentation of the part level on the pixels of the input frame using an efficient graph-based segmentation technique described in [19]. All color clusters are separated from the output of partial segmentation in object-level segmentation. Gabor features are calculated and transferred to SVM [20] for each cluster. The SVM will later classify clusters into human and non-human regions with these features.

3.1.1 Part level segmentation

Consider a graph $G = (V, E)$ having vertices $v_i \in V$, a set of pixels and $(v_i, v_j) \in E$ edges that match the neighboring vertices [21]. The goal of graph segmentation is partitioning the input frame into color clusters.

The input is shown in Figs. 2 and 3 gives the result obtained for input video after applying part level segmentation. First, a graph is constructed with vertices as input frame pixels and edges as the color differences among two adjacent vertices, v_i and v_j as $W_P\left((V_i,\ V_j)\right) = \|I(V_i) - I(V))_j\|$.

Fig. 2. Input first frame.

Fig. 3. Output of part-level segmentation.

Segmentation is initially carried out with each vertex v_i in its own component, say C_i, and the edges are arranged in the ascending weight of W_p. Region merging $D_p(C_1, C_2)$ among C_1 and C_2 is computed using graph segmentation. Merging is done if the $Dif(C_1, C_2)$ difference among the components is huge when compared to the inner difference with at least one of the components, $Int(C_1)$ and $Int(C_2)$. Later, two components are merged when

$$D_p(C_1, C_2) = \begin{cases} True & if \;\; Dif(C_1, C_2) > MInt(C_1, C_2) \\ False & otherwise \end{cases} \tag{1}$$

where

$$Dif(C_1, C_2) = \left| \min_{V_i \in C_1 \,, V_J \in C_2, (V_i, V_J) \in E} W_P\left((V_i, V_J)\right), \right. \tag{2}$$

and

$$MInt(C_1, C_2) = min(Int(C_1) + (C_1), Int(C_2) + (C_2)), \tag{3}$$

$$Int(C) = \max_{e \in MST(C,E)} W_P(e) \tag{4}$$

where the component C's minimum spanning tree is represented by the $MST(C, E)$ and $(C) = \frac{k}{|C|}$. Constant k refers to the value which control the merging and $|C|$ depicts to the number of components.

3.1.2 Object level segmentation

To achieve this segmentation $G' = (V', E')$, a graph is constructed having $V_i' \in V'$ number of vertices, which are the parts of segmentation obtained by part level approach of segmentation $(V_i', V_j') \in E'$. Edges are nothing but the weighted means of confidence values calculated from pairs of neighboring vertices as $W_o(V_i', V_j') = C_f(R(V_i')) + C_f\left(R(V_j')\right)$, where C_f represents the object confidence value obtained by classifiers and minimum rectangular region $R(V_i')$ that consists of the pixels in the segment V_i'. Weight parameters obtained by the sizes of the regions as $= |R(V_i')| / (|R(V_i')| + |R(V_j')|)$ and $|R(V_j')| / (|R(V_i')| + |R(V_j')|)$. The size of the rectangular region of V is represented as $|R(V)|$. After sorting the edge weights in a non-decreasing order, the region comparison predicts $D_o(S_1, S_2)$ between two components S_2 and S_1 is calculated by checking whether the merged segment has a higher confidence

than two segments as in Eq. (5). Later, two components are merged when $D_o(S_1, S_2)$ is false.

$$D_o(S_1, S_2) = \begin{cases} True & if \hat{C}_f(S_1 + S_2) < MCof(S_1, S_2) \\ false & otherwise \end{cases} \quad (5)$$

where

$$MCof(S_1, S_2) = max(C_f(P(S_1)) + C_f(R(S_2)), C_o) \quad (6)$$

$$\hat{C}_f(S) = maxC_f(R(S) +), \sim Uniform(-B, B) \quad (7)$$

where $R(S)$ indicates rectangular region contains all segmented pixels $V_i^{'} \in S.R(S)$ is denoted by four parameters, *x, y, wt* and *ht* of a rectangular box and is equivalent to $|R(S_1)|/(|R(S_1)| + |R(S_2)|)$ and $|R(S_2)|/(|R(S_2)| + |R(S_2)|)$. Where *x, y* indicates the initial point value of the rectangle, *wt* gives the width and *ht* specifies the height of the rectangle box. Gabor features [22,23] are extracted for these rectangular regions and given to SVM classifier [20] as input. By using minimum rectangular region confidence value, $C_f(R(S_1 + S_2))$ is computed. Many region hypotheses are generated from a merged region $R(S_1 + S_2)$ and location sizes of the regions are varied as in Eqs. (6) and (7) [21].The lower and upper bound is obtained and later uniformly the hypotheses are generated.

The comparisons of image segmented with gradient orientation pyramid (GOP) features which are stored in the knowledge base are done. An SVM classifier differentiates between human region and non-human region using trained GOP features to create a single blob. The whole procedure is demonstrated as in Fig. 4.

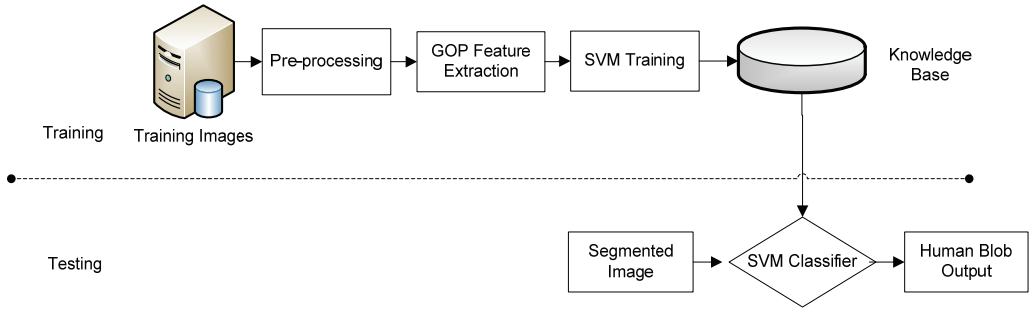


Fig. 4. A flow diagram of human blob generation.

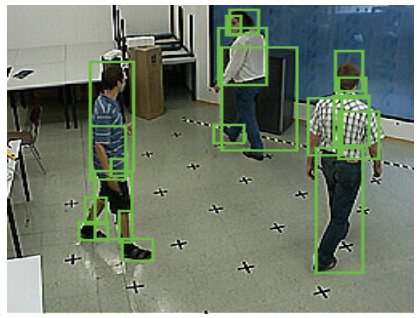


Fig. 5. Human patch generation.

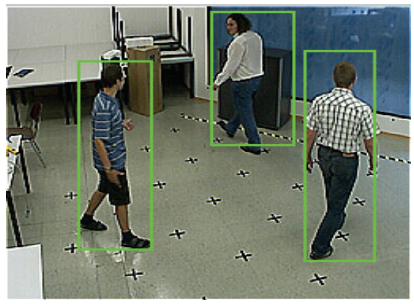


Fig. 6. Human detection by grouping human patches.

As shown in Fig. 5, the patches of human region marked with a rectangular box. All of these patches of same human regions form a human blob, as depicted in Fig. 6.

Human validation is done by comparing the HOF feature of the segmented human with feature stored in the knowledge base. Here the texture feature of the detected human was compared with the HOF features already stored in the knowledge base. During comparison, if the correlation of the features extracted is less when compared to the features already stored in the database then we match that person with the model. If the correlation difference obtained is more, then it is not considered as human. Texture features are chosen because they provide vital information for the classification of the segmented image. It is used in describing the high level semantics for the retrieval of human.

3.2 Human Tracking

3.2.1 Localization of keypoints using scale-invariant feature transformation

SIFT is a system for detecting and extracting local descriptors invariant for variation in scaling, illumination, rotation and little variations in viewpoints. Different phases of SIFT description are summarized below.

Scale space extrema detection

The initial part is a calculation that searches over entire image locations and scales. Efficient implementation is done by using different Gaussian functions to establish potential points of interest that can be invariant in scale and orientation. The point of interest for SIFT function corresponds to the extreme local difference of Gaussian filters on different scales.

Scale space is known by $L(x, y, \sigma)$ and described as below

$$L(x, y, \sigma) = G(x, y, \sigma) * I(x, y) \tag{8}$$

where $I(x, y)$ is input image.

$$G(x, y, \sigma) = \frac{1}{2\Pi\sigma^2} e^{\frac{x^2+y^2}{\sigma^2}} \tag{9}$$

Eq. (9) is a Gaussian variable scale whose result converts an image with a Gaussian filter difference as below

$$D(x, y, \sigma) = L(x, y, k\sigma) - L(x, y, \sigma) \tag{10}$$

This is unique from the Gaussian blurred images on both scales σ and $k\sigma$. The initial phase within the detection of interesting features is the convolution of the image with exclusive Gaussian filters. It produces a difference between Gaussian images and adjacent blurred images. The rotated images are grouped by octave and k value is chosen such that a set of blurred images per octave can be obtained. This ensures that the identical figure for variation of Gaussian images per octave is acquired. Once this difference of Gaussian is located, images are scanned for local extrema over space and scale.

Locating keypoints

Interesting points are identified as local minima or maxima of the difference of Gaussian (DOG) images across scales. Locating keypoints involves few steps as shown in Fig. 7. To find and describe the SIFT feature points, the steps given below must be followed:

Step 1: Input an image ranging from [0, 1].

Step 2: Use a Gaussian kernel variable scale $G(x, y, \sigma)$ to generate scale space $L(x, y, \sigma)$.

Step 3: Difference of Gaussian function is calculated as an approximation to the normalized Laplacian, because studies on this have shown that the normalized Laplacian does not change to the scale change.

Step 4: By comparing one of the pixels of its above, current and below scales in 3×3 areas maxima or minima of change of Gaussian performs value is found.

Step 5: By discarding points below a predetermined value, accurate keypoint's locations are obtained.

$$D(\hat{X}) = D + \frac{1}{2}\frac{\partial \mathrm{D}^T}{\partial \mathrm{X}}\hat{X} \tag{11}$$

where $\hat{X}$ is calculated by setting the derivative $D(x, y, \sigma)$ to zero.

Step 6: The extremas of DOG have huge principle curvatures along edges that can be condensed by comparing

$$\frac{Tr(H)^2}{Det(H)} < \frac{(r+1)^2}{r} \tag{12}$$

H indicates a 2×2 Hessian matrix, r is the ratio between the highest magnitude and the lowest one.

Step 7: To obtain invariance to rotation, $m(x, y)$ the gradient magnitude and $\theta(x, y)$ orientation are recomputed as the equations given below.

$$m(x, y) = \sqrt{(L(x+1, y) - L(x-1, y))^2 + (L(x, y+1) - L(x, y-1))^2} \tag{13}$$

$$\theta(x, y) = \tan^{-1}\frac{L(x, y+1) - L(x, y-1)}{L(x+1, y) - L(x-1, y)} \tag{14}$$

Step 8: This step is calculated taking a feature point and its 16×16 neighbors around it. Next dividing them into sub-regions of 4×4 and histogram of every sub-region with eight bins is calculated. After the keypoint orientation decision is made, the feature descriptor of orientation histograms on 4× 4 pixel neighborhoods is computed. The next step involves

assigning orientation to each and every key factor to for reaching invariance to image rotation. This step is then adopted by using keypoint Matching.

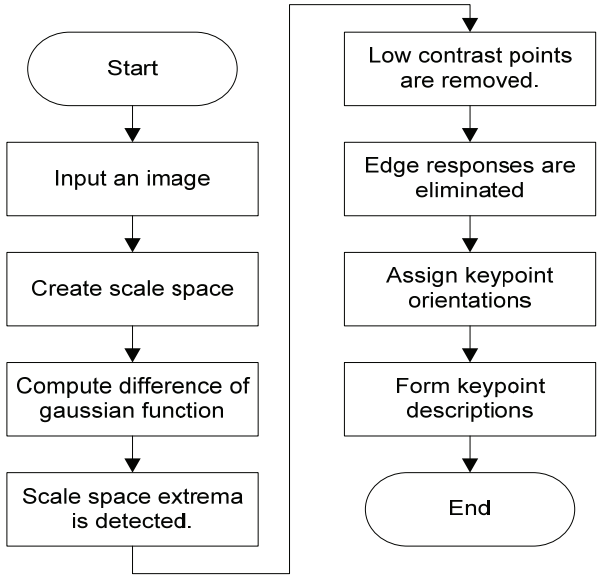


Fig. 7. Flow diagram of keypoints location approach.

Keypoint matching

In this step, keypoints among two images are in comparison through picking out their closest neighbors. The 2nd closest fit may also be near than the 1st in some circumstances. It is going to occur as a result of noise or various further causes. In such case, the ratio of distance closest to the 2nd closest distance is used. Distance is rejected if it is more than 0.8. Elimination of around 90% of false suits is completed even as discarding best 5% correct fits. So this concludes the summary of the SIFT descriptor for important points refer [24].

3.2.2 Consensus based matching and tracking

Consensus based matching and tracking method entails steps that are listed below. As in the Algorithm 1 it is seen that sequence of images $A_1, \ldots, A_n$ and an area $b_{11}, b_{12}, \ldots, b_{1n}$ is initialised in A_1. In every frame, improving pose of the interested object is carried out.

Algorithm 1

Output: $b_{21}, \ldots, b_{nn}$

1: $D_1, \ldots, D_n \leftarrow SIFT\ (A_1, b_{11}, b_{12}, \ldots, b_{1n})$

2: $K_1, \ldots, K_n \leftarrow D_1, \ldots, D_n$

3: **for** $i \leftarrow 2, \ldots, n$ **do**

4: $Df \leftarrow SIFT\ (A_i)$

5: $M_1, \ldots, M_n \leftarrow Match\ ((D_1, \ldots, D_n)\ , Df)$

6: $T_1, \ldots, T_n \leftarrow Tracking\ (M_1, \ldots, M_n, K_{i-1}, \ldots, K_{i-n}, A_{i-1}, A_i)$

7: $V_1, \ldots, V_n \leftarrow Vote\ ((T_1, \ldots, T_n),(D_1, \ldots, D_n))$

8: $V^c{}_1, \ldots, V^c{}_n \leftarrow Consensus\ (V_1, \ldots, V_n\ \ )$

9: $c_{i1}, \ \ldots, c_{in} \leftarrow Bounding\ box\ (V^c{}_1, \ldots, V^c{}_n, b_{11}, b_{12}, \ldots, b_{1n})$

10: $b_{i1}, \ \ldots, b_{in} \leftarrow identify(c_{i1}, \ \ldots, c_{in})$

end for

end

As given in the Algorithm 1 SIFT feature is applied to 1st frame i.e. A_1 and to its entire selected regions $b_{11}, b_{12}, \ldots, b_{1n}$, resulting to get descriptors $D_1, \ldots, D_n$ and keypoints $K_1, \ldots, K_n$. The object model is based on a set of keypoints, each of which represents an initialization of the location. SIFT feature is again applied to the entire second frame to get descriptor Df. Matching takes the first frame descriptor output and Df to get to matched descriptors $M_1, \ldots, M_n$.

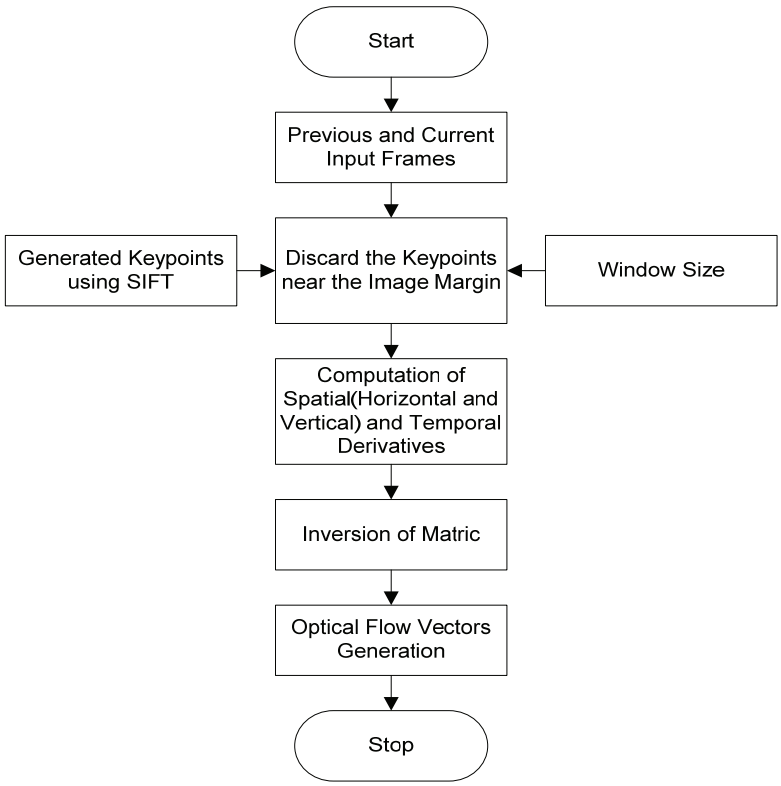


Fig. 8. Flow chart for Lucas-Kanade method.

Tracking of this matched descriptors are done by considering first frame, second frame along with their descriptors to get output $T_1, \ldots, T_n$. By employing the pyramidal variant of the Lucas-Kanade method optical flow is estimated as in Fig. 8. Matching and tracking of keypoints are done as in [25].

To locate the interested objects, each keypoint casts a single vote for the center of object resulting in set

of votes $V_1, \ldots, V_n$. So voting approach presented in [25] is followed along with a Consensus approach to get $V^c{}_1, \ldots, V^c{}_n$. Fig. 9 gives an example of the object tracking result.

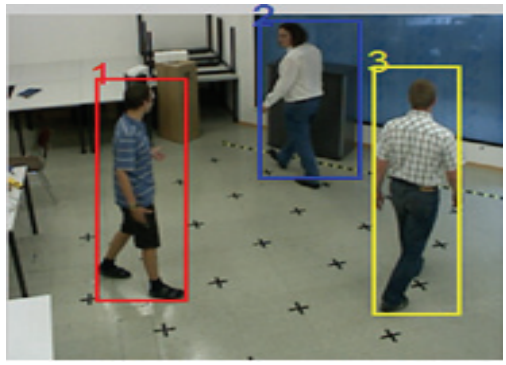


Fig. 9. Human tracking result.

The combination of these algorithms along with the proposed novel validation method produces the better results in detection and tracking quality. Every tracked object is validated using features stored in the database so that if the tracked human is disappeared from the scene for a few frames and reappears. These reappeared humans are tracked again with same id.

4. Experimental Results

The performance of the presented methodology is analyzed by applying the proposed method on the videos generated from the frames available at Learning, Recognition & Surveillance Group dataset (https://www.tugraz.at/institute/icg/research/team-bischof/lrs/). The video sequences contain 3–5 persons and was generated in their laboratory. The videos differ in viewpoint (angle), i.e., each scene is viewed using three static cameras with portion of views being overlapped. Using these videos, 18 different video sequences are generated.

We have also tested our method on CAVIAR dataset which is publicly available [26]. It contains people moving randomly, entering and exiting shops in public places and shopping centers. Enough number of tests cases having different viewpoints are presented in following section.

4.1 Human Detection

The proposed algorithm is applied to the color image of size 256×256. The color-consistent clusters obtained after the part level segmentation on the first frame are shown in Fig. 10(a) and (b). After the separation of the color cluster using object-level segmentation of this output, the classification of the human region uses SVM as in Fig. 10(c). Fig. 10(d) gives grouping of these clusters to outline a human blob.

Fig. 11 depicts the result of part level segmentation and human detection result of frame_1 of video_6 considering three people in a video. Fig. 12 depicts the result of part level segmentation and human detection result of frame_1 of video_7.

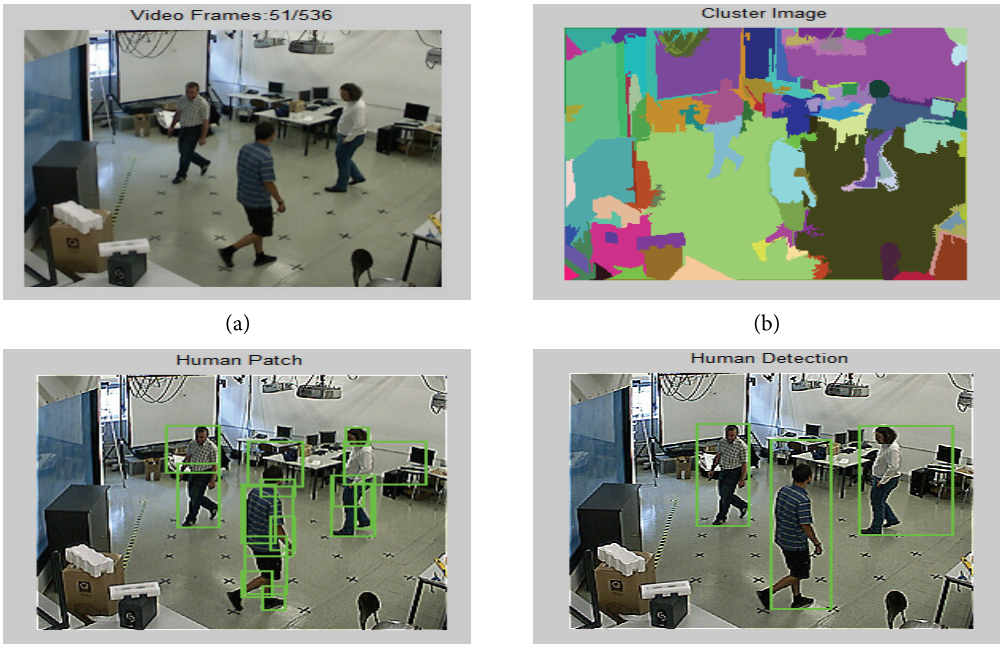


Fig. 10. For frame_1 of video_1. (a) Input frame, (b) part-level segmentation, (c) human patch generation, and (d) human detected by combining all human region clusters.

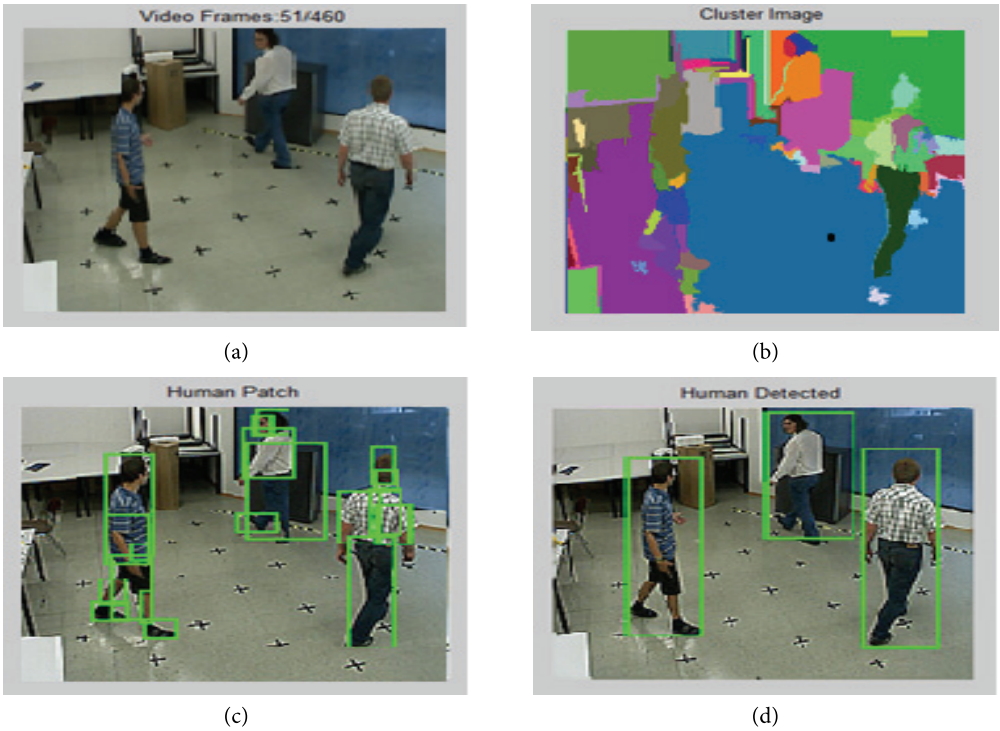


Fig. 11. For frame_1 of video_6. (a) Input frame, (b) part-level segmentation, (c) human patch generation, and (d) human detected by grouping all human region clusters.

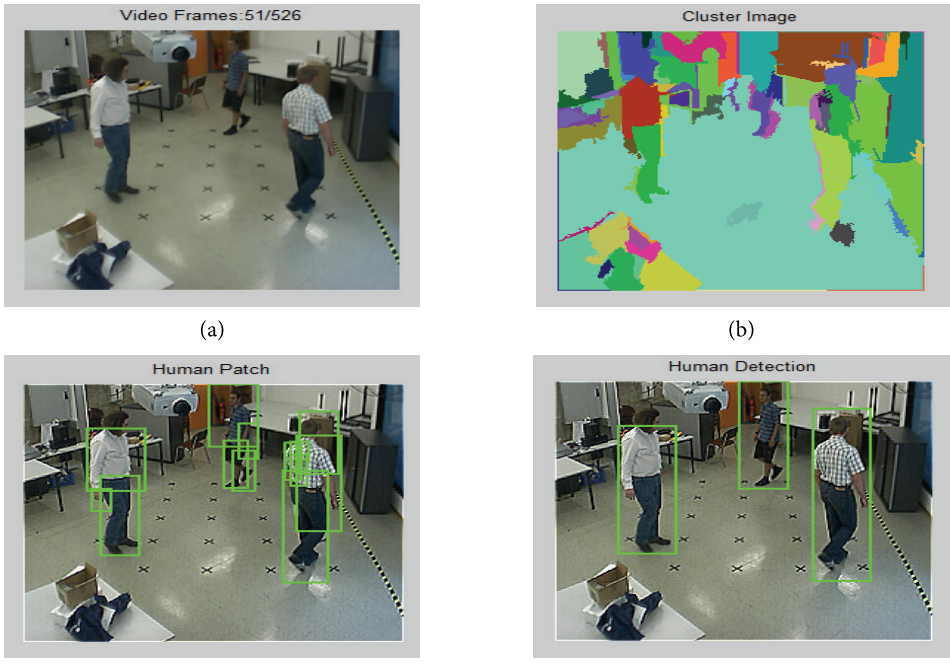


Fig. 12. For frame_1 of video_7. (a) Input frame, (b) part-level segmentation, (c) human patch generation, and (d) human detected by grouping all human region clusters.

Fig. 13. (a) Input frames 51, 173, and 429 of video_1, (b) human validation results of video_1, and (c) human tracking result of video_1.

4.2 Human Tracking

The result obtained from the proposed method of few video sequence of three different videos recorded from three different angles is presented here. Fig. 13(a), (b), and (c) depict the detection and tracking results of video_1 for frames 51, 173, and 429, respectively. In each frame, humans are detected successfully with different colored bounding box for all three people.

Fig. 14(a), (b), and (c) depict the result of frames 51, 147, and 353 of video_6, respectively. This video contains three people moving randomly by occluding one another many times. A proposed method successfully detected humans in each frame. Fig. 15(a), (b), and (c) depict the result of frame 51, 179, and 285 of video_7, respectively, which is the same sequence as video_6 in different camera location.

We have also demonstrated our results for the LeftBag and EnterExitCrossingPaths videos chosen from CAVIAR database [26]. These video clips consist of humans walking around, gathering, and picking and leaving bags, entering, exiting shopping malls, and so on. Fig. 16(a), (b), and (c) depict another result of frames 150, 250, and 870 of video_19 (LeftBag), respectively.

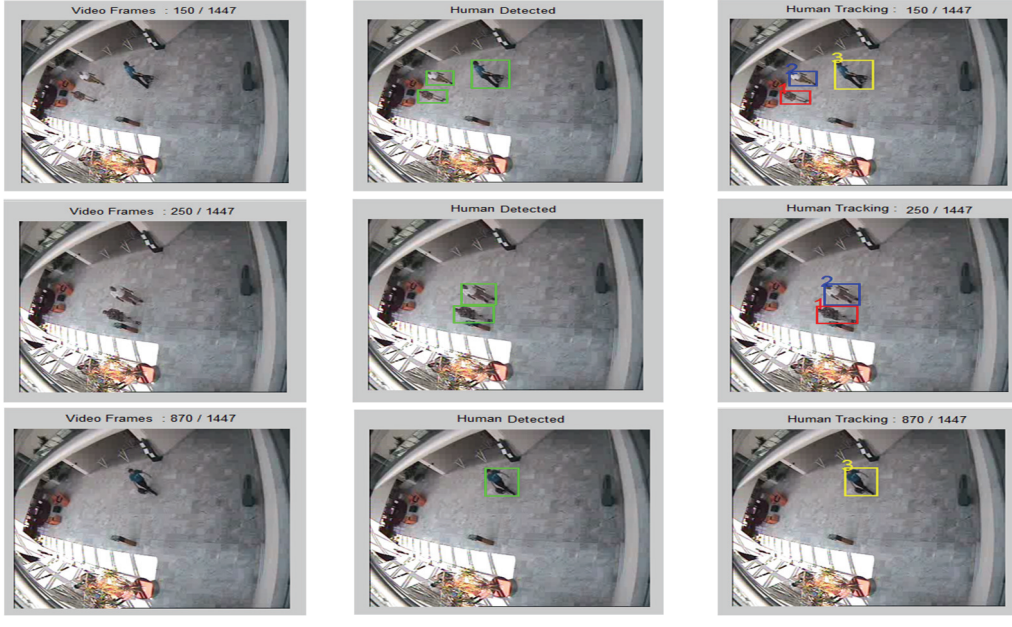
Each person in tracking phase is assigned with a unique id to represent the person. As shown in Fig. 16(c), detected persons are tracked with unique id (identification number) which is displayed above the bounding box. Each person is assigned with unique color bounding box and id.

(a) (b) (c)

Fig. 14. (a) Input frames 51, 147, and 353 of video_6, (b) human validation results of video_6, and (c) human tracking result of video_6.

(a) (b) (c)

Fig. 15. (a) Input frames 51, 179, and 285 of video_7, (b) human validation results of video_7, and (c) human tracking result of video_7.



(a) (b) (c)

Fig. 16. (a) Input frames 150, 250, and 870 of video_19, (b) human validation results of video_19, and (c) human tracking result of video_19.

(a) (b) (c)

Fig. 17. (a) Input frames 55, 73, and 90 of video_20, (b) human validation results of video_20,(c) human tracking result of video_20.

In Fig. 16, a person with green shirt detected with id3 has left the scene from 250th frame. After a few frames, he reenters to the view and the method proposed can track the person with the same id3. This is because the proposed method uses human validation phase, which stores features of each person along with id into the database as soon as the person is detected. Once the person re-enters the scene, features are extracted and validated. If the features are matched with already existing features in knowledge base, same id is assigned. If the feature does not match, then it is treated and assigned with new id.

Fig. 17(a), (b), and (c) depict the result of frames 55, 73, and 90 of video_20 (EnterExitCrossingPaths), respectively. This video contains two humans walking slowly towards the camera and these humans are occluded by humans crossing them. Our method detects and tracks all the humans successfully as shown in Fig. 17 even in occlusion condition.

The algorithm is again evaluated on one more indoor video which is shown in Fig. 18. In this video, three people are moving around the area by occluding one another many times. The proposed method can track the person even in occlusion condition as presented in Fig. 18(a), (b), and (c) for frames 220, 275, 380, and 390 of video_21, respectively.

The computed accuracy for tracking and detection for different videos are shown in the Table 1 using the formula given below:

$$accuracy = \frac{TP + TN}{TP + TN + FP + FN} * 100 \tag{15}$$

where TP indicated true positives, TN indicated true negative, FP indicated false positive, and FN indicates false negative.

(a) (b) (c)

Fig. 18. (a) Input frames 220, 275, 380, and 390 of video_21, (b) human validation results of video_21, and (c) human tracking result of video_21.

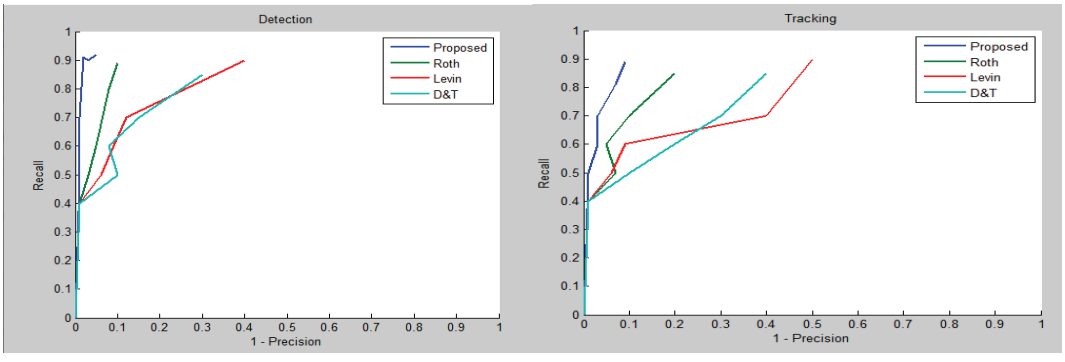
Table 1. Detection and tracking accuracy for video_1, 6, 7, 19, 20, and 21

Video	Accuracy (%)	
	Detection	Tracking
Video_1	86.33271	91.49125
Video_6	95.27831	94.01868
Video_7	88.84615	94.73541
Video_19	89.39876	92.65957
Video_20	87.20430	92.17191
Video_21	89.65000	91.82770
Average	89.45171	92.81742

Fig. 19 presents the comparison graph for different existing methods and proposed methods for detection and tracking rate as in the Table 2. Graph signifies detection and tracking rate obtained for proposed method is more compared to other methods in [27]. The work in [27] also has considered the similar type of indoor videos as used in the proposed methodology, making it more suitable for the comparison of overall accuracy.

Table 2. Recall rate and precision values for human detection and tracking of proposed method for video_1, 6, 7, 19, 20, and 21

Video	Detection		Tracking	
	Recall	Precision	Recall	Precision
Video_1	95.26552	96.81404	94.01427	93.66999
Video_6	88.84615	88.84615	94.73790	94.45710
Video_7	89.40336	92.66953	92.66285	92.12272
Video_19	87.22581	97.12644	92.18508	91.56051
Video_20	89.66667	92.83865	91.83439	91.16466
Video_21	89.45670	92.43792	92.82380	92.28931
Average	86.33271	86.33271	91.50831	90.76087



(a) (b)

Fig. 19. Comparison graph of proposed method with state-of-the-art method: (a) detection rate and (b) tracking rate.

Fig. 19 depicts the comparison graph for detection and tracking of different state-of-the-art tracking methods and our method. Graph shows that our method gives better detection and tracking rate than state-of-the-art method. Table 3 shows that tracking recall and precision average rate is 92.82% and 92.29% and detection average recall is 89.45% and precision rate is 92.44%. Detection and tracking rate of proposed method is better when compared to existing methods as in Table 3.

Table 3. Comparison of precision and recall rate of HOG, SIFT, and combined feature method

Video	Human validation					
	With HOG		With SIFT		With SIFT & HOG	
	Recall	Precision	Recall	Precision	Recall	Precision
Sample 001 (video_1)	81.32344	81.32344	84.35374	84.35374	86.33271	86.33271
LeftBag (video_9)	84.40452	89.0401	87.39922	91.22385	89.40336	92.66953
EnterExitCrossing (video_20)	82.32258	95.7958	85.29032	96.63743	87.22581	97.12644

The recall and precision values are calculated using the formula given below:

$$Recall\ Rate = \frac{TP}{TP + FN} \tag{16}$$

$$Precision\ Rate = \frac{TP}{TP + FP} \tag{17}$$

The proposed method uses HOG and SIFT features to track humans in the scene. Table 3 shows that when single feature is used to track an object, it gives lesser precision and recall rate than proposed combined feature tracking method. When more features are added to track objects, even if one feature fails in certain situation, other feature may still work well to track the object. Figs. 20 and 21 present the precision and recall comparison graph of proposed method with single feature method. Comparison graph shows that combined feature gives better precision and recall rate than single feature methods.

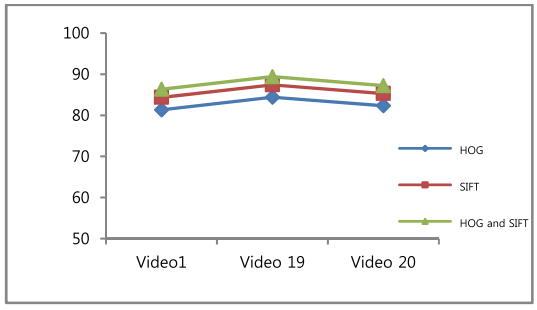


Fig. 20. Precision graph comparison of HOG, SIFT, and combined HOG and SIFT.

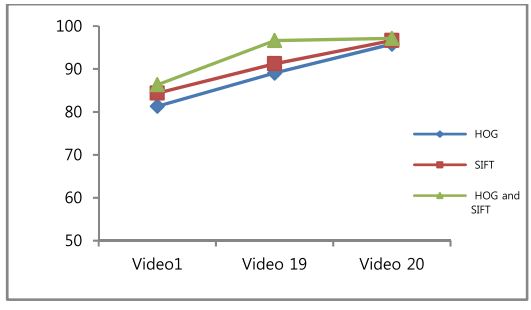


Fig. 21. Recall graph comparison of HOG, SIFT, and combined HOG and SIFT.

5. Conclusion

We have proposed a method for multi object detection and tracking using combined features. The major challenges for proposed tracking technique are unreliable measurements in case of tracking the missed detection and false positives. The contribution of our approach is to discover how this unreliable source of information can be used for tracking multiple human efficiently. Our proposed object tracking method for video images uses hierarchical graph based segmentation and consensus based human

tracking. Our approach uses hierarchical part level approach of segmentation, which aims at dividing an image into color consistent parts and object level segmentation to determine the objects with the help of previously trained SVM Classifier values. Proposed method is an effective approach involving SIFT key point descriptor method and consensus based object tracking technique with validation phase. This work gives good accuracy by detecting and tracking objects even in different conditions, sizes and shapes. Simulation results for frame sequences verify the suitability of the method for robust object tracking. Evolutions on datasets with different camera angle showed that the algorithm gives a good performance on huge number of application scenario compared to other existing methods.

References

[1] N. Dalal and B. Triggs, "Histograms of oriented gradients for human detection," in *Proceedings of IEEE Conference on Computer Vision and Pattern Recognition*, San Diego, CA, 2005, pp. 886-893.

[2] B. Peng, L. Zhang, and D. Zhang, "A survey of graph theoretical approaches to image segmentation," *Pattern Recognition*, vol. 46, no. 3, pp. 1020-1038, 2013.

[3] R. Chaudhry, A. Ravichandran, G. Hager, and R. Vidal, "Histograms of oriented optical flow and binet-cauchy kernels on nonlinear dynamical systems for the recognition of human actions," in *Proceedings of IEEE Conference on Computer Vision and Pattern Recognition*, Miami, FL, 2009, pp. 1932-1939.

[4] J. Santner, C. Leistner, A. Saffari, T. Pock, and H. Bischof, "PROST: parallel robust online simple tracking," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, San Francisco, CA, 2010, pp. 723-730.

[5] M. Andriluka, S. Roth, and B. Schiele, "People-tracking-by-detection and people-detection-by-tracking," in *Proceedings of IEEE Conference on Computer Vision and Pattern Recognition*, Anchorage, AK, 2008, pp. 1-8.

[6] A. Andriyenko, K. Schindler, and S. Roth, "Discrete-continuous optimization for multi-target tracking," in *Proceedings of 2012 IEEE Conference on Computer Vision and Pattern Recognition*, Providence, RI, 2012, pp. 1926-1933.

[7] A. T., Kamal, J. A. Farrell, and A. K. Roy-Chowdhury, "Information consensus for distributed multi-target tracking," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, Portland, OR, 2013, pp. 2403-2410.

[8] A. Milan, K. Schindler, and S. Roth, "Detection-and trajectory-level exclusion in multiple object tracking," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, Portland, OR, 2013, pp. 3682-3689.

[9] S. Duffner and C. Garcia, "PixelTrack: a fast adaptive algorithm for tracking non-rigid objects," in *Proceedings of the IEEE International Conference on Computer Vision*, Sydney, Australia, 2013, pp. 2480-2487.

[10] S. Katragadda and A. Cavallaro, "Neighbour consensus for distributed visual tracking," in *Proceedings of 2015 IEEE 10th International Conference on Intelligent Sensors, Sensor Networks and Information Processing (ISSNIP)*, Singapore, 2015, pp. 1-6.

[11] B. Babenko, M. H. Yang, and S. Belongie, "Robust object tracking with online multiple instance learning," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 33, no. 8, pp. 1619-1632, 2011.

[12] T. P. Tian and S. Sclaroff, "Fast globally optimal 2D human detection with loopy graph models," in *Proceedings of 2010 IEEE Conference on Computer Vision and Pattern Recognition*, San Francisco, CA, 2010, pp. 81-88.

[13] A. P. Ta, C. Wolf, G. Lavoue, A. Baskurt, and J. M. Jolion, "Pairwise features for human action recognition," in *Proceedings of 2010 International Conference on Pattern Recognition*, Istanbul, Turkey, 2010, pp. 3224-3227.

[14] M. Yang, S. Ji, W. Xu, J. Wang, F. Lv, K. Yu, Y. Gong, M. Dikmen, D. J. Lin, and T. S. Huang, "Detecting

human actions in surveillance videos," in *TREC Video Retrieval Evaluation (TRECVID) Workshop*, Gaithersburg, MD, 2009.

[15] W. R. Schwartz, A. Kembhavi, D. Harwood, and L. S. Davis, "Human detection using partial least squares analysis," in *Proceedings of 2009 IEEE 12th International Conference on Computer Vision*, Kyoto, Japan, 2009, pp. 24-31.

[16] B. Peng and L. Zhang, "Evaluation of image segmentation quality by adaptive ground truth composition," in *Computer Vision – ECCV 2012*. Heidelberg: Springer, 2012, pp. 287-300.

[17] M. D. Breitenstein, F. Reichlin, B. Leibe, E. Koller-Meier, and L. Van Gool, "Online multiperson tracking-by-detection from a single, uncalibrated camera," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 33, no. 9, pp. 1820-1833, 2011.

[18] X. Zhang, W. Hu, W. Qu, and S. Maybank, "Multiple object tracking via species-based particle swarm optimization," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 20, no. 11, pp. 1590-1602, 2010.

[19] P. F. Felzenszwalb and D. P. Huttenlocher, "Efficient graph-based image segmentation," *International Journal of Computer Vision*, vol. 59, no. 2, pp. 167-181, 2004.

[20] C. W. Hsu, C. C. Chang, and C. J. Lin, "A practical guide to support vector classification," Department of Computer Science, National Taiwan University, Taipei, Taiwan, 2003.

[21] J. Kim, B. Choi, and I. S. Kweon, "Object detection using hierarchical graph-based segmentation," in *Proceedings of 2013 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP),* Vancouver, Canada, 2013, pp. 1923-1926.

[22] A. K. Jain, N. K. Ratha, and S. Lakshmanan, "Object detection using Gabor filters," *Pattern Recognition*, vol. 30, no. 2, pp. 295-309, 1997.

[23] Z. Sun, G. Bebis, and R. Miller, "On-road vehicle detection using Gabor filters and support vector machines," in *Proceedings of 2002 14th International Conference on Digital Signal Processing*, Santorini, Greece, 2002, pp. 1019-1022.

[24] D. G. Lowe, "Distinctive image features from scale-invariant keypoints," *International Journal of Computer Vision*, vol. 60, no. 2, pp. 91-110, 2004.

[25] G. Nebehay and R. Pflugfelder, "Consensus-based matching and tracking of keypoints for object tracking," in *Proceedings of 2014 IEEE Winter Conference on Applications of Computer Vision*, Steamboat Springs, CO, 2014, pp. 862-869.

[26] CAVIAR dataset [Online]. Available: http://homepages.inf.ed.ac.uk/rbf/CAVIARDATA1/.

[27] P. M. Roth, C. Leistner, A. Berger, and H. Bischof, "Multiple instance learning from multiple cameras," in *Proceedings of 2010 IEEE International Conference on Computer Vision and Pattern Recognition Workshops (CVPRW),* San Francisco, CA, 2010, pp. 17-24.

Sunitha Madasi Ramachandra https://orcid.org/0000-0001-5678-552X

She received B.E. in Computer Science and Engineering from Kalpataru Institute of Technology, Bangalore University in 1997 and M.Tech. degree in Computer Science and and her Ph.D. in the Department of Computer Science and Engineering, Siddaganga Institute of Technology, Visveswaraya Technological University. She has published many research papers in national and international conferences and journals. Currently she is working as Professor in the department of Computer Science and Engineering, Adichunchanagiri Institute of Technology, Chikkamagaluru, Karnataka, India. Her research interests include Image processing, algorithms and cryptography.

Haradagere Siddaramaiah Jayanna https://orcid.org/0000-0002-4342-9339

He received the B.E. and M.E. degrees from Bangalore University in 1992 and 1995, respectively, and Ph.D. degree from prestigious Indian Institute of Technology, Guwahati, India, in 2009. He has published a number of papers in various national and international journals and conferences apart from guiding a number of UG, PG and research scholars. Currently, he is working as a Professor in the Department of Information Science and Engineering, Siddaganga Institute of Technology, Tumkur, Karnataka, India. His research interests are in the areas of speech, limited data speaker recognition, image processing, computer networks and computer architecture.

Ramegowda https://orcid.org/0000-0002-9417-2465

He obtained multidisciplinary degrees such as B.E. from Mysore University (1983), M.E. (1987) and LL.B. from Bangalore University, MCA from IGNOU, New Delhi, and obtained Ph.D. in the field of image processing and analysis applied to geotechnical engineering from Visvesvaraya Technology University, Belagavi, Karnataka, India, in 2010. He has published a number of papers in various national and international journals and conferences apart from guiding a number of UG, PG and research scholars. Currently, he is working as Principal, Bahubali College of Engineering, Shravanabelagola (Jainakashi), Karnataka, India. His research interests are in the areas of soft-computing, image processing and analysis, GIS and remote sensing.