Abstract: In this study, a multiple hypothesis tracking (MHT) algorithm for multi-target multi-camera tracking (MCT) with disjoint views is proposed. The authors’ method forms track-hypothesis trees, and each branch of them represents a multi-camera track of a target that may move within a camera as well as move across cameras. Furthermore, multi-target tracking within a camera is performed simultaneously with the tree formation by manipulating a status of each track hypothesis. Each status represents three different stages of a multi-camera track: tracking, searching, and end-of-track. The tracking status means targets are tracked by a single camera tracker. In the searching status, the disappeared targets are examined if they reappear in other cameras. The end-of-track status does the target exited the camera network due to its lengthy invisibility. These three status assists MHT to form the track-hypothesis trees for multi-camera tracking. Furthermore, they present a gating technique for eliminating of unlikely observation-to-track association. In the experiments, they evaluate the proposed method using two datasets, DukeMTMC and NLPR_MCT, which demonstrates that the proposed method outperforms the state-of-the-art method in terms of improvement of the accuracy. In addition, they show that the proposed method can operate in real-time and online.

1 Introduction

A large number of cameras recently have been deployed to cover wide area. Besides, tracking multiple targets in a camera network becomes an important and challenging problem in visual surveillance systems since in-person monitoring wide area is costly and needs a lot of effort. Hence, it is desirable to develop multi-target multi-camera tracking (MTMCT) algorithm. In this paper, our goal is to develop an algorithm that can track multiple targets (especially for pedestrians in this work) in a camera network. The targets may move within a camera or move to another camera and the coverage of each camera does not overlap. To achieve this goal, we need to solve both single camera tracking (SCT) and multi-camera tracking (MCT). There has been great amount of effort made to SCT whereas relatively smaller amount of effort has been done for MCT with disjoint views. Moreover, most MCT approaches only focus on tracking targets across cameras by assuming solved SCT in advance; thus, jointly tracking multiple targets in both within and across cameras still remains to be explored much further.

The proposed MHT algorithm tracks targets across cameras by maintaining the identities of observations which are obtained by solving SCT that tracks targets in within-camera. Thus, our method jointly tracks targets in both within and across cameras. In this work, we adopt the real-time and online method to produce observations by tracking multiple targets in within-camera. These observations obtained from each camera can be fed into the proposed MHT algorithm which solves MCT problem. The proposed MHT algorithm forms track-hypothesis trees with obtained observations either by adding a child node to hypothesis tree, which describes the association between an observation and an existing track hypothesis, or by creating a new tree with one root node indicating an observation, which describes the initiation of a new multi-camera track. Each branch in track-hypothesis trees represents different across camera data association result (i.e., a multi-camera track). To work in concert with SCT, every node in track-hypothesis trees designates certain observation and all leaf nodes have a status. There are three statuses for the proposed MHT and each of which represents a different stage of a multi-camera track, tracking, searching, and end-of-track. With the status, the MHT can form the track-hypothesis trees while simultaneously solving SCT to produce observations. Then it selects the best set of track hypotheses as the multi-camera tracks from the track-hypothesis trees. Furthermore, we propose gating mechanism to eliminate unlikely observation-to-track pairing; this also prevents track-hypothesis trees from unnecessary growth. We propose two gating mechanisms, speed gating and temporal gating in order to deal different tracking scenarios (tracking targets on the ground plane or image plane).

For the appearance feature of an observation, we used simple averaged color histogram as an appearance model after Convolutional Pose Machine is applied to an image patch of a person in order to capture the pose variation. The experimental results shows that our method achieves state-of-the-art performance on DukeMTMC dataset and performs comparable to the state-of-the-art method on NLPR_MCT dataset. Furthermore, we demonstrate that proposed method is able to operate in real-time with real-time SCT in Section 4.3.

The remainder of this paper organized as follows. In Section 2, we review relevant previous works. The detailed explanation of proposed method is given in Section 3. Section 3.1 describes how the proposed MHT forms track-hypothesis trees while it simultaneously works with SCT. The proposed gating mechanism is explained in Section 3.2. In Section 4, we report experiment results with conducted on DukeMTMC and NLPR_MCT datasets. Finally, we conclude the paper in Section 5.

2 Related Works

Single camera tracking (SCT), which tracks multiple targets in a single scene, is also called multi-object tracking (MOT). Many approaches have been proposed to improve the MOT. Track-by-detection, which optimizes a global objective function over many frames have emerged as a powerful MOT algorithm in recent years. Network flow-based methods are successful approaches in track-by-detection techniques. These methods efficiently optimize their objective function using the push-relabel method.
and successive shortest path algorithms [8][13]. However, the pairwise terms in network flow formulation are restrictive in representing higher-order motion models, e.g., motion model and constant velocity model [11]. In contrast, formulating multi-object tracking with multidimensional assignment (MDA) problem produces more general representations of computed trajectories since MDA can exploit the higher-order information [11][12]. Solutions for MDA are MHT [12][14] and Markov Chain Monte Carlo (MCMC) data association [15]. While MCMC data association exploits the stochastic method, MHT searches the solution space deterministic way [16].

Multiple Hypothesis Tracking (MHT) was first presented in [16] and is regarded as one of the earliest successful algorithms for visual tracking. MHT maintains all track hypotheses by building track-hypothesis trees whose branch represent a possible data association result(a track hypothesis). The probability of a track hypothesis is computed by evaluating the quality of data association result the branch had. An ambiguity of data association which occurs due to either short occlusion or missed detection does not usually matter for MHT since the best hypothesis is computed with higher-order data association information and entire track hypotheses. In this paper, we applied MHT to solve the multi-camera tracking problem.

Multi-camera tracking aims to establish target correspondences among observations obtained from multiple cameras so as to achieve consistent target labelling across all cameras in the camera network [4]. Earlier research works in MCT only try to address tracking targets across cameras, assuming solved SCT. However, researchers have argued recently that assumptions of availability of intra-camera tracks are unrealistic [3]. Therefore, solving MCT problem by simultaneously treating problem of SCT seems to address more realistic problem. Y.T Tesfaye et al. [4] proposed a constrained dominant set clustering (CDSC) based framework that utilizes a three layers hierarchical approach, where SCT problem is solved using first two layers, and later in the third layer MCT problem is solved by merging tracks of the same person across different cameras. In this paper, we also solve the problem of across camera data association (MCT) by the proposed MHT while SCT is simultaneously treated by real-time multi-object tracker such as [8][17].

Multi-camera tracking with disjoint views is a challenging problem, since illumination and pose of a camera changes across cameras as well as a track discontinues owing to the blind area of camera network or miss detections. Some MCT methods try to relax the problem by delaying data association decisions by keeping multiple hypotheses active until data association ambiguities are resolved [15]. With MHT since the best hypothesis is computed with higher-order data association result the branch had. An ambiguity of data association which occurs due to either short occlusion or missed detection does not usually matter for MHT since the best hypothesis is computed with higher-order data association information and entire track hypotheses. In this paper, we applied MHT to solve the multi-camera tracking problem.

For a multi-camera multi-target tracking system with disjoint views, the set $S = \{x_t\}_{t=1:C}$ is the set of all cameras in the camera network, where $C$ is the number of cameras. Let $K$ denote the most recent time. Single camera tracker of each camera generates observations $o_i$ so that they form a set of observations $O = \{o_t\}_{t=1:N}$ where $N$ is the total number of observations observed until the most recent time $K$. An observation contains all the information about the target while it was being tracked by a single camera tracker. Specifically, the $i$-th observation $o_i = \{A_i, X_i, \pi_i\}$ consists of the appearance feature $A_i$, the track $X_i$ and the camera in which it appears $\pi_i \in S$. The track $X_i$ is a collection of all track histories of observation $o_i$, i.e. $X_i = [x_i^t]_{t=1:T_i}$, where $|X_i|$ represents the length of track of $o_i$ recorded until time $K$ and $x_i^t = (t_i^t, u_i^t, v_i^t, w_i^t, h_i^t, y_i^t)$ refers the $t$-th track history of $o_i$ containing time stamp $t_i^t$, position $(u_i^t, v_i^t)$ and size $(w_i^t, h_i^t)$ in image plane of camera $\pi_i$, and $y_i^t = (x_i^t, y_i^t)$ is a position on ground plane. Note that depending on the scenario, $x_i^t$ might not contain $y_i^t$. With this observation set, the MCT system outputs a set of multi-camera tracks, $T = \{T_j\}_{j=1:T}$ where $|T|$ is the size of set $T$ and $T_j = \{o_{i_j}\}_{i_j \subseteq \{1:N\}}$ refers to the $j$-th multi-camera track which consists of observations, i.e. a number of observations(or single observation) which have the same identity composes a multi-camera track. The $I_j$ is an index set whose elements indicate elements of set $O$, hence, $I_j = \{1\}$. A subset of $O$, we introduce a notation $\{I_j\}_{j=1:T}$ enumerating the set $I_j$ and $I_j$ (the $l$-th element of $I_j$) indicates an observation that is also the $l$-th observation of $T_j$. Thus, we can write $T_j = \{o_{i_j}\}_{i_j \subseteq \{1:N\}}$, which means that the $l$-th observation of the multi-camera track $T_j$ is the $o_{i_l}$ in the set $O$. Therefore, $|I_j|$ is not only the number of elements in $I_j$ but also the number of observations that the $T_j$ had. Finally, for all observations and all multi-camera tracks there is the constraint that one observation must belong to a unique multi-camera track such that:

$$T_j \cap T_k = \emptyset, \forall T_j, T_k \in T, j \neq k.$$  (1)

i.e all tracks in the set $T$ do not conflict each other.

### 3.1 Multiple Hypothesis Tracking for MCT

In this section, we introduce how a track-hypothesis tree is formed for a multi-camera tracking system. The tree of our method maintains multi-camera tracks by initiating, terminating, updating with new observations. A node of the tree represents an observation that is generated by SCT. All branches of the tree represent all possible hypotheses that originate from a single observation or root node. A key strategy of MHT is to delay data association decisions by keeping multiple hypotheses active until data association ambiguities are resolved [12]. As new observations are received, MHT forms new trees to initiate tracks for each new observation. Then existing tracks are updated with new observations that were within the gate. Moreover, all existing tracks are updated with dumply observations in order to describe the hypothesis that they are not updated with any current observation(missing detection). Consequently, the number of track hypotheses continues to expand and many of the tracks are inconsistent since the same observations are used for more than one track.

In the tracking literature, a scan is the time interval and sensor FoV (field of view) where observations are collected. [12]. With previous MHT algorithms for vision based target tracking systems [12][13], images were scanned frame-by-frame to gather observations using a feature detector such as person detector [27] and corner

$$\text{draw line}$$
Now we introduce the three statuses \( \{ w_1, w_2, w_3 \} \). Every leaf node of our track-hypothesis trees should have a status in order that a branch, or a track hypothesis, has a status. Note that intermediate nodes have no effect on the status. The first status, \( w_1 \), is tracking status, meaning that the target is being tracked by a single camera tracker. Therefore, all tracks are initiated with status \( w_1 \) and the track hypothesis in this status is not updated with new observations. If the target disappears from a camera, the status of leaves which refer to this track hypothesis should manage the incompatible track lists. For example, if the track hypothesis set is computed \( (N = 2) \), the dummy nodes are marked with rectangles and can be any of the three statuses. For a more detailed explanation, refer to Section 3.1.

Detector \( 23 \). Hence, the depth of their track-hypothesis trees grow with every frame. However, unlike their approach, we gather observations by scanning the entire camera network within a fixed amount of time; consequently, our tree is extended after a scan. Setting an appropriate interval for one scan is important, because a scan should not contain multi-camera tracks. For example, if the interval is long enough to have observations that could form a multi-camera track; then the system loses the chance to associate them correctly (Figure 2a). On the other hand, too-short time interval for a scan leads to increased computational overhead as well as deepened track-hypothesis tree due to the frequent update of trees. The amount of time for a scan generally depends on the datasets. Furthermore, to prevent the trees from growing meaningless by appending only dummy node to all branches, trees are extended with dummy only when a scan contains new observations (Figure 2b). This enhances the efficiency of tree formation if pedestrians enter into the camera network sparsely.

![Fig. 1: Example of track-hypothesis tree formation](image)

(a) Observations in camera network
(b) First scan
(c) Third scan
(d) Fourth scan
(e) Sixth scan
(f) N-scan pruning

The last hypothesis set

The first status, \( w_1 \), is tracking status, meaning that the target is being tracked by a single camera tracker. Therefore, all tracks are initiated with status \( w_1 \) and the track hypothesis in this status is not updated with new observations. If the target disappears from a camera, the status of leaves which refer to this track hypothesis should manage the incompatible track lists. For example, if the track hypothesis set is computed \( (N = 2) \), the dummy nodes are marked with rectangles and can be any of the three statuses. For a more detailed explanation, refer to Section 3.1.

Detector \( 23 \). Hence, the depth of their track-hypothesis trees grow with every frame. However, unlike their approach, we gather observations by scanning the entire camera network within a fixed amount of time; consequently, our tree is extended after a scan. Setting an appropriate interval for one scan is important, because a scan should not contain multi-camera tracks. For example, if the interval is long enough to have observations that could form a multi-camera track; then the system loses the chance to associate them correctly (Figure 2a). On the other hand, too-short time interval for a scan leads to increased computational overhead as well as deepened track-hypothesis tree due to the frequent update of trees. The amount of time for a scan generally depends on the datasets. Furthermore, to prevent the trees from growing meaningless by appending only dummy node to all branches, trees are extended with dummy only when a scan contains new observations (Figure 2b). This enhances the efficiency of tree formation if pedestrians enter into the camera network sparsely.

Now we introduce the three statuses \( \{ w_1, w_2, w_3 \} \). Every leaf node of our track-hypothesis trees should have a status in order that a branch, or a track hypothesis, has a status. Note that intermediate nodes have no effect on the status. The first status, \( w_1 \), is tracking status, meaning that the target is being tracked by a single camera tracker. Therefore, all tracks are initiated with status \( w_1 \) and the track hypothesis in this status is not updated with new observations. If the target disappears from a camera, the status of leaves which refer to the target changes to tracking, \( w_2 \). A leaf node with status \( w_2 \) is updated with new observations that satisfy the gating condition. The last status, \( w_3 \), is end-of-track, which means that the target exits the camera network due to its invisibility for a long period of time. A branch with this status will not be updated with new observations except with dummy observations. A leaf with status \( w_2 \) changes to status \( w_3 \) by checking the elapsed time from when the leaf began and the most recent time \( K \) (refer to Section 3.2). Note that only leaf nodes in status \( w_3 \) can be updated with new observations by appending a node referring to a new observation as its children. Otherwise, a leaf node in either status \( w_1 \) or \( w_3 \) only appends a dummy observation indicating the same observation as its parent after the scan that receives any new observations. Thus, once the status of a track hypothesis has changed to \( w_3 \), it can not revert to either \( w_1 \) or \( w_2 \). From here, we explain how we form the tree using the example in Figure 1b. In Figure 1b, after the first scan, a tree (tree 1) with a node indicating observation \( o_1 \) is formed to make a track hypothesis that \( o_1 \) initiates a new multi-camera track. After the second scan, the status of the leaf node of tree 1 changes to \( w_2 \) because \( o_1 \) is no longer seen by Camera 1. In the third scan, two new observations, \( o_2 \) and \( o_3 \), are received; hence, two trees, tree 2 and tree 3, are newly formed for \( o_2 \) and \( o_3 \), respectively. Then, the existing track hypothesis in status \( w_2 \) (the root node of tree 1) is associated with new observations by appending nodes referring to \( o_2 \) and \( o_3 \), respectively. A dummy node is also added in order to describe the hypothesis that both \( o_2 \) and \( o_3 \) are not \( o_1 \) (Figure 1c). Note that we only consider the temporal gating scheme in this example for simplicity and the gating time of \( o_2 \) covers initiation time of both observations, \( t_2^1 \) and \( t_3^1 \). After the third scan, because the gating time of \( o_1 \) is expired, the status of leaf node referring to \( o_2 \) is changed to \( w_3 \) (rightmost leaf of tree 1 in Figure 1c). Next, two targets exit from the Camera 2 in the fourth scan. The status of the leaf nodes related to the exited targets is changed. For \( o_2 \), the status of two leaves are changed to \( w_2 \) because \( o_2 \) occurred twice in the track-hypothesis trees (leftmost leaf node of tree 1 and root node of tree 2 in Figure 1d). This works again for \( o_3 \). Because no new observation is received in this scan, trees are not extended with dummy nodes. Finally, Figure 1e shows the result of track-hypothesis tree formation after the sixth scan. With these track trees, all possible data association hypotheses can be made. The first branch of tree 1 shows a multi-camera track with a hypothesis that describes the observations \( o_1, o_2, \) and \( o_3 \) have the same identity, while the last branch of tree 1 shows that the multi-camera track consists of only one observation, \( o_1 \) (Figure 1e).

To compute the set that satisfy the constraint, each track hypothesis should manage the incompatible track lists. For example, in Figure 1c the incompatible tracks of the first branch of tree 1 are the track hypotheses that have observation \( o_1, o_2, \) and \( o_4 \), i.e., all the other branches in same the tree (because they at least share \( o_1 \)), all track hypotheses in tree 2 (because of \( o_2 \) and \( o_4 \), the first branch of tree 3 (because of \( o_3 \)), and the root node of tree 4 (because of \( o_4 \)). Then the set of best track hypotheses is computed by solving a maximum weighted independent set problem \( 12 \) which will be described in the Section 3.5.
short scan time and ˙

\[ w \in \Phi \] represent the entry and exit point of the observation, respectively, i.e. \( \pi \) presents the total number of entry/exit points in the camera network.

\[ G \] is the maximum speed of a target. These can either be set by the system designer or learned from training samples. This assumed that \( G \) is the maximum speed of a target. These can either be set by the system designer or learned from training samples. Let \( \alpha \) be the transition matrix, a square matrix with size \( |E| \times |E| \). This transition matrix is computed by \( G^{\text{end}} = \sqrt{G_{\text{area}}/\eta_1} \) where \( G_{\text{area}} \) is the area of the ground plane of the camera network and \( \eta_1 \) is the estimated speed of \( o_1 \).

In some situations, it is impossible to locate targets that are in tracking status \( (v_1) \) due to the absence of calibration information and map information. In this case, we use temporal gating instead of speed gating. First, we estimate entry/exit points for each camera either by learning from training samples or by getting information from the system designer. Let \( E = \{ e_j \}_{j=1}^{N} \) be the set of entry/exit points, and \( e_j \) is the \( j \)-th entry/exit point, where \( |E| \) represents the total number of entry/exit points in the camera network. Then \( \pi_j \) of \( o_1 \) has two elements, \( \pi_j^{en} \) and \( \pi_j^{ex} \), which represent the entry and exit point of the observation, respectively, i.e. \( \pi_j = \{ \pi_j^{en}, \pi_j^{ex} \} \), and \( \pi_j^{en}, \pi_j^{ex} \in E \). After that, we learned the transition matrix between each pair of entry/exit points as well as the mean and standard deviation of transition time using training samples. Let \( \Phi \) be the transition matrix, a square matrix with size \( |E| \times |E| \). An element \( \Phi_{i,j} \), \( i \)-th row and \( j \)-th column, is set to one if a transition from \( e_i \) to \( e_j \) exists. Otherwise, it is set to zero. For all \( \Phi_{i,j} = 1 \), we learned the mean and standard deviation of transition time with training samples. Then the temporal gating for the existing track whose leaf node designates \( o_1 \) and newly received observation \( o_j \) checks the followings:

\[ G_{\text{time}}^{\text{min}} < t_j - t_{|X_j|} < G_{\text{time}}^{\text{max}}, \quad \Phi_{x_j,x_w} = 1, \quad (3) \]

where \( G_{\text{time}}^{\text{min}} \) and \( G_{\text{time}}^{\text{max}} \) are the minimum and maximum threshold for temporal gating, and which can be learned from the training samples. Note that middle-term of the inequality is always positive for MCT with disjoint view, otherwise its absolute value is needed. If the mean transition time between \( \pi_j^{en} \) and \( \pi_j^{ex} \) is \( \mu \) and its standard deviation is \( \sigma \), \( G_{\text{time}}^{\text{min}} \) and \( G_{\text{time}}^{\text{max}} \) could be set to \( \mu - \alpha \sigma \) and \( \mu + \alpha \sigma \), respectively, where \( \alpha_0, \alpha_2 \) are set by the system designer.

If an observation-to-track pair does not satisfy the check, then that pair will not be associated. A leaf node in status \( w_2 \) referring to observation \( o_i \) changes the status to \( w_3 \) if the time gap between last observed time and recent time \( K \) is beyond the predetermined gating time, i.e. it will change the state to \( w_3 \) if \( K - t_{|X|} \) is larger than \( G_{\text{end}} \). In this case, \( G_{\text{end}} \) is used for all observations.

3.3 Pruning

Since there is potential for a combinatory explosion in the number of track hypotheses that our MHT system could generate, pruning the track-hypothesis tree is an essential task for MHT. We adopted the standard N-scan pruning technique[12 26]. The standard N-scan pruning algorithm assumes that any ambiguity at \( K \) is resolved by time \( K + N \), i.e. it defines the number of frames to look ahead in order to resolve an ambiguity[13]. It notes that in our case, \( N \) refers not to the number of frames but to the number of scans since our trees grow after the scan where any new observation is received. An example of N-scan pruning is described in Figure[11] where \( N = 2 \). First, finding the best track hypothesis set is needed before pruning the trees. Computing the best track hypothesis set using the track score is described in Section 3.4 and 3.5. After we identify the best hypothesis set, we ascend to the parent node \( N \) times from each selected leaf node to find the decision node. At that node, we prune the subtree that diverged from the best track. Consequently, we have a tree of depth \( N \) below the decision node, while the tree is degenerated into simple list of assignments above the decision node.

3.4 Scoring a track

The evaluation of a track hypothesis should deal all aspects of data association quality that a multi-camera track possess. According to the original formulation[26], we define a likelihood ratio ratio(\( L \)) of a track \( T_j = \{ o_{1j} \} \) to be

\[ L(T_j) = \frac{P(\{ o_{1j} \}|H_1) P_0(H_1)}{P(\{ o_{1j} \}|H_0) P_0(H_0)} \quad (4) \]

where hypotheses \( H_1 \) and \( H_0 \) are the true target and false alarm hypotheses of given combination of data, i.e. \( P(\{ o_{1j} \}|H_i) \) is probability density function evaluated with given data \( \{ o_{1j} \} \) under the

![Fig. 2: (a) An example of long scan time. The system cannot generate hypotheses associating observation \( o_1 \) with \( o_2 \) as well as \( o_3 \) with \( o_4 \), (b) Short scan time](image-url)
assumption that \( H_t \) is correct. The \( P_t(H_t) \) is a prior probability of \( H_t \). The conditional probabilities in Equation \( (3) \) can be partitioned into a product of two terms, \( L_A(T_j) \) and \( L_X(T_j) \), assuming that the appearance and kinematic information of a target are independent each other. Therefore,

\[
L(T_j) = L_0 L_A(T_j) L_X(T_j) = L_0 \prod_{k=1}^{T_j} \frac{p(A_{i_j}|H_0) p(X_{i_j} | \pi_{i_j} | H_0)}{p(A_{i_j} | H_0) p(X_{i_j} | \pi_{i_j} | H_0)}
\]  

(5)

where \( L_0 = \frac{P_t(H_0)}{P_t(H_t)} \), the second term and third term in rightmost side are \( L_A(T_j) \) and \( L_X(T_j) \), respectively. The Equation \( (5) \) can be further factorized by chain-rules:

\[
L(T_j) = L_0 \prod_{k=1}^{T_j} L_A(T_k) L_X(T_k)
\]  

(6)

where assuming that received observations are conditionally independent under the false alarm hypothesis. \( L_A(T_k) \) and \( L_X(T_k) \) are the appearance and kinematic likelihood ratio when the \( k \)-th observation is associated with the existing track \( T_{k-1} \).

For likelihood of the kinematic term, \( L_X \), we define two different measures in order to differentiate the tracking scenarios. The first one is for the scenario that tracking targets on the image plane of each camera is only available. In this case, we assumed that transition time across cameras is normally distributed, i.e.

\[
p(X_{i_j} | \pi_{i_j} | \{X_{i_j+1} \}, \mu, \sigma^2) = N(t_{i_j}^{1:k} - t_{i_j}^{k-1}; \mu, \sigma^2)
\]  

(7)

where the mean \( \mu \) and variance \( \sigma^2 \) would be estimated using training samples which moved from \( t_{i_j}^{k-1} \) to \( t_{i_j}^{k-1} \). Note that we dropped sub-script for \( \mu \) and \( \sigma^2 \) for simplicity. They should be learned for all pairs of possible transitions (i.e., for all \( \Phi_{e_i e_j} = 1 \) where \( e_i, e_j \in E \)). The \( t_{i_j}^{k} \) is the time stamp of the initiation time of \( o_{i_j} \) whereas \( t_{i_j}^{k-1} \) is the time stamp of the last observed time of \( o_{i_j} \).

The other kinematic likelihood function is for the scenario where we can locate a target on the ground plane of the camera network; hence, measuring distance between tracks is feasible. Let \( \hat{y}_i \) be the moving speed of an observation \( o_i \) and it is estimated by averaging:

\[
\hat{y}_i = \frac{1}{|X_i| - 1} \sum_{l=2}^{|X_i|} \frac{||X_i^l - X_i^{l-1}||_2}{t_i - t_i^{l-1}}.
\]  

(8)

Then the likelihood function is also assumed to be Gaussian:

\[
p(X_{i_j}, \pi_{i_j} | \{X_{i_j+1} \}, \mu, \sigma^2)
\]  

(9)

\[
d = \beta|\hat{y}_i^{j-1} - \hat{y}_i^{j-1}| + (1 - \beta)|\hat{y}_i - \hat{y}_i^{j-1}|
\]  

(10)

where \( d \) is the estimated travel distance of \( o_{i_j}^{j-1} \), and \( d \) (which comes from the Equation \( (10) \) distance is distance between \( o_{i_j}^{j-1} \) and \( o_{i_j}^{j-1} \). Note that although both \( d \) and \( d \) are function of \( X_{i_j} \) and \( X_{i_j}^{j-1} \), we dropped them for the simplicity of notation. The \( \gamma \) is the precision for the Gaussian distribution. For the false alarm hypothesis of kinematic term, \( p(X_{i_j} | \pi_{i_j} | H_0) \), we set to constant probability, \( 0 < C_1 < 1 \).

To compute the appearance likelihood, we first built an color histogram for the appearance feature \( A_i \) of observation \( o_i \) while it has tracked. The learned appearance model, \( A_j^{l:i} \), is constructed for the track \( T_j^{l:k} = \{o_{j_l} \}_{l=1:k} \), after each association between the existing track \( T_j^{l:k-1} \) and a new observation \( o_{j_l} \) is made. i.e.

\[
\tilde{A}_{j_l} = \frac{k - 1}{k} \tilde{A}_{j_l} + 1 + \frac{1}{k} A_i
\]  

(11)

where \( A_{j_l}^{l:i} \) is learned appearance feature for the track \( T_j^{l:k-1} \). Thus, \( A_{j_l}^{l:i} \) is the averaged feature over \( k \) associated observations. This averaging model is used because if the track hypothesis consistently associated the observations which had the same identity then the averaged feature would have the ability to classify correctly than that of inconsistently associated track due to its stable distribution of colors. Then the appearance likelihood is computed by comparing two histogram:

\[
p(A_i | A_{j_l}^{l:i}, H_0) = p(A_i | A_{j_l}^{l:i}, H_1) = D_i (A_i, \tilde{A}_{j_l}^{l:i}),
\]  

(12)

where \( D_i \) is similarity measure between two histograms and it can be any metric such as, Bhattacharyya, histogram intersection, earth mover distance and so on. However, some metrics should be modified in order to use it as the probability (i.e., \( 0 \leq D_i \leq 1 \)). The false alarm hypothesis of appearance term, \( p(A_i | H_0) \), is the constant probability \( 0 < C_2 < 1 \).

Next, we define the log likelihood ratio, or score, for a multi-camera track, \( T_j = \{o_{j_l} \} \) consists of \( [I] \) observations, which is the sum of \( [I] \) kinematics and \( [I] \) appearance related terms plus the initiation score. That is:

\[
\log L(T_j) = \log L_0 + \sum_{k=1}^{T_j} \left[ \log L_A(T_k) + \log L_X(T_k) \right]
\]  

(13)

where \( \log L_0 \) is track initiation score and we set it to a constant \( C_0 \). Then the score of a track can be computed recursively:

\[
\log L(T_j) = \log L(T_j^{1:[i_j]}) = \log L_0 + \sum_{k=1}^{T_j^{1:[i_j]}} [\log L_A(T_k^{1:[i_j]}) + \log L_X(T_k^{1:[i_j]})]
\]  

(14)

where \( \log L(T_j^{1:[i_j]}) \) is the score of track \( T_j^{1:[i_j]} \) and \( \Delta \log L(T_j^{1:[i_j]}) \) is the increment that occurs upon update with a new observation, \( o_{j_l}^{1:[i_j]} \). Finally, we introduce the weights, \( w_A \) and \( w_X \), that control contribution of appearance and kinematics to the score, respectively:

\[
\Delta \log L(T_j^{1:[i_j]}) = \log \left[ \frac{w_A \log L_A(T_j^{1:[i_j]}) + w_X \log L_X(T_j^{1:[i_j]})}{w_A \log L_A(T_j^{1:[i_j]}) + w_X \log L_X(T_j^{1:[i_j]})} \right]
\]  

(15)

where \( w_A + w_X = 1 \). The score is continuously updated as long as the track hypothesis is updated with a new observation.

3.5 Computing the Best Hypothesis Set

In this section, we describe how the best hypothesis set is computed among all track-hypothesis trees that maintain all the possible multi-camera tracks using all observations. The best hypothesis set is computed every scan if the scan received any new observation, and then tree pruning is performed to avoid exponential growth of trees.
Fig 3: An example of MWIS graph that corresponds to the trees in Figure 1c. The red vertices have been selected for the best hypothesis set.

To compute the best hypothesis set, we adopt the approach of [12]. They change the task to the Maximum Weighted Independent Set (MWIS) problem. MWIS [14] is equivalent to the Multi-Dimensional Assignment (MDA) problem in the context of MHT [12]. MDA is used to compute the most probable set of tracks [28, 29] and its MDA formulation for MHT was introduced in [28, 29].

Let $G = (Z, E)$ be the undirected graph for MWIS, which corresponds to a set of track-hypothesis trees generated by MHT. Then the solution of MWIS can be determined by solving the discrete optimization problem:

$$
\max_x \sum_i c_i z_i \quad \text{s.t.} \quad z_i + z_j \leq 1, \quad \forall (i, j) \in E, \quad z_i \in \{0, 1\}.
$$

Each vertex $z_i \in Z$ is assigned to a track hypothesis $T_i$ as well as a vertex $z_j$ has weight $c_j$ that corresponds to its track score, $\log \ell(T_i)$. There is an undirected edge, $(i, j) \in E$, linking two vertices $z_i$ and $z_j$ if the two tracks are incompatible due to shared observations, i.e. if they are violating constraint $\mathbf{c}$. Therefore, the constraint in Equation (15) is a discretized form of constraint $\mathbf{c}$. In the graph $G$, an independence set is a set of vertices no two of which are adjacent, i.e. all tracks in the independence set are compatible. The maximum weighted independence set in $G$ is an independence set with maximum total weight, i.e. a set of compatible tracks of which the total track score is maxima.

An example of the MWIS graph is shown in Figure 1c. It corresponds to the set of track-hypothesis trees in Figure 1e. Each vertex represents a track hypothesis (each branch of a tree) and the observations used for that track are shown at the vertex with three-digits, where zero denotes a dummy observation. Note that since three scans (first, third and fifth scans) have received observations (among the six scans in Figure 1a) a track can contain at most three observations (Refer to Section 3.1 for more details). The red vertices are selected for the best hypothesis set in the example Figure 3. We used the Gurobi optimizer to solve the above MWIS problem.

4 Experiments

We evaluated our method using two datasets: DukeMTMC [30] and NLPR_MCT [31]. These datasets were designed for a multi-camera tracking system. The DukeMTMC dataset consists of eight synchronized cameras, which was recorded at 1080p resolution and 60 fps. The dataset contains more than 7,000 single camera trajectories and over 2,000 unique identities over 85 minutes for each camera, a total of more than 10 h. We used ID-Measure [32] with the DukeMTMC dataset to evaluate our multi-camera tracking performance.

The NLPR_MCT dataset provides four different videos that are, at most, 25 minutes long, and they all have a resolution of $320 \times 240$. To measure the performance for this dataset, we used the MCTA metric [31]. Table 1 shows the parameter setting for each dataset in this section. Note that $G_{\text{speed}}$ is in meters per second and $G_{\text{time}}$ is in seconds.

### Table 1 Parameter settings

| Dataset          | $N$-pruning | $w_A$ | $C_0$ | $C_1$ | $C_2$ | Scan time | $\beta$, $G_{\text{speed}}$, $G_{\text{time}}$ | $\mu_{\min}$ | $\mu_{\max}$ |
|------------------|-------------|-------|-------|-------|-------|-----------|---------------------------------|---------------|---------------|
| DukeMTMC         | 10          | 0.8   | 0.001 | 0.3   | 0.75  | 1 sec     | $0.7, 0.5, 2.0$, N/A             | $\mu - 2.5\sigma$, $\mu + 2.5\sigma$ |
| NLPR_MCT         | 10          | 0.815 | 0.005 | 0.1   | 0.75  | 1 sec     | N/A                             | \text{N/A}   | \text{N/A}    |

Fig 4. (a) shows the image patches of a person and results of pose estimation, (b) corresponding upper body parts, (c) corresponding bottom body parts.

To solve the SCT problem as well as to generate observations that is inputted to the proposed MHT, we used online and real-time MOT [5] for each camera. Utilizing the online and real-time single camera tracker would not affect the online and real-time capability of unified framework. In Section 4.3 we show that our unified framework works in real-time provided that the SCT problem is solved in real-time.

4.1 Appearance Modeling

In this subsection, the appearance model for an observation is described. The appearance feature $A_i$ is the averaged histogram of observation $o_i$ that is learned while under tracking status ($w_i$) of $o_i$. To be more specific, let $a^i_l$ be the extracted feature from corresponding image location $x^i_l$, which is the $l$-th track history of observation $i$. Then the appearance feature $A^i_l$ that is computed from start to the $l$-th track history is:

$$
A^i_l = \frac{1}{l} - \frac{1}{l-1} A^i_{l-1} + \frac{1}{l} a^i_l
$$

where $l \leq |X_i|$. Note that to keep the online nature of the proposed method, $A^i_0$ is used to compute Equation (11) instead of $A_i$.

For appearance feature, we use an HSV (hue, saturation, value) color histogram for upper and lower body parts where the bin size is 16, 4 and 4 for Hue, Saturation and Value channels respectively. Furthermore, to capture the pose variations of a person, we use the Convolutional Pose Machine [6] to estimate the pose of a given image patch of a person. The estimated pose of a person is depicted in Figure 1a. To extract the upper body part, four joints (right shoulder, right hip, left shoulder, left hip) are used (Figure 1b). Six joints (right hip, right knee, right ankle, left hip, left knee, left ankle) are used for the bottom part of the body (Figure 1c). Once those body parts are extracted from an image patch, an HSV histogram, $A^i_l$, is computed.

Even though we have used color histogram as the appearance feature and simple averaging model in this work, other online discriminative models can be applied to proposed MHT method since it is known that MHT can be extended by including online learned discriminative models without difficulty [12].

4.2 DukeMTMC dataset

The DukeMTMC is a large, fully-annotated, calibrated dataset that captures the campus of Duke University, and was recorded using eight fixed cameras. The dataset has an RoI (region of interest) area for each camera where an evaluation is made. The topology of the fixed cameras is shown in Figure 5 where there is no field overlap between any pair of cameras (Figure 5a). The cameras used to acquire the dataset were synchronized and recorded at 1080p resolution and 60fps. The dataset contains more than 7,000 single camera
Table 2: This table compares the results for each camera and average performance over all single cameras as well as the performance of multi-camera tracking on the Test_easy sequence of DukeMTMC dataset.

| Cam | method | ID-Measure | MOTA† | MOTP† | CLEAR MOT | FAF† | ML† | FP† | FN† | IDs† |
|-----|--------|------------|-------|-------|-----------|------|-----|-----|-----|------|
| 1   | Ours  | 57.3       | 91.2  | 41.8  | 63.0      | 79.0 | 0.03| 24  | 46  | 2,713 | 107,178 | 39 |
|     |        | 76.9       | 89.1  | 67.7  | 69.9      | 76.3 | 0.06| 137 | 22  | 5,809 | 52,152   | 156|
|     |        | 84.3       | 89.7  | 79.6  | 84.9      | 79.5 | 0.04| 191 | 12  | 3,679 | 25,318   | 55 |
| 2   | Ours  | 68.0       | 69.3  | 67.1  | 44.8      | 78.2 | 0.51| 133 | 8   | 47,919| 5,374     | 60 |
|     |        | 81.2       | 90.9  | 73.4  | 71.5      | 74.6 | 0.14| 134 | 21  | 8,487 | 43,912   | 75 |
|     |        | 81.9       | 88.9  | 75.9  | 78.4      | 77.1 | 0.07| 151 | 8   | 6,390 | 33,377   | 81 |
| 3   | Ours  | 60.3       | 78.9  | 48.8  | 57.8      | 77.5 | 0.02| 52  | 22  | 1,438 | 28,692    | 16 |
|     |        | 64.6       | 76.3  | 56.0  | 67.4      | 75.6 | 0.02| 44  | 9   | 2,148 | 21,125    | 38 |
|     |        | 69.3       | 76.2  | 63.5  | 65.7      | 77.0 | 0.06| 58  | 7   | 5,908 | 18,589    | 22 |
| 4   | Ours  | 71.5       | 88.7  | 62.8  | 63.2      | 80.2 | 0.02| 36  | 18  | 2,209 | 19,323    | 7  |
|     |        | 84.7       | 91.2  | 79.0  | 76.6      | 76.6 | 0.03| 45  | 4   | 2,860 | 10,686    | 18 |
|     |        | 80.7       | 84.1  | 77.6  | 79.8      | 80.1 | 0.04| 47  | 3   | 3,633 | 8,173     | 17 |
| 5   | Ours  | 73.2       | 83.0  | 65.4  | 72.3      | 80.4 | 0.05| 101 | 17  | 4,464 | 35,861    | 54 |
|     |        | 68.3       | 76.1  | 61.9  | 68.9      | 74.7 | 0.10| 88  | 11  | 9,117 | 36,933    | 139|
|     |        | 73.7       | 81.4  | 67.3  | 76.6      | 80.0 | 0.05| 110 | 8   | 4,410 | 30,195    | 83 |
| 6   | Ours  | 77.2       | 84.5  | 69.1  | 73.4      | 80.2 | 0.06| 142 | 27  | 5,279 | 45,170    | 55 |
|     |        | 82.7       | 91.6  | 75.3  | 77.0      | 77.2 | 0.05| 136 | 11  | 4,468 | 38,611    | 142|
|     |        | 83.5       | 88.9  | 78.8  | 82.8      | 80.2 | 0.06| 163 | 6   | 5,478 | 27,194    | 69 |
| 7   | Ours  | 80.5       | 93.6  | 70.6  | 71.4      | 74.7 | 0.02| 69  | 13  | 1,395 | 18,904    | 23 |
|     |        | 81.6       | 94.0  | 72.5  | 73.8      | 74.0 | 0.01| 64  | 4   | 1,182 | 17,411    | 36 |
|     |        | 81.3       | 91.4  | 73.5  | 77.0      | 75.5 | 0.01| 69  | 7   | 1,232 | 15,119    | 33 |
| 8   | Ours  | 72.4       | 92.2  | 59.6  | 60.7      | 76.7 | 0.03| 101 | 51  | 2,719 | 52,806    | 46 |
|     |        | 73.0       | 89.1  | 61.0  | 63.4      | 73.6 | 0.04| 92  | 28  | 4,184 | 47,565    | 91 |
|     |        | 79.9       | 90.8  | 71.3  | 71.6      | 75.3 | 0.05| 125 | 21  | 4,850 | 35,288    | 46 |
| Single Cam | Average | 70.1       | 83.6  | 60.4  | 59.4      | 78.7 | 0.09| 663 | 234 | 8,147 | 361,872   | 91 |
|     | Ours  | 80.3       | 87.3  | 74.4  | 78.3      | 78.4 | 0.05| 914 | 72  | 35,580 | 193,253   | 406|
| Multi-Cam |     | 56.2       | 67.0  | 48.4  | 60.0      | 68.3 | 0.03| 65  | 4   | 4,410 | 30,195    | 83 |
|     |        | 60.4       | 68.3  | 53.5  | 65.4      | 71.1 | 0.06| 60  | 6   | 4,410 | 30,195    | 83 |

The evaluation on the Test-easy is shown in Table 2 while the performance on the Test-hard is shown in Table 3. In both tables, the last row is for comparison of multi-camera tracking performance and the rest are for comparison of single camera tracking performance. We used public detection responses as the input to our method. The results of single camera tracking between ours and [30] were different, even though we used the public single camera tracker that is published by E. Ristani et al. [5], because we modified the original one to fit our multi-camera tracking method. The last row of Table 2 shows that the performance of our multi-camera tracking method outperformed the method of [30] even in the complicated video sequence (Test-hard). Even if the [30]'s average IDF1 over single cameras was higher than ours by 1%, the IDF1 of their multi-camera tracking performance was even lower than ours by 2.8% (Table 3).

4.3 NLPR_MCT dataset

The NLPR_MCT dataset consists of four sub-datasets. A sub-dataset is depicted in Figure 6. Each sub-dataset includes 3-5 cameras with non-overlapping scenes and records different situations according to the number of people (ranging from 14 to 255) and the level of illumination changes and occlusions. The videos contain both real scenes and simulated environments. Each video was nearly 20 minutes long (except Dataset 3), with a rate of 25 fps.

In this dataset, the topological connection information for every pair of entry/exit points for each sub-dataset is provided. We split the \( \pi_i \) of an observation \( o_i \) into \( \pi_i^m \) and \( \pi_i^e \), that represent the entry point and exit point of observation \( o_i \), respectively. Because the

Fig. 5: Camera topology of DukeMTMC camera. (a) Each polygon represents the FoV of corresponding camera. (b) Each polygon represents the RoI of corresponding camera.
Table 3 This table compares the results for each camera and average performance over all single cameras as well as the performance of multi-camera tracking on the test_hard sequence of DukeMTMC dataset.

| Cam | method | IDP↑ | IDR↓ | MOTA↑ | MOTP↑ | FA↑ | MT↑ | ML↓ | FP↓ | FN↓ | ID↓ |
|-----|--------|------|------|-------|-------|-----|-----|-----|-----|-----|-----|
| 1   | ID     | 52.7 | 64.6 | 3.78  | 3.76  | 0.03 | 6   | 24  | 1.257 | 78.971 | 55  |
|     | Ours   | 67.1 | 72.2 | 58.4  | 78.1  | 0.08 | 65  | 17  | 2.88  | 44.253 | 408 |
| 2   | ID     | 60.6 | 64.6 | 46.5  | 62.7  | 0.35 | 78  | 11  | 12   | 37.287 | 394 |
|     | Ours   | 66.4 | 72.2 | 58.4  | 76.7  | 0.45 | 86  | 19  | 16   | 54.252 | 323 |
| 3   | ID     | 64.5 | 65.7 | 54.8  | 73.6  | 0.74 | 68  | 12  | 26.326 | 46.898 | 194 |
|     | Ours   | 63.4 | 72.8 | 53.1  | 73.9  | 0.25 | 62  | 16  | 8.653 | 54.625 | 323 |
| 4   | ID     | 56.6 | 61.2 | 50.4  | 54.7  | 0.74 | 66  | 10  | 24.59  | 44.401 | 392 |
|     | Ours   | 58.2 | 61.2 | 52.6  | 74.4  | 0.68 | 66  | 10  | 24.59  | 44.401 | 392 |
| 5   | ID     | 60.4 | 61.2 | 50.4  | 73.6  | 0.74 | 68  | 12  | 26.326 | 46.898 | 194 |
|     | Ours   | 63.4 | 72.8 | 53.1  | 73.9  | 0.25 | 62  | 16  | 8.653 | 54.625 | 323 |
| 6   | ID     | 84.3 | 86.0 | 85.3  | 81.5  | 0.04 | 21  | 0   | 1.215 | 2073   | 1    |
|     | Ours   | 82.7 | 87.1 | 81.1  | 77.7  | 0.05 | 17  | 0   | 1.571 | 3888   | 61   |
| 7   | ID     | 83.2 | 84.4 | 82.1  | 81.6  | 0.05 | 20  | 1   | 1.821 | 2404   | 1    |
|     | Ours   | 81.9 | 90.1 | 78.3  | 80.7  | 0.04 | 21  | 0   | 1.215 | 11644  | 100  |
| 8   | ID     | 82.8 | 91.5 | 78.6  | 76.7  | 0.03 | 47  | 2   | 1.219 | 11644  | 100  |
|     | Ours   | 85.7 | 93.3 | 79.2  | 81.9  | 0.01 | 52  | 2   | 785   | 10017  | 24   |
| Single Cam Average | | 64.1 | 81.7 | 59.4  | 76.7  | 0.14 | 85  | 23  | 5.156 | 77031  | 225  |
| Ours | | 53.1 | 71.2 | 53.3  | 76.5  | 0.17 | 68  | 36  | 5.989 | 88164  | 547  |
| 7   | ID     | 54.7 | 70.0 | 56.1  | 77.8  | 0.22 | 82  | 24  | 7.902 | 80716  | 423  |
|     | Ours   | 53.7 | 70.0 | 56.1  | 77.8  | 0.22 | 82  | 24  | 7.902 | 80716  | 423  |
| 8   | ID     | 59.6 | 81.2 | 50.8  | 73.3  | 0.08 | 43  | 23  | 2.971 | 38124  | 148  |
|     | Ours   | 60.6 | 84.7 | 50.8  | 74.0  | 0.05 | 34  | 20  | 1.935 | 39865  | 266  |
| 9   | ID     | 82.4 | 84.9 | 73.0  | 75.9  | 0.02 | 24  | 5   | 7.106 | 9775    | 10   |
|     | Ours   | 81.3 | 90.3 | 73.9  | 70.0  | 0.06 | 37  | 6   | 2.297 | 9306   | 26   |
| Multi-Cam | | 66.4 | 81.2 | 54.8  | 77.1  | 0.14 | 338 | 10  | 39.399 | 283376 | 652  |
| Ours | | 65.4 | 81.4 | 54.7  | 75.4  | 0.09 | 348 | 99  | 26.643 | 26073  | 1637 |
|     | Single Cam Average | | 63.5 | 73.9 | 55.6  | 79.6  | 0.19 | 400 | 80  | 55.038 | 231527 | 1468 |

Table 4 This table compares the results of each sub-dataset. MCTA metric is used for evaluation.

| Method | NLPR 1 | NLPR 2 | NLPR 3 | NLPR 4 | Avg. Rank |
|--------|--------|--------|--------|--------|-----------|
| USC Vision | 0.9152 | 0.9132 | 0.9162 | 0.7051 | 3.5 |
| UW_P4 [34] | 0.9610 | 0.7965 | 0.7899 | 0.7578 | 1.25 |
| CRF_UCR [35] | 0.8383 | 0.8015 | 0.6645 | 0.7266 | 3.75 |
| EGTracker [31] | 0.8353 | 0.7034 | 0.7417 | 0.3845 | 4.75 |
| DukeMTMC [30] | 0.7967 | 0.7336 | 0.6543 | 0.7616 | 4.25 |
| Ours | 0.9129 | 0.8944 | 0.6699 | 0.6812 | 3.5 |

Fig. 6 A sub-dataset of NLPR_MCT dataset. The numbers in each scene represent the entry/exit points.

4.4 N-scan Pruning

In this section, we report how our MHT algorithm is sensitive to the parameter N of N-scan pruning. The N-scan pruning algorithm assumes that any ambiguity at K is resolved by time K + N. The N-scan pruning was utilized in the multi-scan assignment approach to MHT because it solves the data association problems with recent N scans of the data thanks to N-scan pruning [26, 29]. We evaluated the ID-Measure for various N, 1 ≤ N ≤ 20, on the Trainval-Mini of the DukeMTMC dataset. The Trainval-Mini is a small part of the training/validation set of the DukeMTMC dataset and is about 18 minutes long sequence. The parameter settings were the same as before except the N. The intersection-over-union is fixed to 0.5 for this experiment. The experimental result is shown in Figure 7. The result demonstrates that our MHT algorithm is negatively affected by N in IDP while it is slightly positively affected in IDR when N is increasing. Therefore, the proposed method has a small sensitivity to N in terms of IDF1 because the IDP and the IDR are negatively correlated with respect to the N. The difference between the minimum(58.02%) and maximum(58.87%) values in IDF1 was 0.85%. The minimum and maximum value of IDP was 64.97% and 58.87%, respectively. While that of IDR was 51.98% and 53.08%, respectively.

4.5 Real-time implementation

In this section, a real-time implementation of the proposed method is described. Full implementation of the proposed method was programmed by Matlab with a desktop PC (Intel i7-4790K 4.0 Ghz
4-core CPU, 16GB RAM, Nvidia GTX 770 GPU and Ubuntu 16.04 OS). For the computation efficiency, we have discarded the Convolutional Pose Machine in the appearance modeling and switched the SCT algorithm from [5] to GM-PHD (Gaussian Mixture Probability Hypothesis Density) filter [17] for real-time implementation. There was no other reason for switching the SCT algorithm except that we had already implemented the GM-PHD in C++. Real-time implementation was developed using Visual C++ with the multi-thread programming and OpenCV in Windows 10 OS. The test hardware environment included a PC with Intel i7-7700K 4.5 Ghz 4-core CPU, 32GB RAM. To test the processing speed, we have generated a new dataset consisting of six videos with 6×10 480 resolution and about 7 minutes long (Figure 5). This dataset, which includes appearance of up to 25 targets, was recorded in the campus of Gwangju Institute of Science and Technology. To detect a person in the dataset, we applied the pedestrian detector [36], which processes every frame of each camera to detect pedestrians. The average processing time (includes the processing time of detection, SCT and MCT) was about 15 frames per second for all videos. Note that each video(camera) was processed in parallel by multi-thread programming. Therefore, this result demonstrates the real-time performance of our method. We used the Gurobi optimizer to solve MWIS problems in these implementations.

5 Conclusion

In this paper, we applied a multiple hypothesis tracking algorithm to handle the multi-target multi-camera tracking problem with disjoint views. Our method forms track-hypothesis trees whose branch represents a multi-camera track which describes the trajectory of a target that may move within a camera as well as move across cameras. Furthermore, tracking targets within a camera is performed simultaneously with the tree formation by manipulating a status of each target hypothesis. Besides, two gating schemes have been proposed to differentiate the tracking scenarios. The experimental results shows that our method achieves state-of-the-art performance on DukeMTMC dataset and performs comparable to the state-of-the-art method on NLPR_MCT dataset. We also show that the proposed method can solve the problem under online and real-time conditions, provided that the single camera tracker solves in such conditions as well.

MHT can be extended by including online learned discriminative appearance models for each track hypothesis [12]. Therefore, as for the future work, we will investigate online learning techniques that could learn a model for each hypothesis since we used a simple averaging model for appearance modeling in this work.

Fig. 7: The parameter $N$ of $N$-scan pruning versus performance of ID-Measure

Fig. 8: The dataset used in Section 4.5

6 Acknowledgement

This work was supported by Institute for Information and communications Technology Promotion (ITP) grant funded by the Korea government (MSIP) (No. B0101-15-0525, Development of global multi-target tracking and event prediction techniques based on real-time large-scale video analysis). In addition, this work also supported by Korea Creative Content Agency (KOCCA) and Ministry of Culture, Sports and Tourism (MCST) (No. R20170750052, Developed intelligent UI/UX technology for AR glasses-based docent operation).

7 References

1. X. Chen, K. Huang, and T. Tan, “Object tracking across non-overlapping views by learning inter-camera transfer models,” Pattern Recognition, vol. 47, no. 3, pp. 1126–1137, 2014.
2. O. Javed, Z. Rasheed, K. Shafique, and M. Shah, “Tracking across multiple cameras with disjoint views,” in Proceedings of the Ninth IEEE International Conference on Computer Vision—Volume 2, p. 952. IEEE Computer Society, 2003.
3. Y. Wang, S. Velipassalar, and M. C. Gursoy, “Distributed wide-area multi-object tracking with non-overlapping camera views,” Multimedia tools and applications, vol. 73, no. 1, pp. 7–39, 2014.
4. Y. T. Tesfaye, E. Zemene, A. Prati, M. Polillo, and M. Shah, “Multi-target tracking in multiple non-overlapping cameras using constrained dominant sets,” arXiv preprint arXiv:1706.06136 2017.
5. E. Rostami and C. Tomasi, “Tracking multiple people online and in real time,” in Asian Conference on Computer Vision, pp. 444–459, Springer, 2014.
6. S.-E. Wei, V. Ramakrishna, T. Kanade, and Y. Sheikh, “Convolutional pose machines,” in Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 4724–4732, 2016.
7. X. Wang, E. Tiretken, F. Fleuret, and P. Fua, “Tracking interacting objects using intertwined flows,” IEEE transactions on pattern analysis and machine intelligence, vol. 38, no. 11, pp. 2312–2326, 2016.
8. J. Berclaz, F. Fleuret, E. Tiretken, and P. Fua, “Multi-object tracking using k-shortest paths optimization,” IEEE transactions on pattern analysis and machine intelligence, vol. 33, no. 9, pp. 1806–1819, 2011.
9. L. Zhang, Y. Li, and R. Nevatia, “Global data association for multi-object tracking using network flows,” in Computer Vision and Pattern Recognition, 2008. CVPR 2008. IEEE Conference on, pp. 1–8. IEEE, 2008.
10. H. Pirsiavash, D. Ramanan, and C. C. Fowlkes, “Globally-optimal greedy algo-

11. R. T. Collins, “Multi-target data association with higher-order motion models,” in Computer Vision and Pattern Recognition (CVPR), 2012 IEEE Conference on, pp. 1744–1751, IEEE, 2012.
12. C. Kem, F. Li, A. Cuptadi, and J. M. Rehg, “Multiple hypothesis tracking revisi-

13. I. J. Cox and S. L. Hingorani, “An efficient implementation of real’s multiple hypothesis tracking algorithm and its evaluation for the purpose of visual track-

14. D. Papageorgiou and M. Sálupkas, “The maximum weight independent set prob-

15. S. Oh, S. Russell, and S. Sosy, “Markov chain monte carlo data association for multi-target tracking,” IEEE Transactions on Automatic Control, vol. 54, no. 3, pp. 481–497, 2009.
16 D. Reid, “An algorithm for tracking multiple targets,” IEEE transactions on Automatic Control, vol. 24, no. 6, pp. 843–854, 1979.

17 Y.-m. Song and M. Jeon, “Online multiple object tracking with the hierarchically adopted gm-pf model using motion and appearance,” in Consumer Electronics-Asia (ICCE-Asia), IEEE International Conference on, pp. 1–4, IEEE, 2016.

18 O. Javed, K. Shahzad, Z. Rasheed, and M. Shah, “Modeling inter-camera space–time and appearance relationships for tracking across non-overlapping views,” Computer Vision and Image Understanding, vol. 109, no. 2, pp. 146–162, 2008.

19 B. I. Prosser, S. Gong, and T. Xiang, “Multi-camera matching using bi-directional cumulative brightness transfer functions,” in BMVC, p. 74, 2008.

20 A. Gilbert and R. Bowden, “Tracking objects across cameras by incrementally learning inter-camera colour calibration and patterns of activity,” Computer Vision-ECCV 2006, pp. 125–136, 2006.

21 S. Srivastava, K. K. Ng, and E. J. Delp, “Color correction for object tracking across multiple cameras,” in Acoustics, Speech and Signal Processing (ICASSP), 2011 IEEE International Conference on, pp. 1821–1824, IEEE, 2011.

22 C.-H. Kuo, C. Huang, and R. Nevatia, “Inter-camera association of multi-target tracks by on-line learned appearance affinity models,” in European Conference on Computer Vision, pp. 383–396, Springer, 2010.

23 F. Solera, R. Zou, R. Cucchiara, and C. Tomasi, “Performance measures and a data set for multi-target, multi-camera tracking,” in European Conference on Computer Vision, pp. 17–35, Springer, 2016.

24 W. Chen, L. Cao, X. Chen, and K. Huang, “An equalised global graphical model-based approach for multi-camera object tracking,” IEEE Transactions on Circuits and Systems for Video Technology (TCSVT), 2016.

25 K. Bernardin and R. Stiefelhagen, “Evaluating multiple object tracking performance: The clear mot metrics,” EURASIP Journal on Image and Video Processing, vol. 2008, p. 246309, May 2008.

26 Y. Cai and G. Medioni, “Exploring context information for inter-camera multiple target tracking,” in Applications of Computer Vision (WACV), 2014 IEEE Winter Conference on, pp. 761–768, IEEE, 2014.

27 Y.-G. Lee, Z. Tang, and J.-N. Hwang, “Online-learning-based human tracking across non-overlapping cameras,” IEEE Transactions on Circuits and Systems for Video Technology, 2017.

28 X. Chen and B. Bhanu, “Integrating social grouping for multi-target tracking across cameras in a crf model,” IEEE Transactions on Circuits and Systems for Video Technology, 2016.

29 H. K. W. P. Y. Kim, S.W. Kim and S. Ko, “Improved pedestrian detection using joint aggregated channel features,” in ICEIC, pp. 92–93, January 2016.