Identity Retention of Multiple Objects under Extreme Occlusion Scenarios using Feature Descriptors

Dippal Israni, and Hiren Mewada

Original scientific paper

Abstract— **Identity assignment and retention needs multiple object detection and tracking. It plays a vital role in behavior analysis and gait recognition. The objective of Multiple Object Tracking (MOT) is to detect, track and retain identities from an image sequence. An occlusion is a major resistance in identity retention. It is a challenging task to handle occlusion while tracking varying number of person in the complex scene using a monocular camera. In MOT, occlusion remains a challenging task in real world applications. This paper uses Gaussian Mixture Model (GMM) and Hungarian Assignment (HA) for person detection and tracking. We propose an identity retention algorithm using Rotation Scale and Translation (RST) invariant feature descriptors. In addition, a segmentation based optimum demerge handling algorithm is proposed to retain proper identities under occlusion. The proposed approach is evaluated on a standard surveillance dataset sequences and it achieves 97 % object detection accuracy and 85% tracking accuracy for PETS-S2.L1 sequence and 69.7% accuracy as well as 72.3% precision for Town Centre Sequence.**

Index Terms— **Occlusion and demerging; Detection and Tracking; Person tracking; Surveillance video; Identity Retention.**

I. INTRODUCTION

Multiple object tracking is the primary step in video sequence analysis. As an example, player tracking in sports video [1]-[3], person tracking in surveillance video [4], [5], group of animals (ants [6], birds [7], fish [8], [9], bats [10]), etc. Other application includes action recognition [11], pose estimation [12], behavior analysis [13], human-computer interaction [14], visual surveillance [15] and virtual reality [16]. Despite of a large number of algorithms been developed for video analytics, major constraints to successful MOT are illumination changes, initi-alization and termination of tracklets, occlusion, distortion and/or similar appearance, interaction amongst multiple objects, changes in the size of the object.

This paper primarily focuses on pedestrian tracking. Three main reasons for this selection are: (i) pedestrians are the most non-rigid objects as they suffer deformations (changing its appearance along the time. (ii) a lot of practical applications of pedestrians results in a great commercial potential. (iii) 70% of research in the field of object tracking is focused on pedestrians [17].

In computer vision, occlusion defines the effect of one person blocking another person in the 3D Space from captured view. L. Yong et al. categorizes occlusion as partial occlusion, full occlusion, non-occlusion and long term occlusion. Object tracking is primarily grouped into detection based tracking and detection free tracking. In former tracking method, the algorithm detects object and links into trajectories [19], [20]. In later tracking method manual initialization is done in the first frame of some fixed objects, and then localizes these objects in subsequent frames [21]-[23]. Tracking can also be classified as an online tracking or offline tracking. Online tracking also known as sequential tracking handles the image sequence in a stepwise manner [22]-[25]. However, in offline tracking observation from all the frames are utilized and analyzed to estimate trajectories [5], [26]-[29].

This paper emphasizes on the usage of detection based tracking referred to as 'Tracking-by-Detection'. Further, this paper uses an online tracking method to extend the existing trajectories with current observations. This paper contributes a novel and efficient identity retention method. It proposes a multiple object demerge handling method that segments multiple object precisely. The proposed method is evaluated on standard datasets and it outperforms state of the art trackers.

The rest of this work is organized as follows. A review on non-rigid MOT for pedestrian video is described in section 2. The proposed method is described in section 3. Section 4 showcases the performance of the proposed approach on standard datasets evaluated it by comparing it with other competitive methods. Section 5 contains a conclusion.

Dippal Israni (corresponding author) is with U and P U. Patel Department of Computer Engineering, CSPIT, CHARUSAT (e-mail: dippalisrani.ce@charusat.ac.in). Dr. Hiren Mewada is with V T. Patel Department of Electronics and Communication, CSPIT, CHARUSAT. (e-mail: hirenmewada.ec@charusat.ac.in).

II. Related Works

In automated video surveillance applications, there are several background subtractions algorithms commonly employed for object detection [30]. Object Detection is mostly categorized into three parts. (i) classifier-based approach i.e boosting classifiers [31], deep models [32], Support Vector Machine (SVM) classifiers [33] and Grammar models [34]. (ii) Feature based approach i.e Histogram of Gradient (HOG) [35], co-occurrence features [36], segmentation [37], bag of words [38], etc. (iii) Articulation handling approaches i.e mixture of parts [39], pictorial structures [40], Deformable Part based Models (DPM) [41], etc. D. Hari Hara Santosh et al. proposed detection method using weighted sum of gaussian component densities [42].

In tracking by detection paradigm, first objects are detected in each frame and then associated over time to form trajectories for each object instance. Several approaches for object association have been proposed, such as linear programming [43], energy minimization [44], dynamic programming [45], hierarchical association [46], finding min-cost flow network [47], hybrid-boosting [48], Markov chain Monte Carlo (MCMC) [20], binary quadratic programming [49], Conditional Random Field (CRF) [5], [50], hierarchical Adaptive Boosting (AdaBoosting) [26], and higher-order graph [51]. 3-level data association hierarchy was devised that consist of branch division, candidate assignment detection, candidate target detection to count the target to target occlusions by accumulative motion pairing [52]. Association method using bipartite graph matching and Hungarian algorithm are the most commonly used methods offering higher efficiency in contrast to other approaches [53, 54].

A proper choice of the cost function is essential for effective tracking performance. Zamir et al. [55] and Zhang et al. [56] used information of appearance via color histograms for learning cost without actual data. Also, affinity metrics based methods are used for tracking. A hierarchical approach proposed by Li et. al. [48] creates trajectories by combining long increasing tracklets. Bae and Yoon [57] proposed online affinities learning using Linear Discriminant Analysis (LDA). Alternate online method by training appearance models for isolated trajectories was proposed by Song et al [58].

The probabilistic formulation that finds the MAP estimate of the trajectories using a minimum cost network flow was proposed in [59]. A graph structure that encodes multiple match events as standard one to one was presented by [27]. However, these methods require having the complete video sequence in input which makes it unfeasible for online applications. Particle Framework based tracking approach was proposed using online trained instance specific classifier as a graded observation model [60]. However, it imposes few restrictions by making use of information from the past. K-shortest path based optimization method was proposed in [61] that formalizes the motion of targets along the edges of spatio temporal locations which reduces difficult estimation problem to a standard linear programming one. To discriminate the appearances of objects, Linear Discriminant Analysis (LDA) based incremental approach is proposed [57]. Yang et al. [62] handled occlusion by segmenting persons according to their foreground probability templates.

Pedestrian Tracking using integration of asynchronous HOG detection with KLT tracking and Markov-Chain Monte-Carlo Data Association (MCM-CDA) was proposed in [35]. A method using energy minimization was proposed in [44]. However, this method does not effectively track when there are sudden irregular motions. In tracking to improve data association S. Pellegrini et al. [50] proposed a method of jointly estimating pedestrian trajectories using third order CRF. However, this approach is tested only for short sequences. In surveillance, trajectories of person movement are essential and occlusion is an obstruction. To retain a person's identity, occlusion as well as demerge should be handled. Also, the algorithm should have higher tracking accuracy and precision.

This paper uses GMM based background and foreground extraction method [42] to detect multiple persons. Euclidean distance based Munkre's cost function also known as Hungarian cost assignment [54] is used for association.

A severe limitation is the incapability to model multiple objects that occlude (merge into a single object) and track them as a group while retaining optimality. Pertuz et al. [63] used focus measure operators also known as feature descriptors and computed the focus level of an image. The focus measures were divided among six various families known as Gradient based operators (GRA), Laplacian based operators (LAP), Wavelet-based operators (WAV), Statistics based operators (STA), Discrete Cosine Transform based operators (DCT) and Miscellaneous operators (MIS). They [63] concluded 36 focus measure operators having distinct characteristics. These measures helped to focus the cameras. The same algorithms can be used as descriptive information to recognize people after demerging in surveillance sequences. These feature descriptors are advantageous as less information is needed and there is no requirement to store each frame. The proposed algorithm also uses few of the focus measuring operators from these families to tackle spatially extended and temporally short and long term occlusions.

III. Feature Sets Based Proposed Person Tracking and Retention Algorithm

Most of the existing techniques are able to identify occlusion but may fail to retain the correct person's identity due to long and intense occlusion.
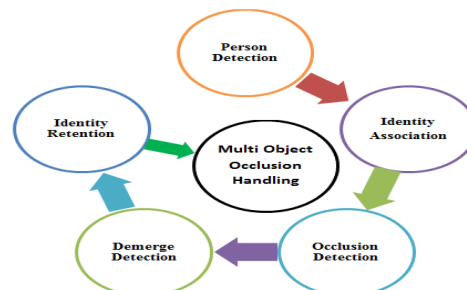

Fig. 1. Block Diagram of the proposed framework

In case of surveillance sequences, velocity and acceleration of a person's movement vary nonlinearly. This nonlinearity

causes identity retention puzzling after occlusion. Therefore, this section describes an efficient algorithm which is capable to retain identity under long and intensive occlusion. The basic steps for multi object detection and tracking are described in Fig. 1. This includes object detection, identity association, occlusion detection, demerge detection and identity retention.

A. Person Detection

Detection of person is a pre-requisite step for any tracking applications. Here, GMM is initialized using 10 frames to detect foreground [42]. GMM is a parametric probability density function represented as a weighted sum of Gaussian component densities. It is computed as,

$$p(x|\lambda) = \sum_{i=1}^{M} \omega_i g(x \mid \mu_i, \Sigma_i) \qquad (1)$$

where X is a D-dimensional continuous-valued data vector, $\lambda = \{\omega_i, \mu_i, \Sigma_i\}, i = 1, \dots, M$, $\omega_i, i = 1, \dots, M$, are the mixture weights, and $g(X|\mu_i, \sum i), i = 1 \dots M$, are the Gaussian component densities with mean vector μ_i, and covariance matrix Σ_i. GMM processed frame and detected blobs (i.e. persons) are presented in Fig. 2. To avoid undesirable noise/blob, morphological operations like opening and closing are applied using a rectangular structuring element of size [3 3] for each.
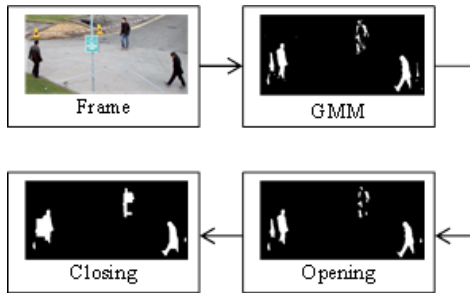

Fig. 2. Person Detection Process

After performing blob detection, centroids are detected for each blob. Then, the cost function is computed using the Euclidian distance between each blob and all other blobs as,

$$d_i = \sqrt{\sum_{j=1}^{n} (r_i - r_j)^2 + (c_i - c_j)^2} \qquad (2)$$

where, r_i, c_i are row and column locations of centroids of detected objects in the current frame, r_j, c_j are row and column locations of centroids of detected objects in the previous frame, d_i, $i=1, \dots, n$ is the distance computed between blobs of detected and current frame and n is a total number of persons (blobs). Based on this cost function, identities are used to assign to each blob.

B. Person Identity Assignment

The possibilities involved in detection are as follows. Few persons are new in the current frame, the same person is available from the previous frame or person is invisible due to occlusion. Therefore once the cost function is computed, person identity assignment is done in three categories.

1) First time appearance of a person

If the person is present for the first time, a new identity has to be assigned. For further analysis, additional details including age (the count of a person present in the scene), number of invisible counts (counter of consecutive frames from when the person got disappeared last) are also used in the proposed framework.

2) Tracking of already present person

Person tracking phase tracks all the persons in each frame as they change their positions continuously. To continuously update their positions, the optimal mapping is done for all observed persons using the Hungarian cost assignment method.

The Hungarian cost algorithm links the process between identified objects and unidentified objects in two consecutive frames. This is achieved by finding the extreme solution of the Euclidean distance in the assignment matrices. The process of position update is mentioned in equation 3.

Hungarian cost assignment finds the assignment matrix $X = [p_{ij}]$ that minimizes the total cost.

$$X = arg\min_{X} \sum_{i,j} \bar{c}_{ij} p_{ij}$$

Subject to:

$$\sum_{i} p_{ij} = 1, \forall j$$

$$\sum_{j} p_{ij} = 1, \forall i$$

(3)

where, $i, j \in \{1, 2 \dots n\}$, and $p_{ij} \in \{0,1\}$. The output X is the adjacency matrix of the track graph. $\bar{c}_{ij}$ represents the Hungarian cost matrix between tracked and detected objects. The equations model the no-overlap restrictions.

To avoid false identities due to partial appearance, the reliability of blob is calculated over a number of frames. This can lead to many identities for a single person as explained below in Fig. 3. Once the blob exceeds a defined threshold, the blob takes shape similar to a person and it can be considered as a reliable person. In Fig. 3, the area covered by a blob since its existence till it recovers as the reliable blob is described for a sequence having resolution 576 * 768. As depicted in Fig. 3, as blob area (Bounding Box area - Length * Width of the blob) of the person exceeds 4423 the person takes shape similar to a person.

Index	1	3	5	7	9
Mask					
Area	663	3807	4264	4423	5259

Fig. 3. Person's visibility in every frame

Hence, this threshold, in general, can be defined as spatial resolution (row * column)/100, i.e in figure 3, dataset having resolution 576*768 the threshold is 4423. Also, when a blob size exceeds the defined threshold, mark that person as a reliable person. Also, obtain its properties such as bounding box, identity, fifteen feature descriptors required to retain person's identity after demerge. This threshold value is optimized based on ten times of minimum blob area for GMM. This is required as the minimal blob is not sufficient to give proper features as described in Fig. 3.

3) The person is not present

Two possibilities cause the invisibility of a person. Either there is no movement of a person in a particular frame or person is occluded by the other person. GMM is a motion based model and hence it will not detect a person who has not done any movement in a particular frame. The latter situation is called as occlusion. Here the person is present in a scene still it is invisible as it is occluded by another person (occluder) in the 3D Space from captured view.

C. Occlusion Detection

As explained in section III-B-3, no movement or occlusion causes invisibility of the person. The discrimination of these two conditions is explained as follow.

```
Begin
For all invisible person
if (occlusion mark is not present) then
      estimate  most  nearest  neighbor  using
      Euclidian distance
if (height  or  width  or  both  is  increased  by
threshold) then
      Mark nearest neighbor as occluder and
      invisible as occluded object
      Update position of occluded object with
      occluder's position
end
end
```

If an object is occluded it will be occluded by its nearest object which is stated as object permanence [64]. Bounding box (height/width/both) of nearest neighbor (NN) is estimated in the proposed method and it is compared with the reliable bounding box. If NN's bounding box (height/width/both) is increased by threshold or more than invisible person, NN is marked as occluder and the invisible person is marked as occluded. If the condition is not satisfied it will be considered as no occlusion i.e no motion. The detected occlusion can be classified as self-

occlusion $S_o(.)$ and full occlusion $f(.)$. This classification can be performed using equations 4 and 5.

$$S_{b,i}(x,y) = \begin{cases} \min_{j=1,\dots N} \sqrt{\sum \left(A_{c(j)}(x,y) - A_{c(i)}(x,y) \right)}, PV = 0 \text{ and } OM = 0 \\ -, \qquad otherwise \end{cases} \quad (4)$$

$$X = \frac{S_{b,i}(x)}{Sf_{b,i}(x)}, Y = \frac{S_{b,i}(y)}{Sf_{b,i}(y)}$$

$$f(z) = \begin{cases} Mark\ Occluder\ and\ Occluded,\ X > 0.5\ or\ Y > 0.5 \\ No\ Occlusion, \qquad\qquad\qquad otherwise \end{cases} \quad (5)$$

Here PV = Person's visibility in current frame, $A_{c(i)}(x,y) = i=1, ...,M$= Centroids of M occluded object, $A_{c(j)}(x,y) j=1, ...,N$ = Centroids of all M detected persons in current frame. For every occluded object i minimum value provides the nearest object. $S_{b,i}(x, y)$ = Occluder's Centroid. $S_{b,i}(x)$, $S_{b,i}(y)$ = horizontal (width) and vertical (height) position of centroids of current frame's probable occluder, $Sf_{b,i}(x)$, $Sf_{b,i}(y)$ = horizontal (width) and vertical (height) position of centroids of reliable frame, $f(z)$ = marking Occluder and Occluded Objects. OM = Occlusion Mark. x and y represent horizontal and vertical ratio between previous centroid and current centroid. If the ratio is greater than threshold (0.50) then occluder and occluded objects should be marked. The threshold (0.5) depicts if the ratio of person's height or width or both with its reliable height width or both are increased by half. The condition does not get true when there is no occlusion and person is not visible in the current frame.

D. Demerge Detection

Demerge detection is an important task to handle occlusion and to retain the identity of persons. The demerge detection is presented in Fig. 4.
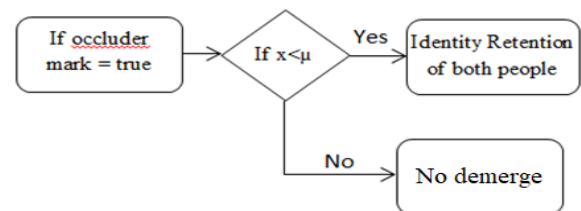

Fig. 4. Demerge Detection

The complete process is explained below.

```
Begin
For all detected persons
if (person's occluded mark is true and the
cost difference is less than the threshold)
then
      consider both people for identity
      retention
else
      no demerge
end
```

$$x = [A] - [B] < \mu$$

$$f(x) =$$
$$\begin{cases} Demerge\ Detected, & x = 1\ and\ OM = 1 \\ No\ Demerge\ , & otherwise \end{cases} \quad (6)$$

Where *[A]* = Hungarian cost function of detected object1, *[B]* = Hungarian cost function of detected object2, μ = threshold = 60, *OM* = Occlusion mark of person.

As mentioned in Fig. 4, if a person is not occluded then it can be verified by computing two minimum cost function. If cost function of two objects is less than a threshold and it's occlude mark is present, it states a particular person is no more occluded from the current frame. It is considered as demerged as presented in equation 6. Any value less than 60 can be due to shadow or incompleteness of blob. Hence, μ = 60 is selected as the a threshold

E. Features Verification and Person Identity Retention

Once the person gets demerged it requires classification. In this paper, an unsupervised approach is proposed in which persons feature descriptors are preserved once it gets reliable. To retain person identity, features of the occluder and occluded person are required to be computed. The feature descriptors are used to retain an identity as there are numerous advantages to it. The selected features are Rotation, Scaling and Translation (RST) invariant, no apriori training is required, requires very few computations and no high storage of features required. Various feature descriptors are proposed for Shape for Focus (SFF) [63]. These SFF descriptors are a mixture of Gradient based operators (GRA), Laplacian based operators (LAP), Wavelet based operators (WAV), DCT based operators (DCT), Statistics based operators (STA) as well as Miscellaneous operators (MIS) [63]. This can be used as information description of an object for recognition at demerge. Fifteen feature descriptors obtained from above operators are selected for preserving information of reliable persons. These measures are selected based on their importance and each having distinct descriptive property. The fifteen features are collected amongst all six families.

TABLE I
LIST OF USED FEATURE DESCRIPTORS

| MIS1-Absoulte Central Moment	GRA1 - Gaussian derivative	MIS7 - Steerable filters
MIS2-Brenner's measure	STA4 - Graylevel local variance	MIS8 - Spatial frequency
STA5 - Normalized gray-level variance	GRA2 – Gradient Energy	GRA6 – Tenengrad
CURV – Image Curvature	MIS5 - Hemli and Scherer's mean	MIS9 - Vollath's correlation
DCT1 – Threshold absolute gradient	LAP1 – Energy of Laplacian	WAV1 - Sum of Wavelet coeffs
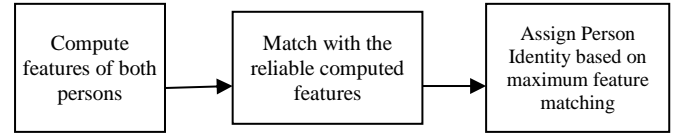
Fig. 5. Feature verification and person identity retention
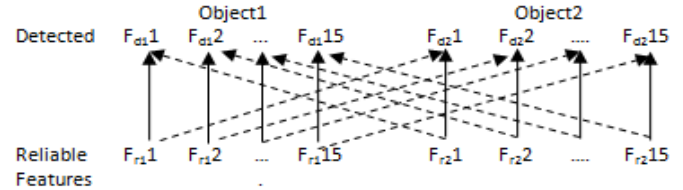

Fig. 6. Feature comparisons with reliable Person

These feature values of the occluder and occluded persons are compared with occluder and occluded features of a reliable person. The features of a person having higher similarity will retain the identity. The flow of feature comparison and person identity retention is mentioned in Fig. 5. Feature comparison for a reliable person is shown in Fig. 6,

$$sum1 = \sum_{i=1}^{15}[f_{d1}(i) - f_{r1}(i)] + \sum_{i=1}^{15}[f_{d2}(i) - f_{r2}(i)]$$

$$sum2 = \sum_{i=1}^{15}[f_{d1}(i) - f_{r2}(i)] + \sum_{i=1}^{15}[f_{d2}(i) - f_{r1}(i)] \quad (7)$$

$$Identity\ retention$$
$$= \begin{cases} f_{d1} = P_1, f_{d2} = P_2, & sum1 < sum2 \\ f_{d1} = P_2, f_{d2} = P_1, & otherwise \end{cases}$$

where $f_{d1}(i)$ = Person 1's feature set in detected frame, $f_{r1}(i)$ = Person 1's feature set in reliable frame, $f_{d2}(i)$ = Person 2's feature set in detected frame, *fr2(i)* = Person 2's feature set in reliable frame. *P1* and *P2* are person's identity in the current frame.

To retain identity, fifteen feature descriptors of demerged person are compared with the preserved reliable feature descriptors of the demerged object and with reliable feature descriptors of another object. If the difference is less for person's position and its feature sets then the same identity is retained else identities are swapped.

F. Features Verification and Person Identity Retention for Multiple Objects

This paper also proposes a unique solution for demerging of multiple people from occlusion area having multiple object. This approach includes detection of a total number of objects demerged and to provide identity to all objects. Fig. 7. depicts occlusion scenarios for a single object as well as multiple objects. For multiple objects, the possible scenarios are i) One person gets occluded with two already occluded. ii) Two persons get occluded with 2 other occluded people. iii) Multiple people get occluded by multiple people. The description of all

three scenarios can be visualized in Fig. 7.

The process of person demerging splits the blob of occluded objects in two blobs. To classify objects in both blob it is important to compute number of persons in a blob and segment that blob into number of computed persons. Hence in the proposed method, the blob having smaller area is used to identify the number of objects and classify those objects into their corresponding identities. The second blob contains remaining identities. The smaller area is chosen as smaller blob will contain less persons in it and it reduces computation time.

Let, total number of objects in small demerged area is n and it can be calculated as:

$$n = \left(\frac{AR1}{AR1 + AR2}\right) * TOB \tag{8}$$

where, TOB = total occluded objects before demerge, $AR1$ = small demerged area, $AR2$ = other demerged area.

	Single Occlusion	Multi Occlusion		
	1+1	2+1	2+2	*+*
Before Occlusion	Frame:127	Frame:97	Frame:66	Frame:337
After Occlusion	Frame:130	Frame:100	Frame:72	Frame:349

Fig. 7. Frames containing single occlusion vs multi occlusion

Once demerge is detected, a number of people demerging are computed. If a number of people demerged is 1 out of multiple objects, the same methodology as depicted in section E is repeated. In another case, when multiple objects are demerged blob of each person is required for the classification which is as computed below.

For detecting persons in a blob for classification

$$[r\ c] = size(blobarea) \tag{9}$$

Here blob area represents small demerged Area $AR1$, r and c indicate row and column of blob area respectively. This is estimated by computing total rows and columns a blob covers.

$$Sobbox = \sqrt{\frac{blobarea}{n}} \tag{10}$$

Here, n is a total number of objects demerged as computed in equation 8.
The ratio will estimate the possible area occupied by each person.

$$A = \forall_{sobbox \in image}\left(\sum_{i=1}^{sobbox}\sum_{i=0}^{sobbox} bframe_{i,j}\right) \tag{11}$$

Here, A is area estimation of a single object. *Sobbox* is the estimated size of the object that lies in the demerged blob.

$$M_1 = \max\{A_{ij}, A_{ij+1}, A_{ij+2}, \dots, A_{i+1j,Ai+1j+1}, \dots, A_{rc}\} \tag{12}$$

Here, M_1 indicates maximum blob density in a demerged area.

A maximum possible blob is selected from all segments of blobarea of size A. Similarly, from remaining segments, all other segments having maximum area are selected upto n (total persons) segments.

$$M_{2..n} = \sum_{i=2}^{n}\max\{A \setminus \bigcup M_{1:i-1}\} \tag{13}$$

Here, $M_{2..n}$ depicts all blobs up to total objects n based on

blob density. Once number of objects in a blob area is computed, the next step is to classify all objects in that blob. In this paper to identify positions of objects in a blob, row and column of bob area is estimated. Also, blob area is divided by total number of persons computed to estimate area of each person. For every estimated area person is classified using feature descriptors.

IV.   COMPARATIVE ANALYSIS AND DISCUSSION

The performance of the proposed algorithm is evaluated on a public dataset PETS2009-S2L1 [65]. This dataset contains eight videos with various challenging conditions like illumination variation and occlusions. From a social behavior point of view, this dataset is very challenging. This is because the subjects often change direction, form a group and split frequently. Sequences of View 001 having 795 frames are selected for processing. The videos are selected based on ground truth observation of a higher number of occlusion activities. The video carries a frame rate of 25 frames per second and resolution of 576*768.

The graph in Fig. 8. depicts the number of objects in ground truth vs Hypothesis (detected objects by the proposed algorithm) for the frames from 1 to 795. The minimum blob area for considering a blob as an object is set to resolution (row * column)/1000. It depicts that the number of objects in the hypothesis is almost leveled off with that in ground truth.
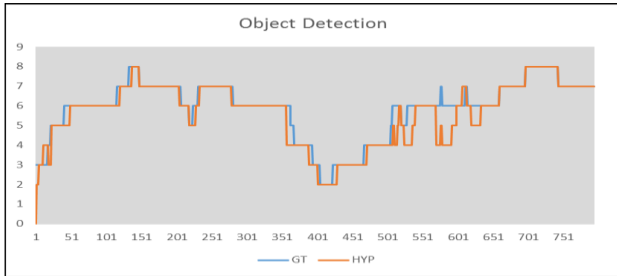

Fig. 8. Object Detection in all frames (GT vs Hyp)

The object detection accuracy is calculated using below equation 14,

$$Accuracy = \frac{2*TP}{(TP+FP)+(TP+FN)} \qquad (14)$$

where, *TP* is one that detects the object when the object is present, *FP* is one that does detects the object when the object is absent, *FN* is one that does not detect the object when the object is present. Obtained person detection accuracy is presented in Fig. 9.
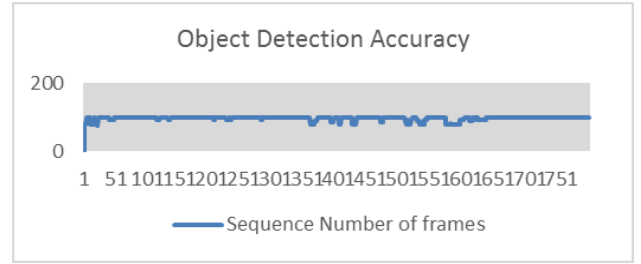

Fig. 9. Object Detection Accuracy

GMM based object detection requires 10 frames for initialization and hence the accuracy has very slow growth in the first few frames. After a few frames, the object detection accuracy rose sharply and for the rest of the frames, it can be seen that the algorithm has very high accuracy. The average object detection accuracy for the complete sequence is 97.32%.

Regarding the evaluation metrics, we adopted the most widely accepted classification of events, activities, and relationships Multi Object Tracking (CLEAR MOT) [66] [67] parameters. CLEAR MOT defined several values:

Multi-object tracking accuracy (MOTA) combines all error types [FP, false negatives/missing detections (FN) and identity mismatch] – the higher the better,

$$\text{MOTA} = 1 - \frac{\sum_t(mme_t + fp_t + m_t)}{\sum_t g_t} \qquad (15)$$

where, $\overline{mme_t}$, $\overline{fp_t}$ and $\overline{m_t}$ are the ratio of mismatches, the ratio of false positives, the ratio of misses respectively for time t and g_t is ground truth for all events. These parameters are as derived below:

Ratio of mismatches:

$$\overline{mme_t} = \frac{\sum_t mme_t}{\sum_t g_t} \qquad (16)$$

Ratio of false positives:

$$\overline{fp_t} = \frac{\sum_t fp_t}{\sum_t g_t} \qquad (17)$$

Ratio of misses:

$$\overline{m_t} = \frac{\sum_t m_t}{\sum_t g_t} \qquad (18)$$

Fig. 10. Tracking result on PETS2009-S2L1. First rows (a) refer to PETS2009-S2L1 reliable persons, whose features are stored, whereas the last row refers to two person occlusion (b) as well as multi person occlusion (c) of same sequence.

Multi Object Tracking Precision (MOTP) measures the normalized distance between tracker output and GT location i.e. the precision in the bounding box (or center of gravity) localization – the higher the better. It is a total error in estimated position for matched object-hypothesis pairs over all frames, averaged by the total number of matches made. It depicts the capability of the tracker to evaluate accurate object position,

$$MOTP = \frac{\sum_{i,t} d_t^i}{\sum_t c_t} \qquad (19)$$

where, d_t^i is the distance between the object i and its corresponding hypothesis and c_t is the number of matches found for time t. MOTP thereby gives the average overlap between all correctly matched hypotheses and their respective objects.

Identity Switches (IDs) is the total number of identity switches throughout the sequence.

The proposed approach is tested for two persons as well as for multi person occlusion scenario. Fig. 10 present occlusion detection and identity retention over possible scenarios presented in section 3.7. Fig. 10(a) represents identity assignment to each detected person. Fig. 10(b) depicts a scenario of frame 127 in which two people are about to occlude. In frame 130 they get occluded and corresponding demerge with identity retention occurs in frame 133. Similarly, Fig. 10(c) presents multi person scenario in frame 148 in which three people are occluded. In frame 157, one person is demerged and retains its identity. Finally in frame 160, other two people are also demerged and they get their original identity.

The proposed approach is compared with state of the art algorithms [59], [61], [27], [60], [62], [57]. Based on the results in Table II, it can be depicted that the proposed approach outperforms existing algorithms in benchmark dataset (PETS2009-S2L1). The proposed approach is computed online, having a monocular camera.

The proposed approach has higher MOTA. Identity switches are least and MOTP is negotiable in comparison to the described methods.

TABLE II
TRACKING RESULTS ON PETS2009-S2L1 SEQUENCE

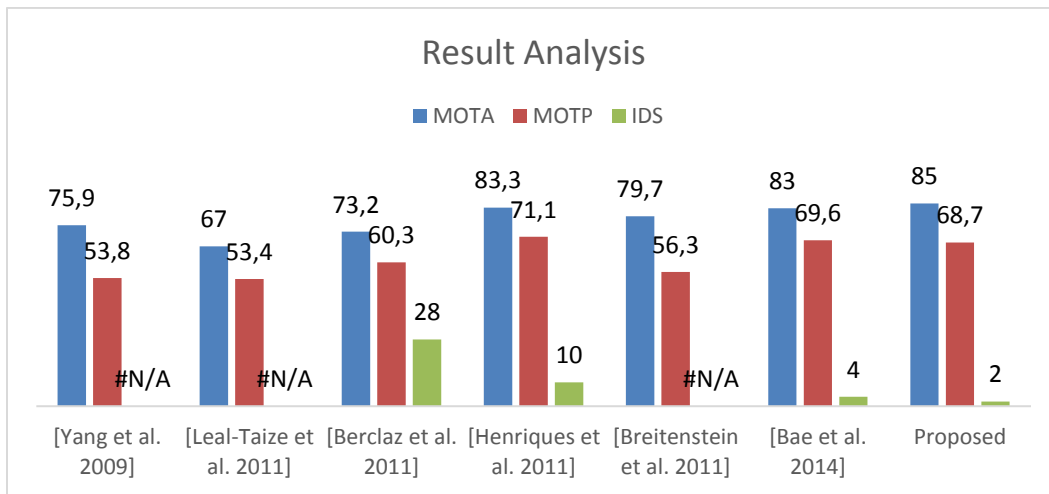| Method	MOTA	MOTP	IDS	Type	Camera	Real Time
Leal-Taixe et al., 2011 [59]	67.00%	53.40%	-	Offline	Multi	NO
Berclaz et al., 2011 [61]	73.20%	60.30%	28	Offline	Multi	NO
Henriques et al., 2011 [27]	83.30%	**71.10%**	10	Offline	Mono	NO
Breitenstein et al., 2011 [60]	74.50%	56.30%	-	Online	Mono	NO
Yang et al., 2009 [62]	75.90%	53.80%	-	Online	Mono	NO
Bae et al., 2014 [57]	83.00%	69.60%	4	Online	Multi	**YES**
Proposed	**85%**	70.70%	**2**	**Online**	Mono	NO

Fig. 11. Analysis of tracking accuracy
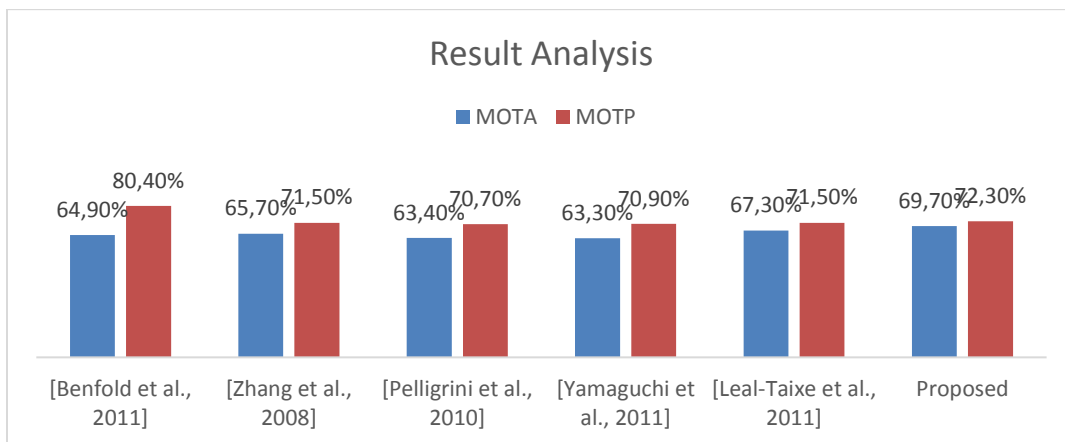
Fig. 12. Analysis of tracking accuracy Town Center Dataset

TABLE III
TRACKING RESULTS ON TOWN CENTRE DATASET

Method	MOTA	MOTP
Benfold et al., 2011 [35]	64.9%	80.4%
Zhang et al., 2008 [56]	65.7%	71.5%
Pelligrini et al., 2010 [50]	63.4%	70.7%
Yamaguchi et al., 2011 [44]	63.3%	70.9%
Leal-Taixe et al., 2011 [59]	67.3%	71.5%
Proposed	69.7%	72.3%

The corresponding graph is presented in Fig. 11. It can be clearly seen that the proposed algorithm outperforms for measures such as MOTA, IDS in online condition for Mono camera. For measures such as MOTP, the proposed approach declines steadily still the difference is moderate. Table III and Fig.12. depicts the promising outcome of the proposed approach on Town Center dataset also.

V. CONCLUSION

This paper proposed the algorithm to retain the identification of the persons under long term occlusion. The major limitation of tracking is to retain the identities of persons when they have similar characteristics like same colors, shape and size. Mixed features based classification proposed in this paper retains identity and has higher MOTA, negotiable MOTP and lower identity switches in contrast to state of the art research papers. The proposed model also includes person assignments; demerge detection along with identity retention for two as well as multi person occlusion scenarios. The experimental results show that the proposed method is efficient to tackle short and long term occlusion and self-occlusion with proper retention of a person's identity. This work can be extended in future to understand the trajectory of persons as well as behavior analysis in surveillance sequences. Also, this algorithm can be enhanced to track the identity of person even after a person completely moves away from the camera and comes back.

REFERENCES

[1] P. Nillius, J. Sullivan, and S. Carlsson, "Multi-target tracking linking identities using bayesian network inference," In Computer Vision and Pattern Recognition, New York City, New York, USA, pp. 2187–2194, 2006, DOI: 10.1109/CVPR.2006.198.

[2] J. Xing, H. Ai, L. Liu, and S. Lao, "Multiple player tracking in sports video: a dual-mode two-way bayesian inference approach with progressive observation modeling," IEEE Transactions on Image Processing, vol. 20, no. 6, pp. 1652–1667, June. 2011, DOI: 10.1109/TIP.2010.2102045.

[3] W. L. Lu, J.-A. Ting, J. Little, and K. Murphy, "Learning to track and identify players from broadcast sports videos," IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. 35, no. 7, pp. 1704–1716, July 2013, DOI: 10.1109/TPAMI.2012.242.

[4] S. Pellegrini, A. Ess, K. Schindler, and L. Van Gool, "You'll never walk alone: Modeling social behavior for multi-target tracking," In Proceedings of IEEE International Conference on Computer Vision, 2009, pp. 261–268, DOI: 10.1109/ICCV.2009.5459260.

[5] B. Yang, C. Huang, and R. Nevatia, "Learning affinities and dependencies for multi-target tracking using a CRF model," In Proceedings of IEEE Conference on Computer Vision and Pattern Recognition, 2011, pp. 1233–1240, DOI: 10.1109/CVPR.2011.5995587.

[6] Z. Khan, T. Balch, and F. Dellaert, "An MCMC-based particle filter for tracking multiple interacting targets," In Proceedings of European Conference on Computer Vision, 2004, pp. 279–290, DOI: 10.1007/978-3-540-24673-2_23.

[7] W. Luo, T.-K. Kim, B. Stenger, X. Zhao, and R. Cipolla, "Bi-label propagation for generic multiple object tracking," In Proceedings of IEEE Conference on Computer Vision and Pattern Recognition, 2014, pp. 1290–1297, DOI: 10.1109/CVPR.2014.168.

[8] E. Fontaine, A. H. Barr, and J.W. Burdick, "Model-based tracking of multiple worms and fish," In Proceedings of IEEE International Conference on Computer Vision Workshops, pp. 1–13, 2007.

[9] C. Spampinato, S. Palazzo, D. Giordano, I. Kavasidis, F.-P. Lin, and Y.-T. Lin, "Covariance based fish tracking in real-life underwater environment," In Proceedings of International Conference on Computer Vision Theory and Applications, pp. 409–414, 2012.

[10] M. Betke, D. E. Hirsh, A. Bagchi, N. I. Hristov, N. C. Makris, and T. H. Kunz, "Tracking large variable numbers of objects in clutter," In Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition, pp. 1–8, 2007, DOI: 10.1109/CVPR.2007.382994.

[11] W. Choi, and S. Savarese, "A unified framework for multi-target tracking and collective activity recognition," European Conference on Computer Vision. Springer, Berlin, Heidelberg, pp. 215-230, 2012, DOI: 10.1007/978-3-642-33765-9_16.

[12] T. Pfister, C. James and A. Zisserman, "Flowing convnets for human pose estimation in videos," In Proceedings of the IEEE International Conference on Computer Vision, pp. 1913-1921, 2015, DOI: 10.1109/ICCV.2015.222.

[13] W. Hu, T. Tan, L. Wang, and S. Maybank. "A survey on visual surveillance of object motion and behaviors," IEEE Transactions on Systems, Man, and Cybernetics, Part C (Applications and Reviews) vol. 34, no. 3, pp. 334-352, 2004, DOI: 10.1109/TSMCC.2004.829274.

[14] J. Candamo, M. Shreve, D. B. Goldgof, D. B. Sapper, and R. Kasturi, "Understanding transit scenes: A survey on human behavior-recognition algorithms," IEEE transactions on intelligent transportation systems vol. 11, no. 1, pp. 206-224, 2010, DOI: 10.1109/TITS.2009.2030963.

[15] X. Wang, "Intelligent multi-camera video surveillance: A review," Pattern Recognition Letters, vol. 34, no. 1, pp. 3–19, Jan. 2013, DOI: 10.1016/j.patrec.2012.07.005.

[16] H. Uchiyama and E. Marchand, "Object Detection and Pose Tracking for Augmented Reality: Recent Approaches," In Proceedings of Korea-Japan Joint Workshop on Frontiers of Computer Vision, pp. 721–730, 2012.

[17] W. Luo, X. Junliang, X. Zhang, X. Zhao, and T. Kim, "Multiple object tracking: A literature review," arXiv preprint arXiv: 1409.7618, 2014.

[18] B. Lee, L. Liew, W. Cheah, Y. Wang, "Occlusion Handling in Videos Object Tracking: A Survey," In IOP conference series: earth and environmental science, vol. 18, no. 1, p. 012020. IOP Publishing, Sarawak, Malaysia, pp: 12-20, January 2014, DOI: 10.1088/1755-1315/18/1/012097.

[19] B. Bose, X. Wang, and E. Grimson, "Multi-class object tracking algorithm that handles fragmentation and grouping," In IEEE Conference on Computer Vision and Pattern Recognition CVPR'07 IEEE, pp. 1-8, 2007, DOI: 10.1109/CVPR.2007.383175.

[20] Q. Yu, G. Medioni, and I. Cohen, "Multiple-target tracking by spatiotemporal monte carlo markov chain data association," IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. 31, no. 12, pp. 2196-2210, 2008, DOI: 10.1109/TPAMI.2008.253.

[21] M. Yang, T. Yu, and Y. Wu, "Game-theoretic multiple target tracking," In 11th International Conference on Computer Vision, ICCV IEEE, pp. 1-8, 2007, DOI: 10.1109/ICCV.2007.4408942.

[22] W. Hu, X. Li, W. Luo, X. Zhang, S. Maybank, and Z. Zhang, "Single and multiple object tracking using log-euclidean Riemannian subspace and block-division appearance model," IEEE Transactions on Pattern Analysis and Machine Intelligence 34, no. 12 (2012): 2420-2440, DOI: 10.1109/TPAMI.2012.42.

[23] L. Zhang and L. van der Maaten, "Structure preserving object tracking," In IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp. 1838-1845, 2013, DOI: 10.1109/CVPR.2013.240.

[24] J. Zhang, L. L. Presti, and S. Sclaroff, "Online multi-person tracking by tracker hierarchy," In Ninth International Conference on Advanced Video and Signal-Based Surveillance (AVSS) IEEE, pp. 379-38, 2012, DOI: 10.1109/AVSS.2012.51.

[25] Y. Xiang, A. Alahi, and S. Savarese, "Learning to track: Online multi-object tracking by decision making," IEEE international conference on computer vision (ICCV) IEEE, pp. 4705-4713, 2015, DOI: 10.1109/ICCV.2015.534.

[26] C. H. Kuo, C. Huang, and R. Nevatia, "Multi-target tracking by on-line learned discriminative appearance models," In IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp. 685-692. IEEE, 2010, DOI: 10.1109/CVPR.2010.5540148.

[27] J. F. Henriques, R. Caseiro, and J. Batista, "Globally optimal solution to multi-object tracking with merged measurements," In IEEE International Conference on Computer Vision (ICCV), pp. 2470-2477. IEEE, 2011, DOI: 10.1109/ICCV.2011.6126532.

[28] Z. Qin and C. R. Shelton, "Improving multi-target tracking via social grouping," In IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp. 1972-1978, 2012, DOI: 10.1109/CVPR.2012.6247899.

[29] B. Yang and R. Nevatia, "An online learned CRF model for multitarget tracking," in Proc. IEEE Conference on Computer Vision and Pattern Recognition, pp. 2034–2041, 2012, DOI: 10.1109/CVPR.2012.6247907.

[30] S. Shantaiya, K. Verma, K. Mehta, "Multiple Object Tracking using Kalman Filter and Optical Flow," European Journal of Advances in Engineering and Technology, vol. 2, vo. 2, pp:34-39, 2015.

[31] P. Doll´ar, Z. Tu, P. Perona, and S. Belongie, "Integral channel features". In British Machine Vision Conference (BMVC), pp. 91.1-91.11, 2009.

[32] W. Ouyang and X. Wang, "A discriminative deep model for pedestrian detection with occlusion handling," In IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp. 3258-3265, 2012, DOI: 10.1109/CVPR.2012.6248062.

[33] P. Felzenszwalb, R. B. Grishick, D.McAllister, and D. Ramanan, "Object detection with discriminatively trained part based models," IEEE Transactions on Pattern Analysis and Machine Intelligence, Vol. 32, no. 9, pp. 1627–1645, 2010, DOI: 10.1109/TPAMI.2009.167.

[34] R. Girshick, P. Felzenszwalb, and D. McAllester, "Object detection with grammar models," In Advances in Neural Information Processing Systems, 2011.

[35] B. Benfold, and I. Reid. "Stable multi-target tracking in real-time surveillance video," In IEEE Conference on Computer Vision and Pattern Recognition (CVPR) IEEE, pp. 3457-3464, 2011, DOI: 10.1109/CVPR.2011.5995667.

[36] W. Schwartz, A. Kembhavi, D. Harwood, and L. Davis, "Human detection using partial least squares analysis," In 12th International Conference on Computer Vision (ICCV), pp. 24-31, 2009, DOI: 10.1109/ICCV.2009.5459205.

[37] M. Enzweiler and D. M. Gavrila, "A multilevel mixture-of experts framework for pedestrian classification," IEEE Transactions on Image Processing, vol. 20, no. 10, pp.2967–2979, 2011, DOI: 10.1109/TIP.2011.2142006.

[38] C. Lampert, M. Blaschko, and T. Hofmann, "Beyond sliding windows: object localization by efficient subwindow search," In IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp. 1-8, 2008, DOI: 10.1109/CVPR.2008.4587586.

[39] Y. Yang and D. Ramanan, "Articulated pose estimation with flexible mixtures-of-parts," In IEEE Conference on Computer Vision and Pattern Recognition, pp. 1385-1392, 2011, DOI: 10.1109/CVPR.2011.5995741.

[40] P. F. Felzenszwalb and D. P. Huttenlocher, "Pictorial structures for object recognition," International Journal of Computer Vision, vol. 61, no. 1, pp. 55–79, 2005, DOI: 10.1023/B:VISI.0000042934.15159.49.

[41] L. Zhu, Y. Chen, A. Yuille, and W. Freeman, "Latent hierarchical structural learning for object detection," In IEEE Conference on Computer Vision and Pattern Recognition, pp. 1062-1069, 2010, DOI: 10.1109/CVPR.2010.5540096.

[42] D. Reynolds, "Gaussian mixture models," Encyclopedia of biometrics, pp. 659-663, 2009, DOI: 10.1007/978-1-4899-7488-4_196.

[43] A. Andriyenko, K. Schindler, "Globally optimal multi-target tracking on a hexagonal lattice," In 11th European Conference on Computer Vision, pp. 466-479, 2010, DOI: 10.1007/978-3-642-15549-9_34.

[44] K. Yamaguchi, C.B. Alexander, E. O. Luis, and L. B. Tamara, "Who are you with and where are you going?", In IEEE Conference on Computer Vision and Pattern Recognition (CVPR) IEEE, pp. 1345-1352, 2011, DOI: 10.1109/CVPR.2011.5995468.

[45] J. Berclaz, F. Fleuret, and P. Fua, "Robust people tracking with global trajectory optimization," Proceedings of IEEE conference on Computer Vision and Pattern Recognition, pp. 744-750, 2006, DOI: 10.1109/CVPR.2006.258.

[46] C. Huang, B. Wu, and R. Nevatia, "Robust object tracking by hierarchical association of detection responses," In 10th European Conference on Computer Vision, pp. 788-801, 2008, DOI: 10.1007/978-3-540-88688-4_58.

[47] H, Pirsiavash, D. Ramanan, C. Fowlkes, "Globally-optimal greedy algorithms for tracking a variable number of objects," In Proceedings of IEEE conference on Computer Vision and Pattern Recognition, pp. 1201-1208, 2011, DOI: 10.1109/CVPR.2011.5995604.

[48] Y. Li, C. Huang, and R. Nevatia, "Learning to associate: Hybrid boosted multi-target tracker for crowded scene," In Proceedings of IEEE conference on Computer Vision and Pattern Recognition, pp. 2953-2960, 2009, DOI: 10.1109/CVPR.2009.5206735.

[49] B. Leibe, K. Schindler, and L. V. Gool, "Coupled detection and trajectory estimation for multi-object tracking," In Proceedings of IEEE International Conference on Computer Vision, pp. 1-8, 2007, 10.1109/ICCV.2007.4408936.

[50] B. Yang, R. Nevatia, "Multi-target tracking by online learning of non-linear motion patterns and robust appearance models," In Proceedings of IEEE conference on Computer Vision and Pattern Recognition, pp. 1918-1925, 2012, 10.1109/CVPR.2012.6247892.

[51] S. Pellegrini, E. Andreas and L. Gool, "Improving data association by joint modeling of pedestrian trajectories and groupings," In European conference on computer vision, Springer, Berlin, Heidelberg, pp. 452-465, 2010, DOI: 10.1007/978-3-642-15549-9_33.

[52] Y. Yi and H. Xu, "Hierarchical Data Association Framework with Occlusion Handling for Multiple Targets Tracking," IEEE signal processing letters, vol. 21, no. 3, pp. 288-291, 2014, DOI: 10.1109/LSP.2014.2300497.

[53] H. W. Kuhn, "The Hungarian Method for the Assignment Problem," Naval Research Logistics Quarterly, vol. 2, pp. 83–97, 1955, DOI: 10.1002/nav.3800020109

[54] J. Munkres, "Algorithms for the Assignment and Transportation Problems," Journal of the Society for Industrial and Applied Mathematics, vol. 5, no. 1, pp.32–38, 1957, DOI: 10.1137/0105003.

[55] A. R. Zamir, A. Dehghan, and M. Shah, "GMCP-Tracker: Global Multi-object Tracking Using Generalized Minimum Clique Graphs". In Computer Vision- ECCV, pp. 343-356, 2012, DOI: 10.1007/978-3-642-33709-3_25.

[56] L. Zhang, Y. Li, and R. Nevatia, "Global Data Association for Multi-Object Tracking Using Network Flows," Proceedings of IEEE conference on Computer Vision and Pattern Recognition, pp. 1-8, 2008, DOI: 10.1109/CVPR.2008.4587584.

[57] S. H. Bae and K.-J. Yoon, "Robust Online Multi-Object Tracking based on Tracklet Confidence and Online Discriminative Appearance Learning," Proceedings of IEEE conference on Computer Vision and Pattern Recognition, pp. 1218-1225, 2014, DOI: 10.1109/CVPR.2014.159.

[58] X. Song, J. Cui, H. Zha, and H. Zhao, "Vision-based Multiple Interacting Targets Tracking via On-line Supervised Learning," In European Conference on Computer Vision Springer, Berlin, Heidelberg, pp. 642-655, 2008, DOI: 10.1007/978-3-540-88690-7_48.

[59] L. Leal-Taixe, G. Pons-Moll, and B. Rosenhahn, "Everybody needs somebody: Modeling social and grouping behavior on a linear programming multiple people tracker," In International Conference on Computer Vision Workshops, pp. 120-127, 2011, DOI: 10.1109/ICCVW.2011.6130233.

[60] 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, DOI: 10.1109/TPAMI.2010.232.

[61] J. Berclaz, F. Fleuret, E. Turetken, and P. Fua, "Multiple Object Tracking Using K-Shortest Paths Optimization," IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. 33, no. 9, pp. 1806-1819, 2011 DOI: 10.1109/TPAMI.2011.21.

[62] J. Yang, Z. Shi, P. Vela, and J. Teizer, "Probabilistic multiple people tracking through complex situations," International Workshop on Performance Evaluation of Tracking and Surveillance, pp. 79-86, 2009.

[63] S. Pertuz, D. Puig, M. A. Garcia, "Analysis of Focus Measure Operators for Shape-from-Focus," Pattern Recognition, Elsevier, vol. 46, no. 5, pp. 1415-1432, 2013, DOI: 10.1016/j.patcog.2012.11.011.

[64] R. Baillargeon, "Object Permanence in Five-Month-Old Infants," Cognition, Vol. 20 pp. 191-208, 1985, DOI: 10.1016/0010-0277(85)90008-3.

[65] J. Pan, Bo. Hu, "Robust Occlusion Handling in Object Tracking," IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Los Angeles CA USA, 2007, DOI: 10.1109/CVPR.2007.383453.

[66] K. Bernardin, R. Stiefelhagen, "Evaluating Multiple Object Tracking Performance: The CLEAR MOT metrics," EURASIP Journal of Image and Video Processing, vol. 2008, no. 1, pp. 1–10, 2008, DOI: 10.1155/2008/246309.

[67] A. Milan, L. Leal-Taixe, I. Reid, S. Roth, and K. Schindler, "MOT16: A Benchmark for Multi-Object Tracking," arXiv:1603.00831 [cs], May 2016.

Dippal P. Israni received his M.Tech in Computer Engineering from Dharmsinh Desai University. He has been pursuing his Ph.D in Computer Engineering from U and P U Patel Department of Computer Engineering, CSPIT, CHARUSAT. He is also Assistant Professor in same department and Assistant Scientist in Charusat Space Research and Technology Centre, CHARUSAT. His major research interests include Computer Vision, Image Processing and Machine Learning.

Dr. Hiren Mewada received his Ph.D and M.Tech from Sardar Vallabhbhai National Institute of Technology, Surat. He is Associate Professor in V T. Patel Department of Electronics and Communication Engineering, CSPIT, CHARUSAT. His major research interests include Computer Vision, Image Processing, FPGA based Signal Processing and Machine Learning.