Robust Multi-Objects Detection and Tracking Algorithm under Complex Circumstance

Yumin Tian, Mingqian Tang, Ankui Meng and Yuejiang Li
School of Computer Science and Technology, Xidian University, Xi’an, Shaanxi, China

Abstract - A MSPF algorithm based on multi-feature fusion and an inferring match algorithm applied to multi-objects tracking are proposed in order to tackle the problems such as the inter-object occlusion, occlusion by background obstacles and objects deformation, which are the key factors that decrease the tracking accuracy. The brief way in which the algorithm works is as follows: represent objects by multi-feature fusion of color and texture features; establish MSPF motion model to estimate the possible locations of objects; decide if there are splits, merges and other complicated motions among objects through inferring matching with detection and tracking information; use cost function to solve the correspondence problem after splits of objects. Experimental results show that this advanced algorithm has the advantages of functioning effectively for the detection and tracking of multiple moving objects under complicated situations while meeting real-time demand.

Keywords: MSPF, multi-feature fusion, inferring match, multi-objects tracking, occlusion

1 Introduction

Detection and tracking of multiple objects is an integrated field of the study of computer vision. A robust multiple objects detection and tracking system should be capable of tackling the complex movement situations among multiple targets, such as inter-object occlusion, occlusion by background obstacles and objects deformation etc. In consideration of fulfilling these requirements, much endeavor have been done by many researchers and great achievements have been made. Xin Li et al. [1] use Kalman filter to address the problem of correspondence among moving objects. Zhong Xiaopin et al. [2] solve the multi-feature problem of particle filter by adopting an adaptive multiple-cue fusion strategy. Nguyen et al. [3] use Kalman filter to set up motion model in a multi-objects moving distributed surveillance system. Medioni et al. [4] propose a multiple-moving-objects-tracking method that based on graph theoretic algorithm. Changjiang Yang [5] uses mean-shift method to track multiple objects. Although the methods mentioned above can be robust to a certain degree aiming at some kinds of disturbance in the process of detection and tracking of multiple objects, varies other conditions may still affect the tracking and matching performance. Therefore, a trajectory prediction method based on particle filter and an inferring match method applied to multiple objects tracking are proposed in this paper. The more advanced method fuses color and texture features, establishes multi-objects trajectory predictor by mean-shift based particle filter, updates the state of objects through the inferring matching of features, and resolves the complex motions among objects such as the appearance/disappearance of objects, merges and splits etc.

2 Framework of detection and tracking of multi-objects

To precisely detect all the objects in the scene, the background substraction is adopted. After obtaining the foreground blobs in current video sequences, we denote each independent object’s boundary rectangle $B_m^n$ by measuring its contour’s centroid, width and height.

We assume that the N foreground object blobs are denoted by $\{R_m^n\}_{m=1}^N$ and M objects are denoted by $\{O_m^n\}_{m=1}^M$ at time t in video sequences I(x, y, t), and the status of mth object $O_m^n$ includes boundary rectangle $B_m^n$, color histogram $C_m^n$, textual histogram $E_m^n$, cue degenerate matrix $CU_m^n$, suspending flag $BE_m^n$, lifetime $L_m^n$. Hence $O_m^n = B_m^C, C_m^n, E_m^n, CU_m^n, BE_m^n, L_m^n$, $m = 1, 2, \ldots, M$.

Mean-shift based particle filter is taken as motion predictor to analyze the color histogram $C_m^n$, textual histogram $E_m^n$ and figure features of the objects, so as to predict the possible locations of objects and contribution degree changes of each feature precisely and efficiently, and obtain the list of tracked objects.

With inference matrix analysis and object’s cue degenerate analysis of predicted objects list and tracked objects list, we can infer many different motion situations such as the appearance/disappearance of objects, merges and splits with respect to the similarities among objects and the changes of the features. Finally, the update of the states of objects can be accomplished according to motion situations.
3 Multi-Feature Fusion Based Trajectory Prediction

3.1 MSPF

With the nonparametric Monte Carlo sampling methods, particle filter can realize Bayesian filtering[6,7]. Mean-shift based particle filter improves the weights selecting method of the classic particle filter [10]. We firstly predict the obtained particles according to prior probability distribution. Then the optimal location of particles can be obtained through iterations. With a comparison of the similarity between the shifted particles and the obtained values, more precise weights can be achieved according to the observation probability after the multi-feature fusion. In this paper, we choose high robust features and similarity calculating function and adopt multi-feature fusion to automatically update the validity of each feature and to enhance the robustness of the tracking of objects.

3.2 Adaptive Multi-feature Fusion

3.2.1 Object feature and Its Similarity Function

Color is always an important information. For the purpose of adapting to any situations, this paper applies RGB spatial weighted histogram, as HSV histogram is unsuitable when color becomes too dim.

Let L be quantization scale, then the weighted histogram of the i-th particle in the corresponding region can be defined as:

\[
p^{(i)}(\tau) = \frac{1}{s} \sum_{u} k\left(\frac{\|\tau - u\|}{h}\right) \delta(\psi(\tau, -u))
\]

Here, \( k\left(\frac{\|\tau - u\|}{h}\right) \) is normalizing factor; \( s \) represents the number of pixels inside the region; \( \tau \) represents pixel in region center; \( \tau_i \) represents a certain pixel point in the region; \( \|\tau - \tau_i\| \) represents the distance between \( \tau \) and \( \tau_i \); \( h \) is the kernel band width; \( k(\cdot) \) is kernel function, and the smaller the distance between pixel and the region center, the higher weight the function assign to the corresponding color feature, thus reducing disturbance which may come from background pixel points in the surrounding regions; \( \delta \) is Kronecker cofactor function; \( \psi(\tau, \tau_i) \) represents color order at \( \tau \); \( u \) is color histogram order, with the range \( 1, 2, \ldots, s \).

Let \( q^{(i)}_0 \) be the object color histogram model and \( p^{(i)}_0(\tau) \) be the corresponding histogram of i-th color particle with the center of \( \tau \) at time t. And its corresponding color-feature observation probability is:

\[
p(\tau_i | x_{color}^{(i)}(t)) = \sum_{u} \sqrt{q^{(i)}_0(u)} p^{(i)}_0(\tau)(t)
\]

And the updated weight of i-th color particle will be:

\[
w_{color}^{(i)} = w_{color}^{(i)-1} p(\tau_i | x_{color}^{(i)}(t))
\]

Because of the complimentary advantage of texture and color, textual information can still tell the similarity of objects when the color cue degenerates. With Sobel operator, texture information and its histogram of objects can be obtained. With the same calculation as color feature, we can also obtain the corresponding observation probability \( p(\tau_i | x_{edge}^{(i)}) \) and weight \( w_{edge}^{(i)} \) \( t \) of i-th edge particle at time t.

3.2.2 Adaptive Multi-features Fusion

As tracking features of objects, color histogram is one of the most direct distinctive features, because of its simple calculation and insensitivity to deformation, rotation, size change and partial occlusion of objects. Particularly, edge histogram possesses a better separability when background and objects share the similar color features or when color cue degenerates due to light variation. Therefore, this paper fuses the two feature particles according to their complementarity to complete state estimation of moving objects.

State estimation of moving object \( s \) at time t:

\[
X_t^{(s)} = \sum_{j=1}^{N_{color}} \delta_{(i)}^{(s)}(X_t^{(s)} - X_i^{(s)}) + \sum_{j=1}^{N_{edge}} \delta_{(i)}^{(s)}(X_t^{(s)} - X_i^{(s)})
\]
Here, \( N_{\text{color}} \) and \( N_{\text{edge}} \) represent particle number of color feature and edge feature respectively; \( x_{\text{color}}^{(i)} \), \( x_{\text{edge}}^{(i)} \) represents the state of i-th color particle at time \( t \); \( y_{\text{color}}^{(j)} \), \( y_{\text{edge}}^{(j)} \) represents the state of j-th edge particle at time \( t \); \( \delta_{t-1,k}^{(s)} + \delta_{t-1,k}^{(e)} = 1 \)

The contribution degree function \( \delta_{t-1,k}^{(s)} \) of kth feature of object s at time t is updated as below:

\[
\delta_{t-1,k}^{(s)} = \frac{\delta_{t-1,k}^{(s)} \cdot d_{t-1,k}^{(s)}}{\sum_{j=1}^{N_{k}} \delta_{t-1,k}^{(j)} \cdot d_{t-1,k}^{(j)}} \tag{5}
\]

\[
d_{t-1,k}^{(s)} = \frac{p(z_{t-1} | x_{t-1,k}^{(s)})}{p(z_{t-2} | x_{t-2,k}^{(s)})} \tag{6}
\]

\( N_{k} \) represents the number of object features referenced, and in this paper it is 2; \( d_{t-1,k}^{(s)} \) reflects effectiveness variation of feature cue. When one feature cue degenerates, its feature contribution degree will decrease correspondently. And the contribution degree of each robust feature shall be increased for tracking to improve the accuracy of particle filter estimation and the adaptivity.

### 4 Strategies of Inferring Match

In real application, besides the most basic motion situation of simple, isolated and complete single object, tracking algorithm may need to deal with several other situations such as appearance of new objects, disappearance of old objects, object deformation, object occlusion by background obstacles, merges and splits of multiple moving objects. This paper would deal with all the situations above based on the results of match with tracking object.

#### 4.1 Fundamental Structure of Inferring Match

Three structures are needed in inferring match in the process of moving objects tracking: inference relationship matrix, cue degeneration matrix and object suspending flag.

#### 4.1.1 Inference Relationship Matrix

Inference relationship matrix \( P \) describes most basic matching relation between objects tracking list \( \{T_{\text{ref}}^{(m)}\}_{m=1}^{M} \) and detection list \( \{R_{\text{det}}^{(n)}\}_{n=1}^{N} \) at time t. Matrix \( P \) is \( M \times N \), and its elements is determined by the similarity of overlapped area between \( T_{\text{ref}}^{(m)} \) and \( R_{\text{det}}^{(n)} \). Element \( P(m,n) \) can be calculated by the formula as follow:

\[
P(m,n) = \begin{cases} 
\frac{\text{Area}(T_{\text{ref}}^{(m)} \cap R_{\text{det}}^{(n)})}{\text{min} \left( \text{Area}(T_{\text{ref}}^{(m)}), \text{Area}(R_{\text{det}}^{(n)}) \right)} & \text{if } \delta_{\text{color}}^{(m,n)} + \delta_{\text{edge}}^{(m,n)} \geq \gamma \\
0 & \text{else}
\end{cases}
\]

\( \gamma \) is the threshold parameter obtained from experiments; \( \text{Area}(T_{\text{ref}}^{(m)} \cap R_{\text{det}}^{(n)}) \) is the overlapped area between \( T_{\text{ref}}^{(m)} \) and \( R_{\text{det}}^{(n)} \); \( q(x_{\text{color}}^{(m,n)}) \) and \( q(x_{\text{edge}}^{(m,n)}) \) respectively represent the color histogram and edge histogram distance between objects \( T_{\text{ref}}^{(m)} \) and \( R_{\text{det}}^{(n)} \), and its calculation is similar to that of the observation probability in particle filter module; \( \delta_{\text{color}}^{(m,n)} \) and \( \delta_{\text{edge}}^{(m,n)} \) are feature contribution degree functions of object \( T_{\text{ref}}^{(m)} \) and object \( R_{\text{det}}^{(n)} \) matches.

Besides, row accumulated value \( P_{m} \) and column accumulated value \( P_{n} \) of matrix \( P \) are calculated as follow:

\[
P_{m} = \sum_{n=1}^{N} p(m,n) \tag{8}
\]

\[
P_{n} = \sum_{m=1}^{M} p(m,n) \tag{9}
\]

\( P_{m} \) represents the number of detected objects which match with tracking object \( T_{\text{ref}}^{(m)} \); \( P_{n} \) represents the number of tracked objects which match with detected object \( R_{\text{det}}^{(n)} \). Then based on the results of \( P(m,n) \), \( P_{m} \) and \( P_{n} \), we can deduce according to Fig.3:

![Fig.3 Inference of the Relationship Matrix](image)

C1 is the tracking and detection information of the current frame. If \( P_{n} = 0 \), it can be inferred that no matching \( R_{\text{det}}^{(n)} \) is detected for object \( T_{\text{ref}}^{(m)} \), which indicates \( T_{\text{ref}}^{(m)} \) exits from the scene; If \( P_{n} > 1 \), it means splits happen among multiple objects, and the split may be caused by partial occlusion by background obstacles, e.g. railings and trees or separation among multiple merging objects. Then according to the result of \( P_{n} \), it can be inferred as follows: \( P_{n} = 0 \) means no matching \( T_{\text{ref}}^{(m)} \) is detected for object \( R_{\text{det}}^{(n)} \), which indicates \( R_{\text{det}}^{(n)} \) enters as a new object; \( P_{n} = 1 \) means \( T_{\text{ref}}^{(m)} \) is an isolated moving object and matches with \( R_{\text{det}}^{(n)} \); \( P_{n} > 1 \) means several single moving objects merge together and a larger merged object is formed, and occlusion always exists among those moving objects under this circumstance.
4.1.2 Cue Degeneration Matrix

Cue degeneration matrix $CU_{tm}^n$ is established to represent cumulative degeneration degrees of $N_t$ cues of object $T_t^n$ at time $t$, and the object is updated through the formula below at $t-1$ frame.

$$CU_{t-1,k} = \begin{cases} CU_{t-1,k} + 1, & \text{if } \frac{p(x_{t-1,k} | x_{t-1,k})}{p(x_{t-2} | x_{t-1,k})} > \eta \\ 0, & \text{else} \end{cases}$$

(10)

Here, $p(x_{t-1,k} | x_{t-1,k})$ is the observation probability of the $k$th feature at time $t-1$; threshold $\eta$ is obtained from experiment; if $CU_{t-1,k} = 0$, it indicates each feature cue of object remains the same and can stably represent the moving state of object; if $CU_{t-1,k} \geq \lambda$ (where $\lambda$ is accumulating threshold value for reference), it indicates certain feature cue of object degenerates in $\lambda$ sequential frames. According to the cue matrix $CU_{tm}^n$ of each tracking object $T_t^n$ at time $t$, we can deduce as follows:

If all $CU_{t-1,k} = 0$, it means isolated object is in simple motion, and all feature cue can stably represent the state of object; if there exists any $CU_{t-1,k}$ not equal to zero and not any $CU_{t-1,k}$ is greater than $\lambda$, it infers that partial feature cue degenerates, which means possible deformation of object. Then we need adjust each feature's contribution degree function dynamically and automatically to improve the contribution degree of effective feature; if all $CU_{t-1,k}$ are not less than $\lambda$, it infers that all the cues of object degenerate in continuous $\lambda$ frames, which means an occurrence of occlusion which may be occlusion by static background obstacles or mutual occlusion among multiple objects.

4.1.3 Object Suspending Flag

Object suspending flag $BL_{t}^m$ is established to indicate whether all cues of object $T_t^n$ is valid at time $t$. If $BL_t^m = TRUE$, this means all the cues of object are invalid, indicating the object exits the scene or the possible occurrences of partial occlusion and mutual occlusion of multiple objects. If $BL_t^m = FALSE$, this means objects behave normally and all the cues of object are valid or partially valid. If all $CU_{tm}^n$ of the tracking object are not less than $\lambda$, this means all the feature cues are invalid, and thus the object suspending flag is set TRUE. Meanwhile, object’s lifetime $L_t^n$ decreases 1 each frame. If object lifetime $L_t^n \geq \beta$, this means object does exit from the scene and this object can be removed from the tracking list. If object lifetime $L_t^n < \beta$, this means feature cues are partially resumed, thus the object suspending flag is set False and the tracking of object is continued.

4.2 Comprehensive Inferring Match of Objects

The three basic inferring structures above can indicate a certain moving situation of object to some extend, but none of which can independently analyze kinds of complicated situations completely. A comprehensive inferring match method is proposed in this paper by taking the advantage of the three structures above to correctly analyze the complex motions of the moving objects.

We establish the inference relationship matrix $P$ between the tracked objects list $\{T_t^n\}_{n=1}^m$ and the detected objects list $\{R_t^n\}_{n=1}^n$ at time $t$, cue matrix $CU_{tm}^n$ of each object, suspending flag $BL_{t}^m$ of object and lifetime $L_t^n$. Then all complicated motions can be handled by the following deduction.

On the basis of Fig.3, we further infer the motions according to the inference chart as is illustrated in Fig.4. Seven complex motions of the objects can be figured out with the detection and tracking information. And the analysis of the corresponding situations is shown as the following.
Objects exit from the scene: $P_e = 0$ indicates the disappearance of object $T_n$ or the complete occlusion of it, otherwise, we suspend the objects, thus $BL_n = \text{TRUE}$, $L_n = L_n + 1$.

Objects enter the scene: here $P_e = 0$ and the information of $R_n$ is inserted into the list of tracked objects.

Merges of multiple objects: here $P_e > 1$ and multiple independent moving objects merge together, forming a larger object. Under this circumstance, there always exists mutual occlusion between moving objects; then suspended object $T_n$, $BL_n = \text{TRUE}$, update the edge matrix of $T_n$ through the velocity information of $R_n$ without updating other features.

Simple motion of isolated objects: here $P_e = 1$, $P_s = 1$, $BL_n = \text{FALSE}$, $\forall CU^m_n = 0$, $T_n$ matches with $R_n$, update and correct feature information of $T_n$ with $R_n$.

Splits of multiple objects: here $P_e > 1$ and $BL_n = \text{TRUE}$, it means splits of multiple objects happen, and this may be the partial occlusion by background obstacles, e.g. railings and trees or splits among multiple moving objects which merge together, and these objects split has been tracked before. If $P_e > 1$ and $BL_n = \text{FALSE}$, it means splits of multiple objects happen and the new objects split has not been tracked before. The information of $R_n$ is inserted into the list of tracked objects as a result.

Occlusion: here $P_e = 1$, $P_s = 1$, $BL_n = \text{FALSE}$, $\exists CU^m_n = 0$, $\neg \forall CU^m_n \geq \lambda$, this means partial occlusion or deformation of objects is undergoing. As the degree of occlusion or deformation is slight, the feature information of tracking is still valid. Then update and correct feature information of $T_n$ with $R_n$; if $P_e = 1$, $P_s = 1$, $BL_n = \text{FALSE}$ and $\forall CU^m_n$, this means the occlusion by the static background obstacles happens. Thus update and correct edge matrix information of object through $R_n$ without updating other features. Meanwhile, $BL_n$ is set TRUE and object lifetime $L_n$ increases 1 each frame.

Resume tracking after occlusion: here $P_e = 1$, $P_s = 1$ and $BL_n = \text{TRUE}$, this means $R_n$ is the object once occluded by background obstacles, known as $T_n$. As the object appears again, its tracking can be resumed. Then feature information of $T_n$ can be updated and correct with $R_n$ and the tracking of $T_n$ resumed and $BL_n$ is set FALSE.

Through the above inferring match strategy, the match and update of complex motions can be completed. In order to describe the motions accurately and sufficiently, two specific complicated motions will be analyzed by this paper as below.

Fig. 5 is the instance of the occlusion by static background obstacles. When the moving object $A$ is occluded by door the background obstacle, the cues of the object degenerate and the value of $CU^m_n$ begins to increase. When the object is completely occluded by the door, $CU^m_n \geq \lambda$ and $P_e = 0$ can be observed. Then the object is suspended, and $L_n$ increases one each frame. When the object appears again, the correction and update of the feature information of $T_n$ is completed with $R_n$. Thus the tracking of the object resumes eventually.

Fig. 6 is the instance of mutual occlusion among moving objects. Object $A$ and object $B$ merge together to a single target while moving and $P_e > 1$ is observed. Then objects $A$
and B are suspended. After the two objects split, $P_{t+1}$ and $BU_t = TRUE$ shall be observed. With the observed information, their features are updated accordingly and the tracking of object A and B is resumed.

5 Conclusion

In this paper, an effective algorithm of multi-objects tracking is proposed based on particle filter and inferring matrix. The algorithm is capable of predicting the trajectory of moving objects through MSPF and feature fusion, and adaptively correcting the multi-feature contribution degree of moving objects. With the analysis of inference relationship matrix and object cue degeneration matrix, precise matching of any complicated movements of objects can be achieved. Experimental results indicate that the algorithm has a stable tracking performance of multiple moving objects in complicated surveillant scenes.

6 Acknowledgments

This research was partially supported by the Natural Science Foundation of Shaanxi Province(Grant No.2010JM8011).

7 References


