Extended MHT Algorithm for Multiple Object Tracking

Long Ying¹, Changsheng Xu¹
¹National Lab of Pattern Recognition, Institute of Automation, Chinese Academy of Sciences, Beijing, China
{lying, csxu}@nlpr.ia.ac.cn

Wen Guo¹,²
²Electronic Engineering Department, Shandong Institutes of Business Technology, Yantai, China
wguo@nlpr.ia.ac.cn

ABSTRACT
In this paper, we propose an improved efficient MHT algorithm integrated with HSV-LBP appearance and repulsion-inertia model for multi-object tracking. Simultaneously tracking multiple objects is critical to video content analysis and virtual reality. The main issues we want to address in this paper are integration of video image patch information into data association and ambiguous observations caused by objects in close proximity. A likelihood function of HSV-LBP histogram with strategy of template updating is constructed. A repulsion-inertia model is adopted to explore more useful information from ambiguous detections. Experimental results show that the proposed approach generates better trajectories with less missing objects and identity switches.

Categories and Subject Descriptors
1.4.8 [Image Processing and Computer Vision]: Scene Analysis – motion, stereo, tracking.

General Terms
Algorithms, Experimentation

Keywords
Multi-object Tracking, Data association, MHT, Patches tracking

1. INTRODUCTION
Tracking is the most fundamental task in computer vision. The ability to simultaneously track multiple objects is critical to video content analysis and virtual reality, which provides important information for higher level analysis and decision.

Multi-object tracking aims at inferring trajectories for each object from video sequence, which can be considered as a spatiotemporal grouping problem. All image regions are classified as detection of specific object or background. In real senses, there are many conditions such as variance of number of objects, similar appearances of different objects, variance of object appearance, complex interactions, long time occlusions and clutter background, which generate the uncertainty and make multi-object tracking a challenging problem.

Several frameworks that incorporate appearance, motion and arising-disappearing models have been proposed to find a joint solution, which make state space of multi-object tracking very complex. To efficiently resolve, the state space is usually restricted to a finite set of candidate locations, either by thresholding the observation likelihood or by regularly discretizing the location space.

Unlike single object tracking which focus on appearance presentation and dynamic search, the core issue in multi-object tracking is data association. Earlier works only used current and past observations to estimate the current state. Joint Probabilistic Data Association Filter (JPDA) [1] based on Kalman filter tried to estimate states for fixed number of objects frame by frame by summing different hypotheses. Label switch easily occurred in crowded scenes. Methods base on particle filter [2, 3] explored multiple hypotheses simultaneously, partly resolving this problem. However, the parameters needed to be carefully set to ensure convergence.

Using both past and future information to estimate the current state is usually more effective to overcome the ambiguities of occlusions, spurious observations and missing observations. An intuitive idea is multiple hypotheses tracking (MHT) [4] which is first used in radar signal processing. Original MHT algorithm enumerates all possible associations in a period of time to find out current global optimal solution. However, this edition will suffer exponentially increasing hypotheses. Various heuristics are developed to overcome this complexity, such as pruning, gating, clustering, N-scan-back logic, and k-best hypotheses [5].

Base on the formulation of multi-object tracking proposed in MHT algorithm, many approaches have been proposed recently. MCMC [6] data association uses Metropolis-Hastings algorithm to construct an irreducible andaperiodic Markov chain, which is employed to efficiently sample from the posterior distribution of state space and obtain approximate optimal solution.

Some approaches construct graph models for global association, converting multi-object tracking to quadratic integer program [7] which is resolved by graph cut, or to integer linear program (ILP) [8] which is approximately solved through LP-relaxation. Others construct network flow models [9, 10], mapping the maximum-a-posteriori (MAP) data association problem into cost-flow network with non-overlap constraint on trajectories, which is solved by network optimization.

Although many new approaches have been proposed, efficient MHT is the most classical approach for data association in tracking by detection framework. It is practical in real-time
tracking with moderate computational complexity, and performs robustly for application.

In this paper, we propose an improved MHT algorithm for multiple pedestrian head-shoulder structures tracking. The main contribution of our work is to adapt motion model for patches tracking, integrate appearance model into probability calculation, and propose a repulsion-inertia model similar to [3] to explore more dynamic information for data association from ambiguous observations when objects are in close proximity.

The rest of the paper is organized as follows: Section 2 gives the overview of our tracking framework; Section 3 reviews the efficient MHT algorithm which is the basis of our method; Section 4 demonstrates our improved MHT algorithm; Section 5 reports experimental results on surveillance video and analysis; We conclude the paper in section 6.

2. OVERVIEW

The tracking framework proposed in this paper concentrated on data association, which includes four steps: (1) construction of appearance descriptor for each observation; (2) track trees growing and likelihood calculation; (3) state recalculation of ambiguous hypotheses; (4) generation of k-best hypotheses and pruning. In the first stage, HSV-LBP histogram is constructed for each observation in newly received video frame. In the second stage, track trees grow according to affinities between trajectories at previous time and observations, and appearance template of new trajectory hypotheses is constructed. In the third stage, trajectory hypotheses associated with ambiguous observations are found, and a repulsion-inertia model is adopted to recalculate states of these hypotheses. The final stage consists of remainder steps of MHT algorithm including clustering, k-best hypotheses, and pruning. This framework is illustrated in Figure 1.

![Figure 1. The framework of proposed method](image)

3. EFFICIENT MHT ALGORITHM

Our work is based on MHT algorithm. To make this paper self-contained, we review the efficient MHT algorithm [5].

3.1 Formulation of Tracking Task

In tracking by detection framework, measurements can be received at every moment. Each measurement may either (1) belong to a previously known object, (2) be the start of a new object, (3) be a spurious measurement. For objects that are not assigned measurements, there is the possibility of (4) termination, e.g. moving out of the field of view and alternatively the possibility of (5) continuation of an object.

Let $Z^k$ be the set of all measurements until time $k$. A single trajectory hypothesis is defined as an ordered list of measurements. An association hypothesis $\Theta^i_k$ is defined as a set of trajectory hypotheses. An additional trajectory $r_0$ contains all spurious measurements.

When measurements $Z(k)$ at time $k$ are received, a particular global hypothesis $\Theta^i_k$ can be derived from certain hypothesis $\Theta^i_{(k-1)}$ at time $k-1$. $\Theta^i_k$ denotes the specific set of assignments of the origins of all measurements received at time $k$ with all the object assumed by the parent hypothesis $\Theta^i_{(k-1)}$, which is defined to consist of $r$ measurements from known objects, $\nu$ measurements from new objects, $\phi$ spurious measurements, and $\chi$ obsolete object. A constraint that one observation originates from at most one object and one object has at most one associated observation is imposed to reduce the size of the search space.

The probability of a global hypothesis can be calculated using Bayes’ rule,

$$P\{\Theta^i_k | Z^k\} = \frac{1}{c} P\{Z(k) | \Theta^i(k), \Theta^i_{(k-1)} Z^{i-1}\} P\{\Theta^i_{(k-1)} | Z^{i-1}\}$$

where $c$ is a normalization constant. Multi-object tracking task is to fine out the MAP hypothesis.

It is assumed that a measurement obeys Gaussian distribution if it is associated with object $j$,

$$N_j = N\left[z_i(k); z_i(k-1), S_j(k)\right]$$

where $z_i(k|k-1)$ denotes the predicted measurement for object $j$, and $S_j(k)$ is the innovation covariance.

If the probabilities of detection and termination of track $j$ are $P^d_j$ and $P^t_j$ respectively, and the numbers of spurious measurements and new objects are assumed to obey Poisson distributed with densities $\lambda_c$ and $\lambda_n$ respectively, the posterior probability of an association hypothesis can be express as:

$$P\{\Theta^i_n | Z^k\} = \frac{1}{c} \lambda_c^n \gamma_j^n \delta_j^n \chi_j^n \prod_{i=1}^{m} \left[ N_j \left( z_i(k) \right) \right]^{\delta_i} \prod_{j=1}^{r} \left( P^d_j \right)^{\delta_j} \left( 1 - P^d_j \right)^{1 - \delta_j} \prod_{j=1}^{\nu} \left( P^t_j \right)^{\delta_j} \left( 1 - P^t_j \right)^{1 - \delta_j} \prod_{j=1}^{\phi} \left( 1 - P^d_j \right)^{\delta_j} \left( 1 - P^t_j \right)^{1 - \delta_j} \prod_{j=1}^{\chi} \left( 1 - P^d_j \right)^{\delta_j} \left( 1 - P^t_j \right)^{1 - \delta_j}$$

where $\gamma_i$, $\delta_j$, and $\chi_i$ are indicator variables defined by:

$$\gamma_i = \begin{cases} 1 & \text{if } z_i(k) \text{ comes from a known object} \\ 0 & \text{otherwise} \end{cases}$$

$$\delta_j = \begin{cases} 1 & \text{if object } j \text{ in } \Theta^i_{(k-1)} \text{ is detected at time } k \\ 0 & \text{otherwise} \end{cases}$$

$$\chi_i = \begin{cases} 1 & \text{if object } j \text{ in } \Theta^i_{(k-1)} \text{ terminates at time } k \\ 0 & \text{otherwise} \end{cases}$$

3.2 Efficient Solution

In practical application, track trees are constructed instead of keeping an explicit hypothesis tree. The computational complexity of MHT algorithm is mainly caused by dramatic increase of combinations. Tracks that do not compete for common
measurements can be partitioned into separate clusters which makes a significantly reductions in number of hypotheses.

Enumerating all possible global hypotheses for optimal solution is impractical. Generating the k-best hypotheses directly can achieve more efficient performance and an approximately optimal solution. This is implemented by optimizing Murty’s algorithm [11] in \( O(kN^2) \).

To reduce computational complexity which increases with accumulation of time, “N-scan-back” pruning which assumes that any ambiguity at time \( k \) is resolved by time \( k+N \) is used. Moreover, hypotheses whose probability is much lower than the best hypothesis can be pruning by a threshold of probability ratio.

4. IMPROVED MHT ALGORITHM

MHT algorithm is widely used in radar and sonar signal processing. When this method is applied in image patches tracking such as pedestrian head-shoulder structures, the affinity metric and dynamic model should be adapted to the task.

Kalman filter is used as discrete-time dynamic of the object in classical MHT, which estimates current motion state, calculates motion affinity and predicts motion state when no observation is received. In pedestrian tracking, linear dynamic model is sufficient. However, additionally scale variables should be integrated into motion states for more discriminative description.

The motion state is defined as \( x = [x, dx, y, dy, w, h]^T \), where \( x \) and \( y \) are the position coordinates, \( dx \) and \( dy \) are velocity components, \( w \) and \( h \) are half width and half height of the detected patch.

4.1 Integrate Appearance Model

Classical MHT only utilizes motion information for data association. However, integrating appearance affinity into video tracking can help to overcome some ambiguous in data association.

Compared with RGB color space, the coupling among three components of HSV is much lower, which generates a better decomposition of color space. We construct a histogram with 110 bins, in which 100 bins for counting joint occurrence of hue and saturation component, and 10 bins for value component. In order to efficiently calculate histograms of detected patches in current image, integral image is used.

Local Binary Patterns (LBP) [12] is a non-parametric kernel which summarizes the local special structure of an image. Moreover, it is invariant to monotonic gray-scale transformations, hence less sensitive to changes in illumination. The discrete occurrence histogram of LBP computed over an image patch is shown to a very powerful texture feature. We using 8-bit LBP operator to get a histogram with 256 bins.

As the HSV feature and LBP feature are independent, we stitch these two vectors together, obtaining a histogram with 366 bins.

Every trajectory hypothesis at previous time maintains a template histogram to calculate appearance likelihood of currently received observations. Bhattacharyya coefficient is used to measure the affinity of two histograms. The templates of derived new trajectory hypotheses which associated with certain observation will be updated according to observation histogram.

Integrated with appearance model, the likelihood between currently received observations and previous trajectories can be defined as:

\[
P(p(z_k | r_i(k))) = (1 - \alpha)N_z(z(k)) + \alpha P_z(z(k+1) | r_i(k))
\]

Because motion likelihood obtained in Kalman filter is probability density while Bhattacharyya coefficient is the cosine of angle between two normalized vectors, a Logistic function should be used to map Bhattacharyya coefficient to confidence of particular object. Another coefficient which corresponds to the Gaussian distribution coefficient should be multiplied, so that the two parts of likelihood can complement each other.

\[
P_a(z_k(k+1) | r_i(k)) = \frac{1}{1 + \exp(-a + b \cdot \text{bha}(\vec{h}_i, \vec{h}_z) - \text{bha}(\vec{h}_i, \vec{h}_z))}
\]

where \( \text{bha}(\vec{h}_i, \vec{h}_z) \) is Bhattacharyya coefficient between template histogram and observation histogram; \( a \) and \( b \) are parameters of Logistic function.

4.2 Repulsion-Inertia Model

Unlike points tracking, image patches in close proximity may cause ambiguous observations due to occlusions. For example, observations of two objects may merge or one observation will miss, as shown in Figure 2. In these cases, the Bhattacharyya coefficient will decrease and fall in steep interval of Logistic function. The histogram template of involved new trajectory hypotheses should not be updated. To overcome these ambiguities, trajectory prediction is important.

![Figure 2. Ambiguous observation and association model](image)

However, because of low detection rate and consistent occlusions, trajectory prediction usually fails. Exploring more information from ambiguous detections may help to achieve better effect. We propose a repulsion-inertia model to estimate motion state of hypotheses associated with common ambiguous detection, which is inspired by [3].

When trajectories compete for common observation and the overlap rate of estimated object patches is above a threshold, it is assumed that there are repulsive forces between observations corresponding to estimated states, preventing error merging. To obtain new trajectory hypotheses, interactive observations of observation \( z_i \) (observation of object \( i \) at time \( t_i \)) should be considered, denoted as \( z_i^t \). A new density \( p(z_i^t | x_i, z_i^t) \) is introduced to characterize the interaction among observations.

\[
p(x_{n1}^t | z_i^t, z_i^{t+1}) = k \cdot p(z_i^t | x_i^t) \cdot p(x_{n1}^t | z_i^{t+1})
\]

Particle filter is used for each trajectory to get MAP estimate. The posterior density is characterized by set \( \{x_{n1}^t, n = 1...N_t\} \), where \( \{x_{n1}^t, n = 1...N_t\} \) is a set of support particles with associated
weight $w_i^\alpha$. However, rough estimate using just fewer particles is enough.

$$w_i^\alpha \propto w_i^\alpha \exp \left( -\frac{1}{\alpha} \cdot \frac{d_i^2}{\sigma_i^2} \right)$$

where $d_i$ is a normalization term, $\sigma_i$ is prior constant, and $d_i$ is the distance between observation of current particle and the interactive observation $z'$. Inertia model is constructed to make trajectories tend to keep original dynamic states, preventing object identities switching.

$$p(x_i^\alpha | x_i^{t-1}^\alpha, z_i^{t-1}) = p(x_i^\alpha | x_i^{t-1}^\alpha) \cdot \phi_i(x_i^{t-1}^\alpha, x_i^{t-1}^\alpha)$$

where $\alpha$ is a normalization term, $\sigma_i$ and $\sigma_i$ are prior constants, $\psi_i = x_i^\alpha - x_i^{t-1}^\alpha$, $\psi_i = x_i^\alpha - x_i^{t-1}^\alpha$; $\theta_i$ is the angle between $\psi_i$ and $\psi_i$. For trajectories which compete for common ambiguous observation currently, interactive virtual observations are located recursively according to states estimated in the last iteration. For each two of these trajectories, states are estimated by taking repulsion and inertia process iteratively for several times. When all pairs of trajectories have been calculated, optimal states of trajectories associated with ambiguous observation are obtained, which extract more information from original observation. This process is illustrated in Figure 3. However, if motion or appearance likelihood of an estimated trajectory hypothesis is too low, it will be deleted from track trees.

**Figure 3. State estimate by repulsion-inertia model**

### 5. EXPERIMENTS

In order to demonstrate the effectiveness of our approach, two surveillance video sequences from TRECVID 2011, which are taken in an airport hall, with solution 720 x 576, containing frequently occlusion of moving or static travelers with luggage, are used in the experiment. The ground truths of these videos are available.

#### 5.1 Experimental Setup

Multi-object tracking is a practical systematic project which takes videos as input, automatically obtains trajectories of objects preliminarily defined. Our work concentrated on data association, hence only a public head-shoulder structure detector based on HOG feature and adaboost classifier is used. The detection result is unsatisfactory, having large amount of missing observations in consistent frames and some spurious observations caused by luggage, which constrains the performance of tracking algorithm.

There are many parameters in proposed algorithm. (1) Prior constants of tracking framework, such as $\lambda_a$, $\lambda_i$, $P_a$ and $P_i$; (2) Parameters of Kalman filter, such as process covariance, measurement covariance and initial state covariance; (3) Appearance model parameters, such as likelihood weight and parameters of Logistic function; (4) Parameters of repulsion-inertia model. These parameters are relevant to the video sense and the behavior and appearance of objects, rough range of which can be estimated by priori knowledge. Then we experimentally set the parameters in several stages, which remain identical for all video sequence.

Our approach is implemented in C++ with OpenCV library. On i5 3.10GHz PC with 4.00GB RAM, tracking can be executed in real time (25fps) for about 10 image patches without parallel processing.

#### 5.2 Tracking Result and Analysis

We adopt the most commonly used CLEAR metrics [13] for evaluation. CLEAR metrics consist of MOTP which calculates the ratio of intersection over union of bounding boxes, rate of misses, rate of false positives, rate of identity switches, and MOTA which evaluates the overall situation of object identities by considering misses, false positives and identity switches.

![Figure 4. Tracking Results 1: sample frames of head-shoulder detection (first row), points MHT (second row), and patches MHT with appearance (third row).](image-url)
Patches MHT with appearance performs much better than points MHT. In most cases, this approach correctly keeps the identities of trajectories when intersections occur, such as trajectory 1, trajectory 5 and trajectory 8. It also correctly associates some missing observations, such as trajectory 8. However, due to unsatisfactory detection, the performance of data association is constrained.

Table 1. Performance of head-shoulder structures tracking

<table>
<thead>
<tr>
<th>Method</th>
<th>MOTA</th>
<th>Miss</th>
<th>False</th>
<th>Switch</th>
<th>MOTP</th>
</tr>
</thead>
<tbody>
<tr>
<td>Points MHT</td>
<td>0.1628</td>
<td>0.7562</td>
<td>0.0736</td>
<td>0.0073</td>
<td>0.6144</td>
</tr>
<tr>
<td>MHT with appearance</td>
<td>0.1770</td>
<td>0.7528</td>
<td>0.0639</td>
<td>0.0063</td>
<td>0.6103</td>
</tr>
<tr>
<td>MHT with RI model</td>
<td>0.1780</td>
<td>0.7542</td>
<td>0.0629</td>
<td>0.0048</td>
<td>0.6113</td>
</tr>
</tbody>
</table>

Ambiguous observations in tracking by detection framework are of special concern. Repulsion-inertia model is integrated into patches MHT algorithm to address identity switches caused by ambiguous observations. Qualitative results of the proposed method and patches MHT without repulsion-inertia model are contrastively shown in Figure 5. It is obvious that the proposed method greatly reduces identity switches. For example, trajectory 11 in the second row of Figure 5 is successfully formed, contrasting with trajectories 10 and 11 in the first row. So is trajectory 19 in the second row of Figure 5, contrasting with trajectories 20 and 21 in the first row.

Figure 5. Tracking Results 2: sample frames of patches MHT with appearance (first row) and patches MHT with repulsion-inertia model (second row)

Meanwhile, quantitative evaluation is also taken using CLEAR metrics, as shown in the third rows of Table 1. The rate of identity switches declines obviously. However, the rate of misses increases slightly due to additional interaction in certain complex situations of unsatisfactory detection. The overall metrics such as MOTA and MOTP slightly increase, which means better performance of the proposed method.

6. CONCLUSION

We have proposed an improved MHT algorithm which incorporates HSV-LBP appearance and repulsion-inertia model for image patches tracking. In our approach, appearance affinity of image patch is integrated with motion affinity to calculate hypotheses likelihood. Ambiguous observations in tracking by detection framework are of special concern. Repulsion-inertia model are used to make trajectories competing for common observations repel each other and tend to keep original dynamic states. Experimental results demonstrate the effectiveness and prospect of our approach. However, due to unsatisfactory detection, the performance of data association is constrained. Our future work is to couple object detection with data association for better performance in surveillance video.

7. REFERENCES