Volume 5, Issue 9, September 2015 ISSN: 2277 128X

International Journal of Advanced Research in Computer Science and Software Engineering

Research Paper

Available online at: www.ijarcsse.com

Optimization of Multi-Target Tracking and Occlusion Handling Using Mean Shift Method

Michael Kamaraj
Research and Development Centre,
Bharathiar University, Coimbatore,
Tamil Nadu, India

Balakrishnan
Department of Computer Science Engineering,
Indra Ganesan College of Engineering, Trichy,
Tamil Nadu, India

Abstract- Multi target tracking is an interesting and challenging task in finding an optimal set of path within a temporal window. The problem of multi target tracking comprises of few distinct challenges, the naturally discrete problem of data association, and continuous problem of trajectory estimation. Many recent approaches often perform multi-target tracking as discrete optimization which need a pre-computation and time. Alternatively, a framework is designed to focus on complete representation of the problem. In this work, an energy term is formulated as minimization of continuous energy in multiple target tracking. The energy function includes the dynamic model of target, mutual exclusion, track persistence and regularization. In addition, the occlusion and ambiguous targets of appearances are handled using the mean shift clustering which is largely invariant to both partial and full occlusions, complex backgrounds and change in scale. The mean shift introduces a feature space analysis for finding local maxima and minima of a density function from given discrete data samples. This paper also includes the quantitative evaluation of the experimental result on several challenging data sets on tracking to validate the proposed framework.

Keywords—Multi-object tracking, Mean shift clustering, tracking-by-detection, visual surveillance, continuous optimization,

I. INTRODUCTION

Object tracking is of vast significance in computer vision and is used in many applications such as video surveillance, computer vision, pattern recognition, multimedia and intelligent transport systems. Although numerous tracking methods were proposed in literature, now a days, still it has been a very challenging task to keep track of multiple targets in a video while preserving their identities. Nevertheless, the current algorithms provide reasonable performance only in comparably easy conditions with few targets. Several modern approaches to tracking follow a tracking-by-detection strategy, where the targets are detected in a pre-processing stage, usually either by background subtraction or using a discriminative classifier, from which the trajectories are later estimated.

Some recent multi-target tracking formulations aim to obtain a (nearly) globally optimal set of trajectories within a temporal window [1-7]. In order to make (near) global optimization possible and efficient, the state space is reduced by restricting the possible target locations to a finite set and the energy function is simplified. While global optimality undoubtedly has many benefits, we must also not lose sight of the actual purpose of formulating multi-target tracking as an energy minimization problem. The energy should adequately reflect the task at hand so that low energy solutions are close to the true situation. Unfortunately, in the realm of multi-target tracking typical specifications of the desirable aspects do not lead to models that can be globally optimized.

In contrast to previous work, we attempt to design the objective function such that it offers a more complete representation of the various aspects of the problem. Our energy is defined in continuous space. The energy depends on the locations and motions of all targets in all frames, including cases where image evidence is missing, and explicitly includes physical constraints, such as smoothness of motion and mutual exclusion. It is beneficial to model these terms in the continuous domain, since they describe the true situation more closely than ones that operate in a discrete setting. The price to pay is having to forgo global optimality, since such a complex model of multi-target tracking is unlikely to be convex. Nevertheless, local optima of our energy yield better results in practice, both visually and in terms of quantitative evaluation with respect to ground truth.

II. RELATED WORK

In recent times, tracking-by-detection methods associate multiple input detection responses in different frames to generate the trajectory of targets. Some researchers formulate the data association task as a matching problem, which matches the detections with similar appearance and motion patterns in consecutive frames. In this review it is thus concentrated on recent advances in visual multi-target tracking. Leibe et al. [8] couple the tasks of object detection and trajectory estimation through a quadratic binary program, which is then solved to local optimality by custom heuristics. Jiang et al. [9] cast the task of tracking multiple targets as an integer linear program (ILP) with linear constraints to enforce that the layout between targets does not change in adjacent frames. The solution is then obtained by LP-

relaxation, which cannot guarantee global optimality in general, but achieves it in most cases nonetheless. A network flow approach for global multi-target tracking was introduced by Zhang et al. [10]. Observation and transition edges between individual detections form a graph where their capacity represents the likelihood of target presence and motion. An optimal set of trajectories without occlusion handling is found by a min-cost flow algorithm. Head-based tracking in dense crowds is also employed by Rodriguez et al. [11], where the solution is obtained by minimizing a binary energy function with a constraining term to enforce the correct number of targets. While a high camera viewpoint can be assumed in many surveillance scenarios, this is generally not feasible in other applications (e.g., driver assistance or entertainment).

Xing et al. [12] generate short tracklets without occlusion reasoning and then connect tracklets to longer trajectories such that the connections can bridge gaps due to occlusions.

Wojek et al. [13] extend a full-body detector with six part detectors to enrich the space of target hypotheses. Each detection is then weighted by its expected visibility computed from a 3D scene model. Breitenstein et al. [14] increase the target likelihood if another target exists nearby. However, occlusion reasoning provides an accurate approximation to the actual fraction of the target visibility. Anton Milan et al.[15] presented a continuous energy minimization which includes occlusion reasoning and appearance model for global optimization of tracking multiple objects and the non convex energy is minimized by descent gradient method and a set of discontinuous jump moves. However the resulting tracking performance is not able to achieve higher recall percentage.

According to Y.Cheng et. al. [17] The Mean Shift method has been used in a number of computer vision problems, these include line fitting , image segmentation and object tracking . A number of improvements to the traditional formulation of the Mean Shift method for object tracking have been investigated by Comaniciu et. al. [18-19] . J. Wang et. al. [20] and A. Babaeian et. al. [21] had investigated on multiple features to gain a more descriptive representation of the target object, thereby various RGB colour spaces ,texture information and edge directions are used as descriptive features, feature localization weights are determined according to the similarity between background features and features present in the target model.

Collin et. al.[22] adopted Scale space theory in order to successfully determine the target object's scale during tracking. The Mean Shift method was applied to Gaussian kernels at various scales to determine the target object's scale. Image moments have been used with the similarity weights (between the target model and candidate) to determine the scale and orientation of the target object by J. Ning et. al [23]. Multiple ellipsoidal, asymmetric kernels with asymmetric centres have been used to effectively track target position, scale and orientation simultaneously was performed by Zhang et. al [24]. In order to remove background features from the target model and candidate model, a level set function has been used along the contour of the target object by Yilmaz et. al. [25]. The level set function defines an asymmetric kernel over the target region which does not contain any background features. Mean Shift is used to track the target object's position, scale and orientation. de Villiers et. al. [26] designed a Mean Shift framework which is largely invariant to both partial and full occlusions, complex backgrounds and change in scale. Multiple features are used to gain a descriptive representation of the target object. Eventhough, the occlusion are handled accurately, it cannot give the complete representation to the problem.

In this research work, it is aimed to find whether it is really beneficial for multi-object tracking to restrict the energy function in order to guarantee global optimality. To make the optimization efficient, all energy terms are formulated as functions that can be computed and differentiated in closed form. Hence, computationally efficient gradient-based optimization methods can be applied. Therefore, the Mean Shift method is a nonparametric, variable step-size, statistical density estimator which iteratively determines the nearest mode of a point sample distribution using gradient ascent whereas gradient descent is associated with minimization of energy. Multiple features are used to gain a more descriptive representation of the target object. An adaptive feature weighting method is used to maximize the better localization of target object. Image moments are used in conjunction with the similarity weight to determine the scale of target object and kalman filter is used to improve the tracking performance during occlusions.

Extensive experiments on various public datasets and show state-of-the-art results quantitatively measured by standard multi-target tracking metrics. This paper proposes a global occlusion model closely integrated into continuous tracking framework, and can easily handle a large number of targets. Moreover, it is able to accurately estimate pairwise visibility dependencies between all targets. This paper is arranged as follows, section III, IV and V describes the multi target tracking energy terms, Mean shift occlusion handling and the energy minimization respectively, section VI explains the quantitative evaluation of the experiment and section VII concludes the paper.

III. TRACKING OF MULTIPLE TARGET

A. Energy Terminologies

For better understanding, let us first introduce the general structure and notation used in this paper. State vector X consists of the (X, Y) co-ordinates of all N targets in a sequence of F frames. Here we assume that all targets move on a common ground plane. The location of each target i at frame t $\{s_i , ... , e_i \}$ is denoted by X_i^t. In our formulation the position of each target is defined even in the case of occlusion is present. The temporal length of trajectory is $F(i) = e_i - s_i + 1$. Where, s_i be the first frame and e_i be the final frame. Note that, the number of targets may get vary from frame to frame. Therefore the number of targets N be denoted by N (t). Also D(t) be the number of detections in frame t and D_g^t be the location of detection g in frame t.

There are many possibilities to define an energy function which rewards more plausible configurations and penalizes unreasonable ones. Our energy function is made up of six terms

$$E' = E'_{om} + iE'_{am} + jE'_{dm} + kE'_{me} + lE'_{tp} + mE'_{r} \qquad (1)$$

E'_{om} is the data term which keeps the solution close to the observation. E'_{am} captures the appearance of the different objects to disambiguate data association. E'_{dm}, E'_{me} and E'_{tp} promote plausible motion and enforce physical constraints. E'_{r} is the regularizer term keeps the solution simple and overfitting. From an optimization perspective, it would certainly be beneficial to have a complex function, which by definition only has a single minimum and can be globally optimized independent of initial values. This is achieved by finding every possible solution a cost and then finds a state with the lowest cost. Then the aim is to find the state E' that minimizes the high dimensional continuous energy from Eq. (1)

$$X' = arg \min_{X \in R^d} E'(X) \qquad (2)$$

Depending on the length of the sequence and the number of targets, the dimension of the search space d takes on values between 10^3 and 10^4. In the remainder of this section each individual term and its functionality are explained in more depth.

B. Observational model

Tracking by detection has proven to be a reliable one in the last few years and is applicable in unconstrained environments with a moving camera. Here, pedestrians with a sliding window approach using both HOG features and histograms of relative optic flow are detected. The energy should be minimized when the location of each target precisely matches detection. That means the data term is to keep the trajectories close to the observations. In order to get the localization uncertainty the energy smoothly increases with the distance between the estimated object location X_i^t and the detection location D_g^t. this characteristic is modelled as follows

$$E'_{om}(X) = \sum_{i=1}^{N} \sum_{t=s_i}^{e^i} [V_i^t . \varepsilon - \sum_{g=1}^{D(t)} \omega_g^t \frac{S_g^2}{\left\| X_i^t - D_g^t \right\|^2 + S_g^2}] \qquad (3)$$

Where, ω be the quantity that weights of detection, S be the scalar quantity that accounts for the object size. ε be the offset which is being uniformly added to all existing targets to penalize all those with no image evidence. This penalty is not applied if the target has occlusion. Therefore it is scaled by the fraction of visibility V_i^t as follows

C. Dynamic model

The representation of dynamic model of the system is as follows:

$$E'_{dm}(X) = \sum_{i=1}^{N} \sum_{t=s_i}^{e_i - 2} \left\| X_i^t - 2X_i^{t+1} + X_i^{t+2} \right\|^2 \qquad (4)$$

The dynamic model defines the objects move slowly relative to the object. The dynamic model can be interpreted as a kind of "Intelligent smoothing" which takes into account the other energy terms rather than blindly soothes the nodes of the trajectory curves. The dynamic model beyond smoothing helps to prevent identity switches between crossing targets.

D. Mutual Exclusion

A further aspect of multi target tracking is collision avoidance. The most natural way to formulate collision avoidance as a pairwise term- if two putative target locations are close to each other, add a high penalty unless at least one of them is labelled as an outlier. We incorporate collision handling by putting penalty when two targets come too close to each other.

$$E'_{me}(X) = \sum_{i=1}^{F} \sum_{j \neq i}^{N(t)} \frac{S}{\left\| X_i^t - X_j^t \right\|^2} \qquad (5)$$

The scale factor s is set to 35cm for people tracking and this value goes to infinity when they share one identical position. This formulation of collision avoidance takes into account the actual overlap of target volumes and can correctly handle two notoriously difficult problems of multi target tracking. On the one hand, overlap between targets is checked at all times, even if both targets are occluded or otherwise missed by the detector. On the other hand, if two targets would collide due to inaccurate observations, the continuous optimization can push them apart just as much as needed, whereas methods based on grid discretization or non-maximum suppression can only "connect the dots" and would have to discard an entire trajectory.

E. Persistence

Object disappearance will lead to abrupt track termination in the middle of the tracking area. In order to push the trajectories to start and end along image borders or along a predefined perimeter, tracks that do not obey this requirement is penalized. The sigmoid penalty is as follows.

$$E'_{tp}(X) = \sum_{\substack{i=1,\dots N \\ t \in \{s_i, e_i\}}} \frac{1}{1 + e(-q.b(X_i^t) + 1)} \qquad (6)$$

where $b(x_i^{st})$ and $b(x_j^{et})$ are distances of the start, respectively end points or trajectory i to the border of the frame.

F. Regularization

A regularaizer is needed to prevent the number of targets from grouping arbitrarily large so as to better fit the data

$$E'_r(X) = N + \sum_{i=1}^{N} \frac{1}{F(i)} \qquad (7)$$

where, F(i) be the temporal length of trajectory i in frame

IV. MEAN SHIFT OCCLUSION REASONING

A. Object Representation

An object is typically defined by an ellipsoidal region or patch surrounding a region of interest in an image. A feature space is chosen (typically the RGB feature space is used) to determine a histogram of the pixel distribution in the target region. The histogram is represented by target model q. The target model is used to describe the appearance of the object located in the target region. The target model q is comprised of m normalized bins .

$$\vec{q} = \{\vec{q}_u\}_{u=1\ldots m} \qquad (8)$$

$$\sum_{u=1}^{m} \vec{q}_u = 1 \qquad (9)$$

$\{x_i^*\}_{i=1\ldots n}$ denotes the n normalized pixel locations in the target region which are centred around zero. Let k(x) denote a convex, monotonically decreasing, isotropic kernel. Let b: $R^2 \rightarrow \{1 \ldots m\}$ be a function which determines the histogram bin $b(x_i^*)$ associated with the pixel location x_i^* . The probability of the feature u = 1 ...m in the target models histogram is determined by

$$\overrightarrow{q_u} = C \sum_{i=1}^{n} K\|x_i^*\|^2 \, \delta[b(x_i^*) - u] \qquad (10)$$

where δ is the Kronecker delta function. The normalization constant C is derived by imposing the condition (9), normalization constant C can therefore be represented by

$$C = \frac{1}{\sum_{i=1}^{n} K(\|x_i^*\|^2)} \qquad (11)$$

B. Candidate Representation

Typically the target model is formed from the target region in the first frame of a video sequence. The target model is compared to candidate regions in the current frame to determine the location and scale of the target in the current frame. A target candidate p(y) is defined by a histogram of the pixel distribution of a region in the current frame. The target candidate p(y) is comprised of m normalized bins .

$$\vec{p}\,(y) = \{\vec{p}_u(y)\}_{u=1\ldots m} \qquad (12)$$

$$\sum_{u=1}^{m} \vec{p}_u\,(y) = 1 \qquad (13)$$

Let $\{x_i^*\}_{i=1\ldots n_h}$ denotes the n_h normalized pixel locations in the candidate region which are centred around y. Let k(x) denote the same convex, monotonically decreasing, isotropic kernel used with the target model only with a different size (based on the scale of the target object) specified by bandwidth h. The probability of the feature u = 1....m in the target candidates histogram is determined by

$$\overrightarrow{p_u}(y) = C_h \sum_{i=1}^{n} K\left(\left\|\frac{y - x_i}{h}\right\|^2\right) \delta[b(x_i^*) - u] \qquad (14)$$

$$C_h = \frac{1}{\sum_{i=1}^{n} K\left(\left\|\frac{y - x_i}{h}\right\|^2\right)} \qquad (15)$$

C. Similarity Model

In order to determine the similarity between the target model and the target candidate a similarity function is determined. The similarity function used is the sample estimate of the Bhattacharyya coefficient [11] between the distributions $\vec{q}$ and $\vec{p}\,(y)$. The similarity function is defined by

$$\overrightarrow{p}(y) = \rho[\vec{p}(y), \vec{q}] \sum_{i=1}^{n} \sqrt{\overrightarrow{p_u}(y)\overrightarrow{q_u}} \qquad (16)$$

Due to the conditions imposed by (9) and (14) the similarity function has a minimum value of 0 (distributions are orthogonal) and a maximum value of 1 (distributions are equal).

D. Mean Shift Vector

The Mean Shift algorithm iteratively samples target candidate locations in an effort to find the local maximum of the similarity function $\vec{p}\,(y)$. By taking the Taylor expansion around the target candidate probability values $\overrightarrow{p_u}\,(\vec{y}_0)$ (where

the target candidate $\vec{p}(y_0)$ is centred around $(\vec{y}_0)$ the estimated linear approximation of the Bhattacharyya coefficient [16] can be described by

$$\rho[\vec{p}(y), \vec{q}] = \frac{1}{2}\sum_{u=1}^{m}\sqrt{\overrightarrow{p_u(\vec{y}_0)}\overrightarrow{q_u}} + \frac{1}{2}\sum_{u=1}^{m}\overrightarrow{p_u}(y)\sqrt{\frac{\overrightarrow{q_u}}{\overrightarrow{p_{u(\vec{y}_0)}}}} \qquad (17)$$

The first term of (18) is independent of position y, therefore to maximize $\rho[\vec{p}(y), \vec{q}]$ it is necessary to maximize the second term of (18), using (15) the second term of (18) denoted by $\rho[\vec{p}(y), \vec{q}]_2$ can be described by

$$\rho[\vec{p}(y), \vec{q}]_2 = \frac{C_h}{2}\sum_{i=1}^{n_h}\omega_i K\left(\left\|\frac{y - x_i}{h}\right\|^2\right) \qquad (18)$$

$$\omega_i = \sum\sqrt{\frac{\overrightarrow{q_u}}{\overrightarrow{p_{u(\vec{y}_0)}}}}\delta[b(x_i^*) - u] \qquad (19)$$

The Mean Shift vector is determined in order to maximize the similarity function $\vec{p}(y)$ by maximizing (19). The Mean Shift vector is determined by

$$E'_{am}(X_1) = \frac{\sum_{i=1}^{n_h}(x_i - \vec{y}_0)\omega_i g\left(\left\|\frac{\vec{y}_0 - x_i}{h}\right\|^2\right)}{\sum_{i=1}^{n_h}\omega_i g\left(\left\|\frac{\vec{y}_0 - x_i}{h}\right\|^2\right)} \qquad (20)$$

where g(x) = k'(x). If we choose k(x) to use the Epanechnikov profile [29] described by

$$k(x) = \begin{cases}\frac{1}{2}c_d^{-1}(d+2)(1-x), & \text{if } x \le 0\\ 0 & \text{otherwise}\end{cases} \qquad (21)$$

the computation of (21) can be simplified as g(x) becomes a constant. Different kernel profiles may be used, they however have little impact on the localization accuracy of the Mean Shift algorithm. These kernel profiles have a higher computational cost as the kernel derivative g(x) must be determined for each computation of the Mean Shift vector. Using the Epanechnikov profile the Mean Shift vector can be described by

$$E'_{am}(X_1) = \frac{\sum_{i=1}^{n_h}(x_i - \vec{y}_o)\omega_i}{\sum_{i=1}^{n_h}\omega_i} \qquad (22)$$

The updated position of the target candidate position $\vec{y}_o$ is simply described by

$$\vec{y}_o = \vec{y}_0 + E'_{am}(X_1) \qquad (23)$$

The Mean Shift algorithm is run recursively until convergence, convergence occurs when the Mean Shift vector is lower than a tolerance $\in$. The tolerance is usually chosen to be the width of a single pixel. Take the inverse of equ (23) to get the gradient descent of energy minimizations.

$$\vec{y}_o = -\eta\,\vec{y}_o E'_{am}(Q_1) \qquad (24)$$

V. ENERGY MINIMIZATION

Each energy component is differentiable in closed form, making the entire formulation well suited for gradient-based minimization. It is chosen to apply a standard conjugate gradient descent to minimize Eq. (1) locally. However, given the highly non-convex nature of the energy, a purely gradient-based optimization would be very susceptible to initialization. Therefore, a set of jump moves are added. These non-local jumps in the energy landscape change trajectory lengths and potentially the number of targets, thus allowing a more flexible probing of the solution space to escape weak minima. Upon convergence of the gradient descent, one of six jump moves described below is executed in a greedy fashion. Then the gradient descent restarts.

1) Growing and Shrinking: each trajectory can be extended by linear extrapolation for an arbitrary number of time steps both forward and backward in time. Similarly, a track is shortened by discarding a fragment of a certain length from either end. Growing is useful for ending new targets, while shrinking weeds out false positives that may have been introduced by noise or during intermediate optimization steps.

2) Merging and Splitting: two existing trajectories are merged into one if the merge lowers the energy. It is noted that the individual energy components, in particular the dynamics and the exclusion terms, assert that this step will not cause physically implausible situations with intersecting trajectories or unlikely motion patterns. A single track may also be split into two at a specific point in time. Both these moves provide a method to bridge over regions with missing sensor responses and to reduce fragmentation of tracks and identity swaps.

3) Adding and Removing: These two moves operate on entire trajectories. Removing a false positive target from the current solution may decrease the overall energy because it results in a more plausible explanation of the data. On the other hand, it is important to allow for inserting new tracks around active sensor locations that do not have a target nearby. This is done conservatively by adding a short tracklet of only three frames. It is noted that it can grow and merge with other existing trajectories at a later optimization step.


```
Input : Q initial results, detections M
Output : Best of less than Q results

for i= 1 to Q
 {
  While (NOT CONVERGED)
   {
    for m in {GROW,SHRINK,ADD,REMOVE,MERGE,SPLIT}
    {
     for j in {1,..,N}
     { Try jump move m on trajectory j
      If E_new(Q_i) less than E_old(Q_i)
       {
        Perform jump move  i
       }
     }
     Perform conjugate gradient descent
    }
   }
 }
Return argmin_Q E(Q_i)
```

Algorithm 1. Greedy Energy Minimization

Applying algorithm 1 the optimum value of tracked locations is obtained and the energy term has the minimum values. After minimization, the trajectories obtained are smooth in nature, in contrast with the sharp edged trajectories obtained before optimization.

VI. EXPERIMENTAL STUDY

A. Implementation Detail

To facilitate the computation of the distance, the target area is defined as rectangular area on the ground. Targets outside its limits are excluded from the solution. The run time detection of the MATLAB 7.8/MEX implementation takes approximately 1s/frame to obtain solution using explicit occlusion reasoning and a simple inexpensive occlusion computation, the optimization runs an order of magnitude faster, achieving near real-time performance. Computing colour information for all pixels significantly slows down the optimization. This can be improved with the help of appearance term (mean shift). To speed up the convergence results the number of iterations is limited to the maximum of 15. During the experimentation the precise parameter values are included for accurate tracking performance and also it is highly dependent on the implementation. The weights for the parameters are set accordingly {i=0.1, j=.02, k=0.5, l=0.7, m=0.7} and λ =0.1.In order to improve the results for various sequences of the dataset ,the parameter terms from i to m can be modified slightly to achieve the best results.

The energy term represented in equ (1) is studied based on the output results of each individual weight of the energy function. The quantitative evaluation of the different tracking methods are measured using the following metrics of MOTP (Multiple Object Tracking Precision) It is total error in estimated position for matched object over all frames averaged by the total number of matches made. It shows the ability of the tracker to estimate precise object position and keep the consistent trajectories and so forth. The scores of the mostly lost (ML) and mostly tracked (MT) are used to evaluate the pose of the tracker output with respect to the groundtruth. MOTA (Multiple Object Tracking Accuracy) are derived from three error ratios. FP are the number of misses of false positives, FN are the mismatch of the object. ID are the total number of the identity switches.

Fig. 1. The sample video sequence of PETS09-S3-MF1

B. Experimental Evaluation

The validation of the proposed framework is applied on openly available video sequences the PETS09 S2L1, PETS09 S2L2. These are recorded outdoors from an elevated viewpoint, corresponding to a typical surveillance setup. The sequence of S2L1and S2L2 is having a frame rate of 7 fps and resolution of each frame is 768X576 is the most widely used in multi-target tracking. It also includes non-linear motion of target, closeness of the target and a sight of occlusions, the tracking accuracy of the results on these sequences is over 90%. The experimentation of the proposed methods is extended to the test data with more difficult video sequences with high crowd density from PETS09 of the

video sequence S2L2 and S2L3 respectively. The evaluations of the tracker are tested to its maximum level, therefore additionally two more complex scenarios from the selected from the PETS09 of the video sequence S1L1- 2 and S1L2-1, which were initially intended for person counting and density estimation, rather than for tracking persons. Finally, the ETH-Bahnhof video sequence shows people walking on a street where, the size of the pedestrians on the image plane varies significantly.

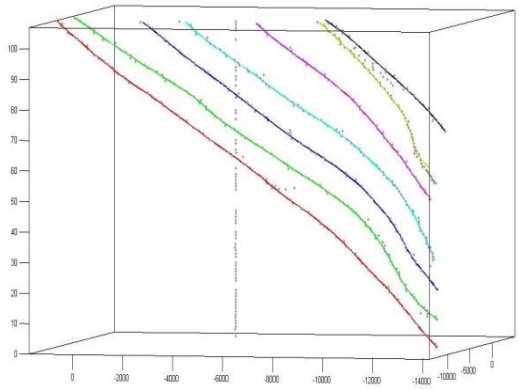

Fig. 2. The Trajectory of the PETS09-S3-MF1

TABLE I 2D & 3D EVALUATION OF THE VIDEO SEQUENCE PETS09-S3-MF1

Eval	Rcll	PRcn	FAR	G T	MT	P T	M L	F P	F N	I Ds	F M	MOTA	MOTP	MOTAL
2 D	98.2	99.8	0.01	7	7	0	0	1	9	1	0	97.9	75.4	98.1
3 D	98.2	99.8	0.01	7	7	0	0	1	9	1	0	97.9	86.2	98.1

The video sequence PETS09-S3-MF1 is thus completely studied. It consists of 107 frame runs at 7fps. Each frame is of size 768X576. The complete execution of these frames takes nearly 0.21 mins, each frame takes 0.12sec for the processing. The optimization is converged after 10 epochs. The fig (1) displays the sample output results of the proposed work. It is clear that the energy function yield a good accuracy in tracking multiple target taking occlusion into consideration and the influence of each energy terms and the optimization of convergence rate is measured. From the Table I it is observed that the evaluation results of the proposed algorithm produce a mean object tracking accuracy is 98.1% is achieved and the tracking precision is 75.2% on 2D evaluation which is improved by 10.8% increase during the 3D evaluation of the dataset. The number of frames a trajectory is grown are optimized independently for each trajectory as given in the fig (2), which is the trajectory moves of the all seven targets in the video sequence.

TABLE II QUANTITATIVE RESULTS OF THREE VIDEO SEQUENCE AND THE MEAN PERFORMANCE OF THE HUMAN ACTION DETECTOR.

Sequence	Method	MOTA	MOTP	GT	MT	ML	FP	FN	ID	FM	Rcll	Prcsn	Fa/F
PETS-S2L1	CEM	90.6	80.2	23	21	1	59	302	11	6	92.4	98.4	0.07
	KSF	80.3	72.0	23	17	2	126	641	13	22	83.8	96.3	0.16
	EKF	68.0	76.5	23	9	1	65	1173	25	30	70.3	97.7	0.08
	Our method	91.2	80.9	23	21	1	56	303	10	5	92.8	98.6	0.06
ETH-BAhnhof	CEM	71.1	65.5	9	7	0	92	108	4	3	84.7	86.7	0.51
	KSF	45.8	56.7	9	5	1	49	172	5	4	63.1	79.2	0.65
	EKF	58.2	58.3	9	3	0	115	172	2	6	75.1	81.9	0.65
	Our method	72.1	66.2	9	7	0	90	106	3	3	85.2	87.0	0.49
PETS-S3-MF1	CEM	96.7	82.7	7	7	0	5	12	0	0	97.7	99.0	0.05
	KSF	83.7	77.8	7	6	1	22	62	0	0	87.9	95.4	0.21
	EKF	66.7	81.9	7	2	0	0	169	0	1	66.7	100.0	0.00
	Our method	96.8	82.9	7	7	0	4	12	0	0	98.0	99.1	0.04
MEAN	CEM	86.1	76.1	13.0	11.7	0.3	52.0	140.7	5.0	3.0	91.6	94.7	0.2
	KSF	69.9	68.8	13.0	8.0	1.3	88.3	321.3	6.0	12.3	78.3	90.3	0.3
	EKF	64.3	72.2	13.0	4.7	0.3	60.0	504.7	9.0	12.3	70.7	93.2	0.2
	Our method	86.7	76.7	13.0	11.6	0.3	50.0	140.3	4.3	2.7	92.0	94.9	0.2

TABLE III QUANTITATIVE RESULTS OF FOUR CHALLENGING VIDEO SEQUENCE AND THE MEAN PERFORMANCE OF THE HUMAN ACTION DETECTOR

Sequence	Method	MOTA	MOTP	GT	MT	ML	FP	FN	ID	FM	Rcll	Prcsn	Fa/F
PETS-S2L3	CEM	56.9	59.4	74	28	12	622	2881	99	73	65.5	89.8	1.43
	KSF	24.2	60.9	74	7	40	193	6117	22	38	26.8	92.1	0.44
	EKF	28.6	60.3	74	2	32	280	5565	74	116	32.9	90.7	0.64
	Our method	57.9	60.0	74	32	8	620	2678	100	70	66.2	89.9	1.40
PETS-S2L3	CEM	45.4	64.6	44	9	18	169	1572	38	27	51.8	90.9	0.70
	KSF	28.8	61.8	44	5	31	45	2269	7	12	30.4	95.7	0.19
	EKF	20.4	63.3	44	1	35	13	2543	8	33	21.1	98.1	0.05
	Our method	46.2	65.1	44	9	17	163	1576	35	26	52.0	91.0	0.69
PETS-S1L1-2	CEM	57.9	59.7	36	19	11	148	918	21	13	64.5	91.8	0.61
	KSF	51.5	64.8	36	16	14	98	1151	4	8	55.5	93.6	0.41
	EKF	34.6	63.2	36	3	17	10	1664	6	18	35.2	98.9	0.04
	Our method	58.4	61.3	36	18	10	146	910	21	12	65.1	92.0	0.59
PETS-S1L2-1	CEM	30.8	49.0	43	7	20	227	2308	61	35	38.5	86.4	1.13
	KSF	19.5	60.6	43	4	29	64	2950	7	11	21.4	92.6	0.32
	EKF	9.5	53.1	43	0	34	38	3326	28	46	11.3	91.8	0.19
	Our method	31.6	50.1	43	7	18	230	2301	60	34	39.2	87.1	0.91
MEAN	CEM	47.8	58.2	49.2	15.8	15.2	291.5	1919.8	54.8	37.0	55.1	89.7	1.0
	KSF	31.0	62.0	49.2	8.0	28.5	100.0	3121.8	10.0	17.2	33.5	93.5	0.3
	EKF	23.3	60.0	49.2	1.5	29.5	85.2	3274.5	29.0	53.2	25.1	94.9	0.2
	Our method	48.5	59.1	49.2	16.5	13.2	289.8	1866.2	54.0	35.5	55.6	90.0	0.9

The main objective of the energy function of multiple object tracking can be approached in two distinct aspects of examining the influence of the individual energy terms on the tracking performance and the robustness of the chosen parameters to variations of their respective value and compare the different optimization strategies and their influence on the convergence result. Table II and III gives the quantitative results for all metrics, computed on all seven sequences individually. The results of the proposed state-of-art method are compared to those of a other well known multiple object tracking methods such as k-shortest paths(KSP)[27]on a regular grid as well as to a well boosted particle filter(BSP)[28] and extended kalman filter(EKF) .Furthermore, the average performance across a number of video sequences is reported for the dataset containing less than 10 targets per frame and a more complex group where up to 42 pedestrians in a frame.

The overall tracking accuracy (MOTA) of the proposed method is computed explicitly taking occlusion into the consideration. Though, for the less dense group occlusion computation cannot show its complete effectiveness because the pedestrians are totally visible in most of the frames. Conversely, for the complex environment the accuracy increases by 2 % on average and over 5 % in most difficult case. Hence, the proposed method does better in performance than the discrete tracker on all sequences.

VII. CONCLUSION

Multi object tracking framework established an effective appearance model, while reasonably describe the target can better distinguish background. The proposed method is fast and efficient to meet the real-time requirement of correctly associated targets, tracks remarkably well even through occlusions, and difficult target interactions with missing detections, crowded scenes and long-term occlusions. A wide experimental evaluation on a number of complex pedestrian datasets prove that the proposed method leads to very competitive results both visually and in terms of quantitative evaluation with respect to groundtruth. Furthermore, the proposed method will be able to reach real-time performance with an even faster optimization based on multi-grid search, as well as a more efficient implementation. This would make the method applicable to real-time applications, by repeatedly solving for only the past frames.

REFERENCES

[1] M. Andriluka, S. Roth, and B. Schiele. *Monocular 3D pose estimation and tracking by detection.* In CVPR 2010.

[2] A. Andriyenko, S. Roth, and K. Schindler. *An analytical formulation of global occlusion reasoning for multi-target tracking*. In 11th International IEEE Workshop on Visual Surveillance, 2011.

[3] A. Andriyenko and K. Schindler. *Globally optimal multi-target tracking on a hexagonal lattice*. In ECCV 2010, vol. 1, pp. 466-479.

[4] A. Andriyenko and K. Schindler. *Multi-target tracking by continuous energy minimization*. In CVPR 2011.
[5] B. Benfold and I. Reid. Stable multi-target tracking in real-time surveillance video. In CVPR 2011.
[6] J. Berclaz, F. Fleuret, and P. Fua. Robust people tracking with global trajectory optimization. In CVPR 2006.
[7] J. Berclaz, F. Fleuret, and P. Fua. *Multiple object tracking using flow linear programming*. In Winter-PETS, Dec. 2009
[8] B. Leibe, K. Schindler, and L. Van Gool. *Coupled detection and trajectory estimation for multi-object tracking*. In ICCV 2007.
[9] H. Jiang, S. Fels, and J. J. Little. *A linear programming approach for multiple object tracking*. In CVPR 2007.
[10] L. Zhang, Y. Li, and R. Nevatia. *Global data association for multiobject tracking using network flows*. In CVPR 2008.
[11] M. Rodriguez, I. Laptev, J. Sivic, and J.-Y. Audibert. , *Density-aware person detection and tracking in crowds*. In ICCV 2011
[12] J. Xing, H. Ai, and S. Lao. , *Multi-object tracking through occlusions by local tracklets filtering and global tracklets association with detection responses*. In CVPR 2009.
[13] C. Wojek, S. Walk, S. Roth, and B. Schiele., Monocular 3D scene understanding with explicit occlusion reasoning. In CVPR 2011
[14] M. D. Breitenstein, F. Reichlin, B. Leibe, E. Koller-Meier, and L. Van Gool., *Robust tracking-by-detection using a detector confidence particle filter*. In ICCV 2009.
[15] A. Milan, S.Roth, and K.Schindler,, *Continuous Energy Minimization for Multi-Target Tracking*, IEEE Transactions on Pattern Analysis and Machine Intelligence, Vol. No. , March 2014.
[16] F. Aherne, N. Thacker, and P. Rockett, "The Bhattacharyya Metric as an Absolute Similarity Measure for Frequency Coded Data," in Kybernetika, vol. 34, no. 4, pp 363 - 368, 1998.
[17] Y. Cheng , *"Mean Shift, Mode Seeking, and Clustering*," in IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. 17, pp. 790 -799, 1995.
[18] D. Comaniciu and P. Meer, "*Mean Shift Analysis and Applications*," in International Conference on Computer Vision, vol. 2, pp. 1197 - 1203, 1999.
[19] D. Comaniciu, V. Ramesh, and P. Meer, *"Kernel-Based Object Tracking," in IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol.25, no. 5, pp. 564 - 577, May 2003
[20] J. Wang and Y. Yagi, *"Integrating Shape and Color Features for Adaptive Real-time Object Tracking*," in IEEE International Conference on Robotics and Biomimetics, pp. 1 - 6, 2006
[21] A. Babaeian, S. Rastegar, M. Bandarabadi and M. Rezaei, "*Mean Shift-Based Object Tracking with Multiple Features*," in 41st Southeastern Symposium on System Theory, pp. 68 - 72, March 2009
[22] R. T. Collins, *"Mean-shift blob tracking through scale space*," in IEEE Conference on Computer Vision and Pattern Recognition, pp. 234 - 240 2003
[23] J. Ning, L. Zhang1, D. Zhang and C. Wu, "*Scale and Orientation Adaptive Mean Shift Tracking*," in Computer Vision, IET, vol. 6, iss. 1, pp. 52 - 61, 2012
[24] S. Zhang and Y. Bar-Shalom, *"Robust Kernel-Based Object Tracking with Multiple Kernel Centers,"* in 12th International Conference on Information Fusion, pp. 1014 - 1021, July 2009
[25] A. Yilmaz, "*Object Tracking by Asymmetric Kernel Mean Shift with Automatic Scale and Orientation Selection,"* in IEEE Conference on Computer Vision and Pattern Recognition, pp. 1 - 6, 2007
[26] B.Z. de Villiers, W.A. Clarke, P.E. Robinson, "*Mean shift Object Tracking with Occlusion Handling"*, Available : **URI:** http://hdl.handle.net/10210/9889.
[27] J. Berclaz, F. Fleuret, E. Turetken, and P. Fua. "*Multiple object tracking using k-shortest paths optimization"*. IEEE Transaction on Pattern Anal. Mach. Intell., 33(9):1806–1819, Sept. 2011.
[28] K. Okuma, A. Taleghani, O. D. Freitas, J. J. Little, and D. G. Lowe. *"A boosted particle filter: Multitarget detection and tracking"*. In ECCV 2004, volume 1, pages 28–39.
[29] D. Comaniciu and P. Meer, *"Mean Shift: A Robust Approach Toward Feature Space Analysis,"* in IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. 24, no. 5, pp. 603 - 619, May 2002.

