International Journal of Engineering

Journal Homepage: www.ije.ir

Improved Object Matching in Multi-objects Tracking Based On Zernike Moments and Combination of Multiple Similarity Metrics

A. Dadgar, Y. Baleghi*, M. Ezoji

Department of Electrical & Computer Engineering, Babol Noshirvani University of Technology, Babol, Iran

PAPER INFO

Paper history:
Received 10 January 2021
Received in revised form 08 April 2019
Accepted 15 April 2021

Keywords:
Multi-objects Tracking
Zernike Moments
Gaussian Mixture Model
Hausdorff Distance
Earth Mover's Distance

ABSTRACT

In video surveillance, multiple objects tracking (MOT) is a challenging task due to object matching problem in consecutive frames. The present paper aims to propose an improved object matching approach in MOT based on Zernike moments and combination of multiple similarity distance metrics. In this work, the object is primarily detected using background subtraction method while the Gaussian Mixture Model (GMM) is applied for object extraction in the next frames. Subsequently, the color histogram and the magnitude of Zernike moments of the objects are calculated. In the next step, the objects are matched in the current and the previous frames based on the Hausdorff distance between objects, Earth Mover's distance (EMD) between their color histograms, and Chi-square distance between their Zernike moments. Then, a voting mechanism is designed to find the best consensus object matching from the aforementioned metrics. Eventually, the location of each object is predicted by the Kalman filter to continue tracking in subsequent frames. The results show that the object tracking and matching performance is improved using the proposed method in the video sequences of the multi-camera pedestrian "EPFL" video dataset. Specifically, errors caused by the merging of targets are reduced in the proposed tracking process.

doi: *10.5829/ije.2021.34.06c.08*

1. INTRODUCTION

Multi-Objects tracking (MOT) is an important task in computer vision and is often one of the first steps for video analysis in surveillance, sports, or industrial applications. In contrast to Single Object Tracking (SOT), the number of objects will vary in this approach and may merge, split, appear, or disappear in the scene over time. Due to challenges such as object shape deformations, brightness changes, and the issues of occlusion and distraction, the existing approaches still do not perform properly in all situations [1-4]. Increasing the number of objects creates a new and major difficulty in detection, data association and tracking. In MOT, special attention should be paid to determine the identity of each object at any time and to maintain the consistency of objects identities during tracking, and to solve multi identeties matching challenges [5-7]. Accordingly, validation criteria must be considered alongside each object to perform a proper object matching during the video sequence. In this regard, selecting suitable features that can separate the objects from the foreground or other scene objects, implementing and updating a robust model for the objects, and occlusion occurrence, are issues that will be addressed in the MOT process [8-9]. In the present study, the data association in the MOT process is based on extracting orthogonal Zernike moments [10-11] and also combination of three similarity metrics e.g., Hausdorff distance [12-13], Earth Mover's distance (EMD) [14] and Chi-square distance to improve object matching.

The present article is arranged as follows; a literature review will be carried out in the second section, and subsequently, the proposed method will be discussed in the third section. The results and evaluation will be presented in the fourth section. The fifth section will discuss the results, and the conclusion will be presented in the sixth section.

*Corresponding Author Institutional Email: y.baleghi@nit.ac.ir (Y. Baleghi)

Please cite this article as: A. Dadgar, Y. Baleghi, M. Ezoji, Improved Object Matching in Multi-objects Tracking Based On Zernike Moments and Combination of Multiple Similarity Metrics, International Journal of Engineering, Transactions C: Aspects Vol. 34, No. 6, (2021) 1445-1454

2. LITERATURE REVIEW

The MOT consists of two main parts; the first is object detection, and the second performs the association between the detected and tracked objects. Challenges such as the occlusion may yield undesirable results during the tracking process. In the following, the related investigations will be discussed regarding tracking and data association.

Object tracking utilizes two groups of non-predictive and predictive algorithms, depending on the situation. In the first group, the tracking is performed based on matching [15]. More specifically, by detecting the target area in each frame, the area of the next frame that most closely resembles the mentioned area is considered as the object area. In other words, no prediction is made about the target position in the next frame according to its current movement (e.g., Mean shift and CAM shfit) [16]. In the second group, the tracking is performed by algorithms that possess predictive features. The stated algorithms use the object position in frame k to predict the target position in frame $k + 1$ (e.g., Kalman filter and particle filter) [16-18]. The tracking problem can be considered as a posterior probability density function (PDF) estimation of the object's state variable [19]. In other words, the target's probability distribution is estimated in the current frame and desired in the next frame. The same framework is used as the basis of the tracker in the present paper.

In the MOT process, the objects are primarily detected, and subsequently, the association of the detected objects in the present frame must be established with the objects in the previous frames. The Nearest Neighbor (NN) and General Nearest Neighbor (GNN) methods are two common approaches of data association. The stated methods may also be inaccurate when the object areas are close to each other or when the number of incorrect measurements increases [20]. The Hungarian method is a combinatorial optimization algorithm that solves the assignment problem in polynomial time [21].

The Joint Probabilistic Data Association (JPDA) method has been proposed to renovate the GNN. Thus, each path is updated by the weighted sum of all measurements. The PDA method encounters several targets independently [22-23].The Multiple Hypothesis Tracking (MHT) method is a statistical association algorithm that may even postponed the data association to the next repetitions to eliminate the ambiguities. Generally, the MHT method is composed of the sections such as the generation of hypothesis matrix, generation of hypothesis, calculation of hypothesis probability, calculation of the Kalman filter associated with the target, and hypothesis management [24]. As a result, several hypotheses will be the output of one hypothesis in cases such as occlusion and noisy situation. However, the computational cost can be high, depending on the application complexity [25].

Although the majority of data association algorithms, such as JPDA and MHT, take the peer-to-peer measurements and objectives into account, the Markov Chain Monte Carlo data association method does not work based on such hypotheses. In general, the Monte Carlo method is an approximate solution that considers the problem as a hybrid optimization problem and examines it through random space exploration, rather than enumerating all association options [26]. However, issues like long occlusions, severe video blur, sudden movements of the camera, and disruption of the state of targets can cause tracking failure. MOT algorithm should be able to establish unique correspondences between objects in each frame of a video sequence.

A new trend is to use object features and object matching for both tracking and association. Image/Object matching has rich meaning in pairing two objects, thus deriving many specific tasks, such as sparse feature matching, dense matching (like image registration and stereo matching), patch matching (retrieval), 2-D and 3-D point set registration, and graph matching [27].

A gradient based corner response uses the first order information in image to distinguish the corner feature. The famous Harris corner detector was introduced to address the anisotropy and computation complexity problems [28]. The goal of the Harris method is to find the directions of the fastest and lowest grey value changes using a second-order moment matrix or an autocorrelation matrix; thus, it is invariant to orientation and illumination and has reliable repeatability and distinctiveness [29].

In methods based on second-order partial derivatives, the Laplacian of Gaussian (LoG) is applied based on scale space theory. The difference of Gaussians (DoG) [30] filter can be used to approximate the LoG filter, and greatly speed up the computations. Another classical blob feature detection strategy is based on the determinant of Hessian (DoH) [31]. This is more affine invariant because the eigen value and eigen vector of the second-order matrix can be applied to estimate and correct the affine region [29].

Interest point detection using DoG, DoH, and both has been widely utilized in recent visual applications. The famous Scale Invarient Feature Transform (SIFT) [30], extracts key point as the local extrema in a DoG pyramid, using the Hessian matrix of the local intensity values. Speeded up Robust Features (SURF) [32] accelerates the SIFT by approximating the Hessian matrix based detector using Haar wavelet calculation, together with an integral image strategy, thus simplifying the construction of a second order differential template [29-33].

Ucar et al. [34] put forward a novel hybrid Local Multiple system based on Convolutional Neural Networks (CNNs) and Support Vector Machines (SVMs)

with the feature extraction capability and robust classification. In the proposed system, they first divided the whole image into local regions using the multiple CNNs. Secondly, they selected discriminating features using principal component analysis (PCA) and imported them into multiple SVMs by both empirical and structural risk minimizations. Finally, they tried to fuse SVM outputs. They worked on the pre-trained AlexNet and also performed object recognition and pedestrian detection experiments on the Caltech-101 and Caltech Pedestrian datasets. Their proposed system generated better results with the low miss rate and improved object recognition and detection with an increase in accuracy. Zhou *et al.* [35] presented the architecture and the algorithms of deep learning in an application of object detection task. They worked on built-in datasets such as ImageNet, Pascal Voc, CoCo and deep learning methods for object detection. They created their own dataset and proved that using CNN for the object detection, the results are improved. Experiments proved that the deep learning is an effective tool to pass the man-made feature with the large qualitative data. Kaushal *et al.* [33] conducted a comprehensive survey on object detection and tracking in videos techniques based on the deep learning. The survey included neural network, deep learning, fuzzy logic, evolutionary algorithms required for detection and tracking. In the survey, they discussed various datasets and challenges for the object detection and tracking based on the deep neural network.

On the other hand, the application of deep learning strategies also burdens high computational cost and requires high performance computing devices and very large datasets to reach proper results. Hence, in this paper we have proposed an improved object matching method for MOT algorithm based on powerful features including Zernike Moments and combination of some distance metrics, EMD, Hausdorff and Chi-square. The results are also compared with a deep learning based approach in this paper.

3. THE PROPOSED METHOD

The main purpose of the present work is to track multiple objects in consecutive frames and solving some MOT challenges with new strategies. The tracking is performed in each frame according to the flow diagram illustrated in Figure 1. In the mentioned procedure the object is primarily obtained using background subtraction method while the Gaussian Mixture Model (GMM) is applied for object extraction in the next frames. Subsequently, the color histograms are considered for object matching and the objects' Zernike Moments are calculated for data association. In the next step, the objects are matched in the current and the previous frames based on the combination of similarity metrics: Hausdorff distance between objects, EMD distance between their color histograms, and Chi-square distance between their Zernike Moments. Eventually, the location of each object is predicted by the Kalman filter to continue tracking in subsequent frames. Thus, the predictors are updated based on the new detections, and the predicted locations are returned to the tracking process again. The same procedure continues until the last frame. Meanwhile, a new tracker starts tracking objects that do not match any of the predictions, and also the stopped trackers will be eliminated. In this paper the following challenges are considered: object fragmentation during the object detection, merging two or more objects and thus sharing a single ID, and the association of multiple identities under conditions of exit, re-entry, as well as occlusion.

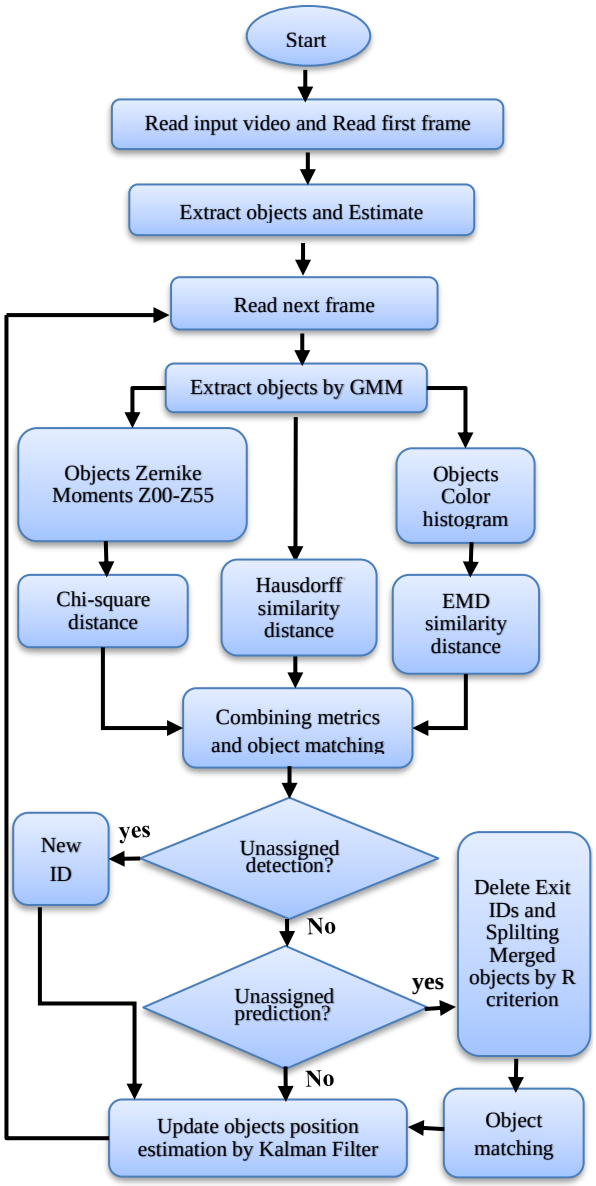

Figure 1. Flow diagram of the proposed method

To deal with the issue, an improved object matching is proposed in this work based on Zernike Moments and combination of three distance metrics e.g., EMD, Hausdorff and Chi-square that increases the multi-objects tracking accuracy. A detailed description of every step of the proposed method is put forth in the following subsections.

3. 1. Target Detection and Feature Extraction In the object tracking systems, the moving areas must be subtracted from the background, and the objects models must be created prior to the initiation of the tracking process.

The present study utilizes the GMM for object detection, which possesses decent prediction capabilities since it can accurately model any types of probability density function with a sufficient number of Gaussian functions. The GMM is among the pattern recognition systems and is defined as Equation (1) [36].

$$p(x) = \sum_{k=1}^{K} \pi_k \, N(x|\mu_k \, , \Sigma \mathrm{k}) \quad , \quad \sum_{k=1}^{K} \pi_k = 1 \qquad (1)$$

where, π_k denotes the weight of the k[th] distribution. Also, μ_k and Σk represent the mean and covariance of the k[th] cluster, respectively. In this paper, Zernike Moments (ZM), the powerful feature extractor, has been used for object matching that significantly improved the preformance of sole color features.

The ZM feature is able to determine the overall object shape in low orders and represent the object details in high orders. ZM is a powerful descriptor with features including orthogonality, low sensitivity to noise, and insensitivity to rotation. The ZM is used in numerous applications, such as character recognition, palm-print recognition, recognition of different languages in old texts, signature-based authentication, and face-based recognition. The Zernike polynomial is defined in a unit radius circle [10-37]. The mixed two-dimensional ZM of order n and with repetition m is defined as Equation (2), [38-39]. In Equation (2), $f(x,y)$ is digital image with the dimension of $M \times N$ related to intensity function of the input image and * denotes the complex conjugate.

$$ZM_{n,m}\big(f(x,y)\big) = \frac{n+1}{\pi} \sum_{i=0}^{M-1} \sum_{j=0}^{N-1} V_{n,m}^{*}(x,y) \; f(x,y) \qquad (2)$$

The order *n* is a non-negative integer and *m* is an integer which satisfies condition $|\, m \,| \leq n$. Zernike polynomials, $V_{n,m}(x,y)$ and Radial polynomials $R_{mn}(r)$, are defined as Equation (3).

$$V_{n,m}(x,y) = R_{n,m}(r)\, e^{jm\theta} \quad , \quad R_{n,m}(r) = \sum_{s=0}^{(n-|m|)/2} (-1)^s \cdot \frac{(n-s)!}{s!\left(\left(\frac{(n+|m|)}{2}-s\right)!\left(\frac{(n-|m|)}{2}-s\right)!\right)} r^{n-s} \qquad (3)$$

where; $r = \sqrt{x^2 + y^2}$ is the length of the vector that connects the origin of the coordinates to the pixel with the coordinates (x,y), and $\theta = \tan^{-1}(\frac{y}{x})$ [38-39]. It should be noted that, Zernike polynomials are defined within a unit circle, and the coordinates of the images $f(x_i, y_j)$ must be normalized into [− 1, 1] by a mapping transformation. $f(x,y)$ is the image function after mapping to a unit circle. The pixels located outside the circle are not involved in the calculations. The center of the bounding box in which the object is detected is the origin of the coordinates.

3. 2. Distance Similarity As displayed in the flow diagram of Figure 1, distance criteria based on local similarity, statistical/non-statistical similarity, or global similarity can be used to address the challenges of data association in a camera view and to perform target matching. Euclidean distance is a common method of calculating the distance between two data sets. In this study, to strengthen the separability of objects and improve the assignment of the object identities, in addition to ZM feature extractor, the EMD similarity criterion and the Hausdorff similarity metric have also been used. A number of criteria that are effective in improving the object matching results of this article are stated in the following.

The Hausdorff Distance (HD) is a similarity metric between two sets of points. The HD between two finite sets of points including A and B is the maximum of minimum distances between each point $a \in A$ to its nearest neighbor $b \in B$. HD can be calculated by, $h(A,B) = Max_{\,a \in A} \{ Min_{b \in B} \{\|a - b\|\}\}$ and $HD(A,B) = Max\,(h(A,B), h(B,A))$. In general, the values of h(A, B) and h$(B,$A) can be substantially different. So the HD, is the maximum of the directed HD in both directions and thus it is symmetric [12-13]. The EMD method begins by answering this question: What is the lowest cost to convert one distribution to another, assuming that two histograms have the same number of columns and frequecies [14].

The EMD can be stated in terms of a linear programming problem: two distributions represented by signatures P and Q, where p_i , q_i are bin centroids with frequencies w_{pi} , w_{qi}; and $D = [d_{ij}]$ is the matrix containing the Euclidean distances between p_i and q_j for all *i,j*. We ensure that *P* and *Q* have the same total mass of unity, equal to one, by normalizing each of distributions. Next, we find a $F = [f_{ij}]$ between p_i and q_j that minimizes the total cost. In Figure 2, dij represents the distance between the two columns that the value is transferred among them, and f_{ij} represents the amount of transferred value. The similarity of the two histograms can be expressed based on Equation (4). In the following equations *n* and *m* are the number of histogram bins. The objective function denotes the set of all feasible flows between bins. Equation (4) will be optimized according to the variables and constraints [14].

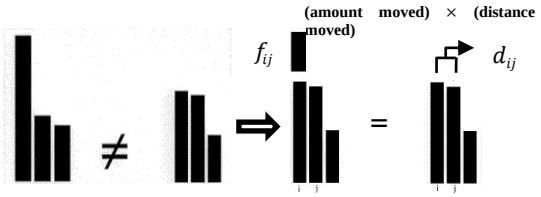

Figure 2. The histogram with different statistical distributions

$$\begin{aligned} &\textstyle\sum_{i=1}^{m}\sum_{j=1}^{n}(distance\ moved)\times(amount\ moved)\\ &Cost(P,Q,F)=\textstyle\sum_{i=1}^{m}\sum_{j=1}^{n}(d_{ij})\times(f_{ij})\\ &\textstyle\sum_{i=1}^{m}\sum_{j=1}^{n}(f_{ij}) \;=\; min(\sum_{i=1}^{n}p_i\ ,\sum_{j=1}^{n}q_j\ )\\ &EMD=\frac{\sum_{i=1}^{m}\sum_{j=1}^{n}(d_{ij})\times(f_{ij})}{\sum_{i=1}^{m}\sum_{j=1}^{n}\ f_{ij}} \end{aligned} \quad (4)$$

Solving the above linear programming problem determines the optimal flow, between the source and destination. In other words, the conversion from P to Q is performed by removing at least parts of the columns in P. Thus, the EMD method goes from an optimization problem to a minimization problem [14].

3. 3. Object Matching Data association and objects matching, in the video sequence are influential topics in the MOT. The tracking process searches for the correct association between the foreground and the predicted objects at time t. The magnitude of the ZM are calculated in the detected objects at time t and the prediction of objects in time $t-1$ and compared by the Chi-square distance according to Equation (5).

Object matching has been done according to combination of similarity distances metrics: Hausdorff distance between objects, EMD distance between their color histograms, and Chi-square distance between their Zernike Moments magnitudes. To improve the assignment of objects' identity, hard voting will be conducted between the three similarity criteria, and for each object, an identity that has been approved by the majority will be assigned.

$$D(O_t,O_{t+1})=\ \sum_{i=1}^{n}\frac{\left(O_{t_n}-O_{t_{n+1}}\right)^2}{O_{t_n}+O_{t_{n+1}}} \quad (5)$$

where; O_{t_n} and $O_{t_{n+1}}$ are the feature values of the objects.

Subsequently, object matching is conducted based on the thresholded values. Since the detected objects are less than the predictions, there is a possibility of occlusion. Hence, the merging and splitting of the bounding boxes must be evaluated. So, the shape bounding box's metric is defined as R. The R criterion is expressed according to Equation (6).

$$\begin{aligned} &R_i=\frac{Height}{Width},\\ &R_i>\tau_{RatioDown}\quad and\quad R_i<\tau_{RatioUp}\ ,\\ &\tau_{RatioDown}<R_i<\tau_{RatioUp} \end{aligned} \quad (6)$$

where; $Height$ and $Width$ are the height and the width of the bounding box of the object, respectively. The R criterion is limited by the upper bound threshold of $\tau_{RatioUp}$, and the lower bound threshold of $\tau_{RatioDown}$.

3. 4. Object Tracking In this paper, the object tracking is performed using the Kalman filter[40]. The Kalman filter is a recursive estimator with the minimum variance. Kalman filter consists of two groups of time update and measurement update equations, which are also referred to as predictive and corrective equations, respectively [40].

Equations (7)-(11) represent the Kalman equations. Equation (7) predicts the state of the system. This prediction is made without observing the current moment. Equation (8) shows the current prediction error, and Equation (9) is responsible for calculating the gain. Similarly, Equation (10) shows the estimation of the Kalman filter of the system state. The error corresponding to this estimation is given in Equation (11).

$$\hat{x}_{k+1}^{-}\ =\Phi_k\hat{x}_k \quad (7)$$

$$P_{k+1}^{-}=\Phi_kP_k{\Phi_k}^T+Q_k \quad (8)$$

$$k_k=p_k^-\ H_k^T\left(H_k\,p_k^-\ H_k^T+R_k\right)^{-1} \quad (9)$$

$$\hat{x}_k=\hat{x}_k^-+K_k(z_k-H_k\hat{x}_k^-) \quad (10)$$

$$P_k=(I-k_kH_k)P_k^- \quad (11)$$

In the above Equations, x_k is the state vector at moment k, Φ_k is the transition matrix, and Y_k is the system output. H_k is the conversion matrix, and vector Z_{k} is the sum of the output of the Y_{k} with the measurement noise of V_{k} (system observations). $\hat{x}_{\mathrm{k}}^-$ is the previous prediction, $\hat{x}_{\mathrm{k}}$ is the subsequent prediction, and $\hat{x}_{\mathrm{k+1}}^-$ is the previous prediction of $K+1$. w_{k} and v_{k} are the measurement and process noises, respectively. K_k is the Kalman gain, and p_{k} is the covariance error matrix.

4. RESULTS AND EVALUATION

In this work, the MOT in a camera view is performed by the modification of data association based on Zernike Moments features, similarity criteria, and Kalman filter. This study attempts to maintain the continuity of tracking each object during the MOT process. For this purpose, a unique and new ID is assigned to each object. The stated ID must remain constant during the tracking process even after short occlusion.

4. 1. Database and Evaluation Metrics The video sequences of the multi-camera pedestrian video

"EPFL" dataset are used in the simulations[1]. In this dataset, several videos have been recorded simultaneously from a specific location using multiple cameras at various angles. This sequence consists of people appearing one after the other, and walking in front of the cameras. It tests the ability of our algorithm to cope with a moderately crowded environment. The calibration information and homography matrices, H, are provided for each camera. The homographies given in the calibration files[4] project points in the camera views to their corresponding location on the top view of the ground plane, that is presented in Equation (12). In this equation, $\boldsymbol{X_{image}}$ is the object position in tracking process and $\boldsymbol{X\_topview}$, is the corresponding location on the top view of the ground plane.

$$H \times X_{image} = X\_topview \tag{12}$$

Accuracy and precision, along with three main parameters of False Negative, False Positive, and Identity switch (IDsw), are among the common criteria for evaluating the quality of the MOT. This study utilized the stated parameters to evaluate the tracking quality.

Multi-Objects Tracking Precision (MOTP) is stated in Equation (13). This criterion represents the overall object position error for the "object- prediction" pair in all frames. It shows the tracker's ability to estimate the precise object position [41]. Another criteria, Multi-Objects Tracking Accuracy (MOTA) is calculated according to Equation (14). In the this Equation, FN, FP, and ID_{sw} are the missing objects, false positives, and identity switches at t, respectively [41]. To evaluate the correctness of any tracker at least three entities are needed to be defined: the tracker output (or hypothesis) H_i, which is the result of the tracking algorithm, the correct result, or ground truth, g_t, and a distance function $d_{i,t}$ that measures the similarity between the true object and the prediction. C_t, is the number of matches at time t and *bbox* stands for bounding box.

$$MOTP = \frac{\sum_{i,t} d_{i,t}}{\sum_t C_t}$$
$$d(H_i, g_t) = \frac{bbox(H_i) \cap bbox(g_t)}{bbox(H_i) \cup bbox(g_t)} \tag{13}$$

$$MOTA = 1 - \frac{\sum_t (FN+FP+ID_{sw})}{\sum_t g_t},$$
$$\overline{FN} = \frac{\sum_t FN}{\sum_t g_t}, \quad \overline{FP} = \frac{\sum_t FP}{\sum_t g_t}, \quad \overline{ID_{sw}} = \frac{\sum_t ID_{sw}}{\sum_t g_t} \tag{14}$$

$\overline{FN}$ is the rate of missing objects calculated among all objects and in all frames. $\overline{FP}$ is the rate of false positives, and $\overline{ID_{sw}}$ is the rate of occurred mismatches [41].

4. 2. Simulation Details This subsection gives an overview of the results of MOT simulations using the Multi-camera pedestrian video "EPFL" dataset. During the MOT, an ID is assigned to each detected object. However, some objects may leave the scene while being tracked, or their lifespan may be less than the threshold level, or be occluded. According to the tracking process the object matching is improved based on the combination of similarity distance metrics, EMD, Hausdorff and Chi-square, each of which calculates histogram similarity distance, images distance and the object's Zernike Moments magnitudes, respectively.

Figure 3, illustrates the moments of the first object while ZM changes due to the occlusion and intersection after the 750th frame. Furthermore, an occlusion can be predicted based on the sudden changes in the magnitude of the Zernike Moments. Multiple objects in the scene may merge and be detected as a single object due to lighting, occlusion, merging of shades, or the intersection of individuals' limbs, Figure 4. Thus, the *R* ratio will be

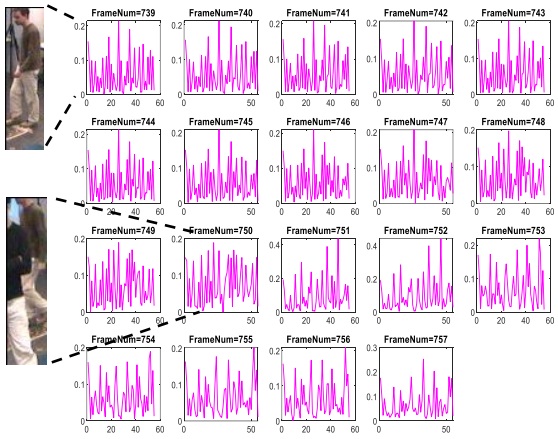

Figure 3. The illustration of the changes in Zernike Moments up to $n = 10$ order, before and after occlusion

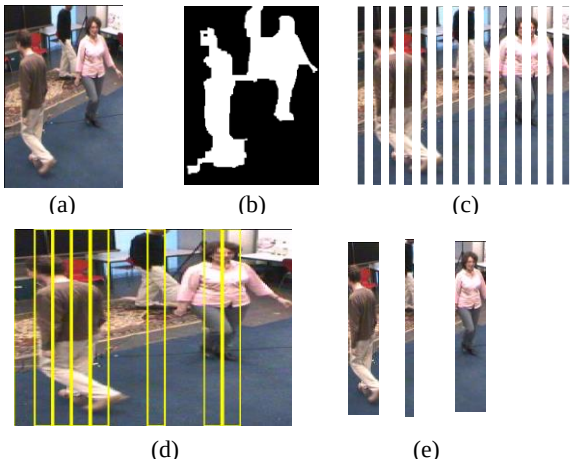

Figure 4. The separation of merged objects in the pedestrian video "EPFL" dataset, frame 1091st; (a) Three merged objects, (b) The silhouette of the merged objects, (c) Columnar segmentation, (d) Separation of merged objects, (e) Representation of the objects

[1] https://www.epfl.ch/labs/cvlab/data/data-pom-index-php/

calculated with respect to Equation (6). If the value of *R* is outside of the set thresholds, it is inferred that several objects are merged and must be separated. Consequently, the merged objects will be segmented into columnar patches, and the Zernike Moments of each patch will be calculated and compared using the Chi-Square distance. The patches with the least differences in moments magnitudes will be merged to form an object.

4. 3. Simulation Results Various studies have been performed object matching based on the SURF features [42, 43], Harris corner [44, 45], and Hungarian method [46, 47]. In Hungarian method, the objects are matched based on their optimal distance from each other. The Hungarian method tries to minimize the local distance between the target and the available predictions in each repetition. In this algorithm, each measurement will be assigned to an object by repeatedly scrolling the list of objects and allocating the closest measurement to each object. The measurement is then invalidated, and the next measuerments are processed. In SURF and Harris corner methods the objects are matched based on their key points. The SURF feature detector applies an approximate Gaussian second derivative mask to an image at many scales. Since the feature detector applies masks along each axis and at 45 degrees to the axis it is more robust to rotation than the Harris corner. The method is very fast because of the use of an integral image where the value of a pixel *(x,y)* is the sum of all values in the rectangle defined by the origin at *(x,y)*.

To detect features, the Hessian matrix, Equation (15), is applied.

$$H = \begin{bmatrix} L_{xx} & L_{xy} \\ L_{xy} & L_{yy} \end{bmatrix} \quad (15)$$

where $L_{xx}, L_{xy}, L_{yx} and\ L_{yy}$ is the convolution of the second derivative of a Gaussian with the image at the point. The Hessian determinate values are used for the range of detector windows. Valid features are found as a local maximum over a 3x3x3 range where the third dimension is detector window size, so a feature must be locally unique over a spatial range and a range of scales. The SURF uses a fast search algorithm to do non-maximum suppression.

The Harris corner detector takes horizontal and vertical derivatives of the image and looks for areas where both are high, this is quantified by the Harris corner descriptor which is defined as the matrix and descriptor, in Equations (16)-(17), respectively.

$$H = \begin{bmatrix} {D_x}^2 & D_x D_y \\ D_x D_y & {D_y}^2 \end{bmatrix} \quad (16)$$

$$c = \frac{\det(H)}{trace(H)} \quad (17)$$

We define a feature as a point that is a local maximum on a 3x3 area and is above a threshold. Also the results of the proposed method are compared with a tracking method that is implemented based on a kind of deep neural network called a Convolutional Neural Networks (CNN) [48]. This framework uses CNN to detect objects within the input frames. A state-of-the-art object detection framework [33], Faster R-CNN [49], is used for the detection of objects. The features used for the tracking are derived from a SURF and serve as a strong basis for object recognition. The approach is to match the extracted features of individual detections in subsequent frames, hence creating a correspondence of detections across multiple frames.

The images (of $227 \times 227 \times 3$ size) are applied as input to the detector model which detects and localizes individual objects [50]. According to the transfer learning, at first, a pre-trained AlexNet network is trained by the CIFAR10 dataset to regularize the network's weights and biases. Once again it is trained based on MOT: ETH-Bahnhof, ETH-Bahnhof, ETH-Linthescher datasets through the Faster R-CNN to detect humans. The developed algorithm uses SURF as a feature to make correspondences of the detections across the frames, and IDs are allocated to individual detections.

Since the results of the mentioned studies are not available in the Multi-Camera pedestrian video "EPFL" dataset, the MOT simulation and the object matching is performed based on SURF features, Harris corner, Hungarian, and Faster R-CNN methods regarding the flow diagram in Figure 5. The results of the performed tracking are presented in Table 1. The proposed method is based on Zernike Moments feature on the frame sequences from 1 to 2000, in the "EPFL" dataset.

5. DISCUSSION

Table 1 shows that tracking and matching objects with the proposed method yielded an accuracy (MOTA) of 81.6%, while the accuracy of the Hungarian, SURF, Harris corner, and Faster R-CNN methods are 74, 71.8, 50.1 and 78.6% respectively. Also, the percision of the proposed method, Faster R-CNN, Hungarian, SURF, and Harris corner methods are 54.5, 65.7 ,69.3, 72.6 and 78.1% respectively. Higher value of MOTP signifies low accuracy of the bounding boxes around the object. Higher values of MOTA signifies high accuracy in tracking. Based on the results, the Hungarian method minimizes the total distance between each object-prediction. So the false negative is less than the matching methods that operate based on points features. Also it is observed that tracking and matching processes based on the SURF or Harris corner have given poorer results compared to the rest.

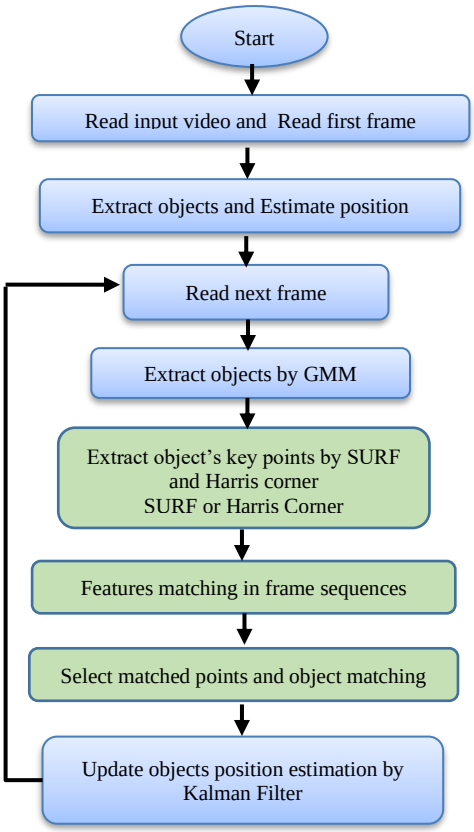

Figure 5. The flow diagram of object tracking and object matching based on SURF and Harris corner

TABLE 1. Evaluation of MOT based on data association simulated methods

MOTA	ID_{sw}	FP	FN	Method
74.4	29.8	11.3	52	Hungarian [48]
71.8	21.9	16	55.1	SURF [44]
50.1	14.6	34	97	Harris corner [45]
78.6	17.8	9.2	42.6	Faster R-CNN [51]
81.6	18	4	34	Proposed Method

The aforesaid is because the changes such as the horizontal movement of the objects, movement of limbs and deformations of object clothing will make the features points/corners extracted in frame $t-1$ disappear in frame t. Thus, the object feature points detected in frame $t-1$ are not detected and matched in frame t, so leading to a false negative error and false positives. In contrast, since the points features are used for data association, the IDsw and wrong identity allocations are less than other methods. The results in Table 1 shows the more accurate performance of the proposed method, even when comparing with a deep learning based approach.

In the proposed method, the recognition is performed based on the Zernike Moments of $n = 10$ order, and the magnitudes are obtained for Z_{00} to Z_{55}. The magnitude of the Zernike Moments indicates the overal shape of the object at low orders and object details at higher orders. Since the object feature extraction is not based on local distance or the points features, the false negatives and IDsw errors are less, resulting in higher accuracy. On the other hand, the false negative error is reduced, and fewer objects are lost in the tracking process due to the separation of merged objects. Finally, the tracking path of the first object is illustrated in Figure 6, in comparison with other methods and the groundtruth which shows the superiority of the proposed method.

It should be noted that the Epfl dataset is very famous that has different modes of challenges that happened while targets are moving. On the other hand the ground truth file, the coordinates of the objects in the scene, their path, and the homography matrix of each view is presented in a file. Hence, the proposed method has been evaluated on this dataset.

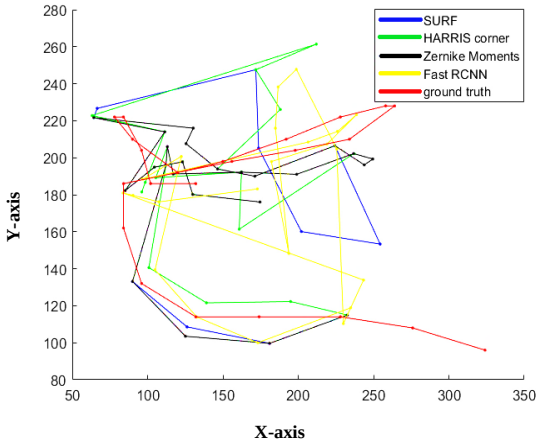

Figure 6. The trajectory of the first object in the 1st to 900^{th} frames; (blue) SURF, (green) Harris corner, (black) Zernike Moments, (yellow) Faster R-CNN and (red) Ground truth

6. CONCLUSION

This study attempted to improve the matchimg and identity association of objects according to the feature-based data association. To this end we designed a method which used Zernike Moments and similarity-based hard voting for data association and objects matching, respectively. We evaluated different criteria in various aspects between detected objects in consecutive frames. In this regard, Hausdorff and EMD distance criteria were used for distance metrics rather than Euclidean distance. Furthermore, separation of the merged objects in this work, leads to reduction of false negatives and can tackle with the dense distribution and mutual occlusion of individuals in the tracking process.

The results of this study can be used to track multiple objects using multiple cameras and detect the desired targets in the future works.

7. ACKNOWLEDGEMENT

The authors acknowledge the funding support of Babol Noshirvani University of Technology through the grant program No. BNUT/370123/00.

8. REFERENCES

1. Asvadi, A., Mahdavinataj, H., Karami, M. and Baleghi, Y., "Incremental discriminative color object tracking", in Artificial Intelligence and Signal Processing, Cham, Springer International Publishing. (2014), 71-81.
2. Ardeshir, G. and Khakpour, F., "Using a novel concept of potential pixel energy for object tracking", ***International Journal of Engineering***, Vol. 27, No. 7, (2014), 1023-1032.
3. Abbass, M.Y., Kwon, K.-C., Kim, N., Abdelwahab, S.A., El-Samie, F.E.A. and Khalaf, A.A.M., "A survey on online learning for visual tracking", ***The Visual Computer***, (2020), doi: 10.1007/s00371-020-01848-y.
4. Manafifard, M., Ebadi, H. and Abrishami Moghaddam, H., "A survey on player tracking in soccer videos", ***Computer Vision and Image Understanding***, Vol. 159, (2017), 19-46, doi: 10.1016/j.cviu.2017.02.002.
5. Asvadi, A., Karami, M. and Baleghi, Y., "Efficient object tracking using optimized k-means segmentation and radial basis function neural networks", ***Itrcjrn***, Vol. 4, No. 1, (2012), 29-39.
6. Asvadi, A. and Karami-Mollaie, M., "Object tracking using adaptive object color modeling. (2013).
7. Luo, W., Xing, J., Milan, A., Zhang, X., Liu, W. and Kim, T.-K., "Multiple object tracking: A literature review", ***Artificial Intelligence***, Vol. 293, (2021), 103448, doi: https://doi.org/10.1016/j.artint.2020.103448.
8. Xing, J., Ai, H., Liu, L. and Lao, S., "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, (2011), 1652-1667, doi: 10.1109/TIP.2010.2102045.
9. Asvadi, A., Karami-Mollaie, M., Baleghi, Y. and Seyyedi-Andi, H., "Improved object tracking using radial basis function neural networks", in 2011 7th Iranian Conference on Machine Vision and Image Processing. (2011), 1-5, doi: 10.1109/IranianMVIP.2011.6121604.
10. Khare, M., Srivastava, R.K. and Khare, A., "Object tracking using combination of daubechies complex wavelet transform and zernike moment", ***Multimedia Tools and Applications***, Vol. 76, No. 1, (2017), 1247-1290, doi: 10.1007/s11042-015-3068-5.
11. Fadaei, S. and Rashno, A., "Content-based image retrieval speedup based on optimized combination of wavelet and zernike features using particle swarm optimization algorithm", ***International Journal of Engineering***, Vol. 33, No. 5, (2020), 1000-1009, doi: 10.5829/ije.2020.33.05b.34.
12. Di Gesù, V. and Starovoitov, V., "Distance-based functions for image comparison", ***Pattern Recognition Letters***, Vol. 20, No. 2, (1999), 207-214, doi: 10.1016/S0167-8655(98)00115-9.
13. Taha, A.A. and Hanbury, A., "An efficient algorithm for calculating the exact hausdorff distance", ***IEEE Transactions on Pattern Analysis and Machine Intelligence***, Vol. 37, No. 11, (2015), 2153-2163, doi: 10.1109/TPAMI.2015.2408351.
14. Rubner, Y., Tomasi, C. and Guibas, L.J., "The earth mover's distance as a metric for image retrieval", ***International Journal of Computer Vision***, Vol. 40, No. 2, (2000), 99-121, doi: 10.1023/A:1026543900054.
15. Asvadi, A., Mahdavinataj, H., Karami, M.R. and Baleghi, Y., "Online visual object tracking using incremental discriminative color learning", ***The Csi Journal On Computer Science and Engineering***, Vol. 12, No. 2-4 (B), (2014), 16-28.
16. Dendorfer, P., Ošep, A., Milan, A., Schindler, K., Cremers, D., Reid, I., Roth, S. and Leal-Taixé, L., "Motchallenge: A benchmark for single-camera multiple target tracking", arXiv preprint arXiv:2010.07548, (2020).
17. Ciaparrone, G., Sánchez, F.L., Tabik, S., Troiano, L., Tagliaferri, R. and Herrera, F., "Deep learning in video multi-object tracking: A survey", ***Neurocomputing***, Vol. 381, (2020), 61-88.
18. Li, X., Wang, K., Wang, W. and Li, Y., "A multiple object tracking method using kalman filter", in The 2010 IEEE International Conference on Information and Automation. (2010), 1862-1866, doi: 10.1109/ICINFA.2010.5512258.
19. Xi, Z., Xu, D., Song, W. and Zheng, Y., "A* algorithm with dynamic weights for multiple object tracking in video sequence", ***Optik***, Vol. 126, No. 20, (2015), 2500-2507, doi: https://doi.org/10.1016/j.ijleo.2015.06.020.
20. Wu, Z., Thangali, A., Sclaroff, S. and Betke, M., "Coupling detection and data association for multiple object tracking, (2012), 1948-1955, doi: 10.1109/CVPR.2012.6247896.
21. Arun Kumar, N.P., Laxmanan, R., Ram Kumar, S., Srinidh, V. and Ramanathan, R., "Performance study of multi-target tracking using kalman filter and hungarian algorithm", in Security in Computing and Communications, Singapore, Springer Singapore. (2021), 213-227.
22. Li, H., Liu, Y., Lin, W., Xu, L. and Wang, J., "Data association methods via video signal processing in imperfect tracking scenarios: A review and evaluation", ***Mathematical Problems in Engineering***, Vol. 2020, (2020), 1-26.
23. Habtemariam, B.K., Tharmarasa, R., Kirubarajan, T., Grimmett, D. and Wakayama, C., "Multiple detection probabilistic data association filter for multistatic target tracking", in 14th International Conference on Information Fusion, (2011), 1-6.
24. Motro, M. and Ghosh, J., "Scaling data association for hypothesis-oriented mht", in 2019 22th International Conference on Information Fusion (FUSION), IEEE. (2019), 1-8.
25. Kim, C., Li, F., Ciptadi, A. and Rehg, J.M., "Multiple hypothesis tracking revisited", in 2015 IEEE International Conference on Computer Vision (ICCV). (2015), 4696-4704, doi: 10.1109/ICCV.2015.533.
26. Songhwai, O., Russell, S. and Sastry, S., "Markov chain monte carlo data association for general multiple-target tracking problems", in 2004 43rd IEEE Conference on Decision and Control (CDC) (IEEE Cat. No.04CH37601). (2004), 735-742 Vol.731, doi: 10.1109/CDC.2004.1428740.
27. Zhao, Z.-Q., Zheng, P., Xu, S.-t. and Wu, X., "Object detection with deep learning: A review", ***IEEE Transactions on Neural Networks and Learning Systems***, Vol. 30, No. 11, (2019), 3212-3232.
28. Bansal, M., Kumar, M. and Kumar, M., "2d object recognition techniques: State-of-the-art work", ***Archives of Computational Methods in Engineering***, (2020), doi: 10.1007/s11831-020-09409-1.
29. Ma, J., Jiang, X., Fan, A., Jiang, J. and Yan, J., "Image matching from handcrafted to deep features: A survey", ***International Journal of Computer Vision***, Vol. 129, No. 1, (2021), 23-79.
30. Lowe, D.G., "Distinctive image features from scale-invariant keypoints", ***International Journal of Computer Vision***, Vol. 60,

No. 2, (2004), 91-110.

31. Mikolajczyk, K. and Schmid, C., "Scale & affine invariant interest point detectors", ***International Journal of Computer Vision***, Vol. 60, No. 1, (2004), 63-86.
32. Bay, H., Tuytelaars, T. and Van Gool, L., "Surf: Speeded up robust features", in European conference on computer vision, Springer. (2006), 404-417.
33. Kaushal, M., Khehra, B.S. and Sharma, A., "Soft computing based object detection and tracking approaches: State-of-the-art survey", ***Applied Soft Computing***, Vol. 70, (2018), 423-464.
34. Uçar, A., Demir, Y. and Güzeliş, C., "Object recognition and detection with deep learning for autonomous driving applications", ***Simulation***, Vol. 93, No. 9, (2017), 759-769.
35. Zhou, X., Gong, W., Fu, W. and Du, F., "Application of deep learning in object detection", in 2017 IEEE/ACIS 16th International Conference on Computer and Information Science (ICIS), IEEE. (2017), 631-634.
36. Cheng, S., Luo, X. and Bhandarkar, S.M., "A multiscale parametric background model for stationary foreground object detection", in 2007 IEEE Workshop on Motion and Video Computing (WMVC'07). (2007), 18-18, doi: 10.1109/WMVC.2007.1.
37. Binh, N., "Human object tracking in nonsubsampled contourlet domain", ***International Journal of Advanced Computer Science and Applications***, Vol. 7, (2016).
38. Khare, M., Binh, N.T. and Srivastava, R.K., Human object classification using dual tree complex wavelet transform and zernike moment, in Transactions on large-scale data-and knowledge-centered systems xvi. 2014, Springer.87-101.
39. Górniak, A. and Skubalska-Rafajłowicz, E., "Object classification using sequences of zernike moments", in IFIP International Conference on Computer Information Systems and Industrial Management, Springer. (2017), 99-109.
40. Farahi, F. and Yazdi, H.S., "Probabilistic kalman filter for moving object tracking", ***Signal Processing: Image Communication***, Vol. 82, (2020), 115751, doi: 10.1016/j.image.2019.115751.
41. Bernardin, K. and Stiefelhagen, R., "Evaluating multiple object tracking performance: The clear mot metrics", ***EURASIP Journal on Image and Video Processing***, Vol. 2008, No. 1, (2008), 246309, doi: 10.1155/2008/246309.
42. Lu, X., Izumi, T., Teng, L. and Wang, L., "Particle filter vehicle tracking based on surf feature matching", ***IEEJ Journal of Industry ApplicationS***, Vol. 3, (2014), 182-191.
43. Wei, H., Takayoshi, Y., Hongtao, L. and Shihong, L., "Surf tracking", in 2009 IEEE 12th International Conference on Computer Vision. (2009), 1586-1592, doi: 10.1109/ICCV.2009.5459360.
44. Qi, Z., Ting, R., Husheng, F. and Jinlin, Z., "Particle filter object tracking based on harris-sift feature matching", ***Procedia Engineering***, Vol. 29, (2012), 924-929, doi: 10.1016/j.proeng.2012.01.065.
45. Salmane, H., Ruichek, Y. and Khoudour, L., "Object tracking using harris corner points based optical flow propagation and kalman filter", in 2011 14th International IEEE Conference on Intelligent Transportation Systems (ITSC). (2011), 67-73, doi: 10.1109/ITSC.2011.6083031.
46. Soleh, M., Jati, G. and Hilman, M., "Multi object detection and tracking using optical flow density-hungarian kalman filter (ofd-hkf) algorithm for vehicle counting. (2018).
47. Sahbani, B. and Adiprawita, W., "Kalman filter and iterative-hungarian algorithm implementation for low complexity point tracking as part of fast multiple object tracking system", in 2016 6th International Conference on System Engineering and Technology (ICSET). (2016), 109-115, doi: 10.1109/ICSEngT.2016.7849633.
48. Kim, B., Yuvaraj, N., Sri Preethaa, K., Santhosh, R. and Sabari, A., "Enhanced pedestrian detection using optimized deep convolution neural network for smart building surveillance", ***Soft Computing***, Vol. 24, (2020), 17081-17092.
49. Ren, S., He, K., Girshick, R. and Sun, J., "Faster r-cnn: Towards real-time object detection with region proposal networks", ***IEEE Transactions on Pattern Analysis and Machine Intelligence***, Vol. 39, No. 6, (2016), 1137-1149.
50. Acharya, D., Khoshelham, K. and Winter, S., "Real-time detection and tracking of pedestrians in cctv images using a deep convolutional neural network", in Proceedings of the 4th annual conference of research@ locate. Vol. 1913, (2017), 31-36.

Persian Abstract

چکیده

در نظارت ویدئویی ، رهگیری چند هدف به دلیل مشکل تطابق اهداف در فریم های متوالی ، یک مساله چالش برانگیز است. مقاله حاضر با هدف ارائه یک روش تطبیق بهبود یافته در ردیابی چند هدف بر اساس گشتاور زرنیک و ترکیبی از معیارهای فاصله شباهت ارائه شده است. در این تحقیق، ابتدا اهداف با استفاده از روش تفریق پس زمینه آشکارسازی می‌شود، همچنین مدل گاوسی مخلوط، برای استخراج اهداف در فریم های بعدی اعمال می شود. پس از آن، هیستوگرام رنگ و اندازه گشتاورهای زرنیک هر یک از اهداف محاسبه می شوند. در مرحله بعد، اهداف براساس فاصله هاسدورف ، فاصله Earth Mover بین هیستوگرام های رنگ و فاصله Chi-square بین گشتاورهای زرنیک در فریم حاضر و فریم های قبلی تطبیق داده می شوند. سپس، مکانیزم رأی گیری برای یافتن بهترین تطابق با معیارهای فوق طراحی شده است. درنهایت ، مکان هر یک از اشیا توسط فیلتر کالمن پیش بینی می‌شود تا رهگیری در فریم های بعدی ادامه یابد. نتایج نشان می دهد که رهگیری اهداف و عملکرد تطبیق با استفاده از روش پیشنهادی در توالی فریم های ویدئویی بر روی مجموعه داده ویدیویی چند دوربین عابر پیاده سایت"EPFL" ، بهبود یافته است. به طور خاص، خطاهای مثبت و منفی کاذب در رهگیری روش پیشنهادی کاهش یافته اند.