Novel Seed Selection for Multiple Objects Detection and Tracking

Zailiang Pan and Chong-Wah Ngo
Department of Computer Science
City University of Hong Kong
Tat Chee Avenue, Kowloon, Hong Kong
{zerin, cwngo}@cs.cityu.edu.hk

Abstract

This paper proposes a unified approach for initializing, detecting and tracking of multiple moving objects. Object initialization is achieved through novel seed selection which is adaptively activated, depending on the quality of tracking, to select the best possible frames along the temporal direction for object detection. EM algorithm is then employed to robustly segment and detect multiple objects in a selected frame. Each detected object is represented by an appearance-based model and mean shift tracking procedure is adopted to rapidly and effectively track the target objects.

1. Introduction

The effective initialization, detection and tracking of multiple objects in a sequence is a challenging task. Popular approaches include energy minimization [5], condensation [3] and mean shift tracking [1, 2]. Very often, manual initialization of object locations is necessary for these approaches, in particular when camera motion exists. In this paper, we propose an automatic motion-based approach for object initialization through novel seed selection (NSS). NSS is a procedure to search for the best possible frame, not necessary the first frame, in a sequence to start object detection and tracking. This is motivated by the fact that not every frame is appropriate for object localization since 1) some objects may cease moving in some frames; 2) some objects may be occluded; 3) some objects may have same moving direction as camera motion; 4) some frames may not be stable due to camera shaking artifacts.

NSS is achieved through motion discriminant analysis, more specifically, NSS looks for seeds (or frames) that contain the most distinctive motion clusters. Once a seed is located, EM algorithm is initialized to detect and segment possible objects in the frame. Mean shift tracking procedure is then employed to rapidly track multiple detected objects in both temporal forward and backward directions. In our framework, the process of initialization, detection and tracking is represented as a finite state machine (illustrated in Fig 1). Given an image sequence, NSS is adaptively activated whenever the quality of tracking starts to degrade, for instance, due to occlusion.

2. Motion Discriminant Analysis

$3D$ structure tensor is used for motion representation. The computed optical flows and their fidelity measures are utilized directly for motion clustering and discrimination.

2.1. 3D Tensor Representation

Let $I(x, y, t)$ be the intensity of a point in 3D image volume. By assuming point intensity remains constant in a short time. A constraint condition can be derived as

$$\frac{dI}{dt} = \frac{\partial I}{\partial x}u + \frac{\partial I}{\partial y}v + \frac{\partial I}{\partial t} = \epsilon \tag{1}$$

where u and v represent the local spatial velocity along the x and y coordinates respectively. ϵ is assumed to be zero-mean Gaussian noise. The total sum of ϵ^2 over a 3D image volume R can be represented as

$$E = \sum \epsilon^2 = V^T \left(\sum_{x,y,t \in R} (\nabla I)(\nabla I)^T \right) V \tag{2}$$

where $V = [u, v, 1]^T$ and $\nabla I = \left[\frac{\partial I}{\partial x}, \frac{\partial I}{\partial y}, \frac{\partial I}{\partial t}\right]^T$. The central term, which is a symmetric tensor representation of the local structure of R, has the form

$$\begin{aligned} \Gamma &= \begin{bmatrix} J_{xx} & J_{xy} & J_{xt} \\ J_{yx} & J_{yy} & J_{yt} \\ J_{tx} & J_{ty} & J_{tt} \end{bmatrix} \\ J_{mn} &= \sum_{x,\,y,\,t \in R} \frac{\partial I}{\partial m}\frac{\partial I}{\partial n} \quad m, n = x, y, t \end{aligned} \tag{3}$$

Given the tensor representation in Eqn (3), the optical flow $\mathbf{v} = [u, v]^T$ can be estimated by minimizing the cost function E in Eqn (2). The diagonal components of a tensor which represent the intensity variation in spatio-temporal coordinate can be exploited for fidelity measure. Thus, our proposed fidelity term λ, which depicts the certainty of estimated optical flow in R, is defined as

$$\lambda = 1 - \frac{E}{E + J_{xx} + J_{yy}} \tag{4}$$

Proceedings of the 17th International Conference on Pattern Recognition (ICPR'04)
1051-4651/04 $ 20.00 IEEE

The fidelity term has following favorable properties: 1) It is maximal for ideal flows, i.e., $E = 0$; 2) It is minimal if no spatial intensity variation, i.e., $J_{xx} + J_{yy} = 0$; 3) Its value is normalized in the range $[0, 1]$.

2.2. Motion Clustering and Discriminant Analysis

Given the flows $\{\mathbf{v}_i\}$ and their fidelities $\{\lambda_i\}$ at time t as described in Section 2.1, we adopt k-mean algorithm with robust estimator for outlier-tolerated clustering:

1. Choose an initial classification $\{u_{ij}\}$, where $u_{ij} = 1$ if $\mathbf{v}_i$ belongs to j^{th} group and $u_{ij} = 0$ otherwise.
2. Calculated class probability $P_j(k)$, sample means $M'_j(k)$ and covariance matrices $\Sigma'_j(k)$ from samples $\mathbf{v}_i$ weighted by fidelities λ_i at k^{th} iteration.
3. Recalculate sample means $M_j(k)$ and covariance matrices $\Sigma_j(k)$ from $\mathbf{v}_i$ and λ_i which satisfy the constraint $(\mathbf{v}_i - M'_j)^T \Sigma_j^{'-1} (\mathbf{v}_i - M'_j) < c \times \sigma_j$, where σ_j is estimated by robust estimator as $\sigma_j = 1.4826 \times$ median$(\mathbf{v}_i - M'_j)^T \Sigma_j^{'-1} (\mathbf{v}_i - M'_j)$, and $c = 2.5$ is a empirical parameter.
4. Reclassify every $\mathbf{v}_i$. If there is any change in the class label of $\mathbf{v}_i$, repeat steps 2 to 4. The classification of $\mathbf{v}_i$ is based on $\min_j \{(1/2)(\mathbf{v}_i - M_j)^T \Sigma_j^{-1} (\mathbf{v}_i - M_j) + (1/2) \ln |\Sigma_j| - \ln P_j\}$.

Class separability is utilized to determine novel seed selection. The more separable the classes are, the more likely the objects can be detected. The class separability is defined as

$$\begin{aligned} M &= tr(S_w^{-1} S_b) \qquad (5) \\ S_w &= \sum_{j=1}^{g} P_j \Sigma_j \\ S_b &= \sum_{j=1}^{g} P_j (M_j - \sum_{k=1}^{g} P_k M_k)(M_j - \sum_{k=1}^{g} P_k M_k)^T \end{aligned}$$

The initial number of cluster is set as $g = 10$. Merging of clusters is desirable if any two classes, say 1 and 2, satisfy the following constraint.

$$tr(S_w^{-1} S_b) = P_1 P_2 (M_2 - M_1)^T S_w^{-1} (M_2 - M_1) < s \qquad (6)$$

where s is an empirical parameter.

3. Seed Point Selection

Based on Eqn (5), good starting points (or seeds) are adaptively located along the temporal dimension. Seed point selection is represented as a finite state machine as shown in Figure 1. The transition among states is based on the degree of matching, occlusion and motion intensity as follows:

Seed select. A state is transferred to *seed select* to initialize object location or when the quality of tracking starts to degrade. The selection criteria of a seed is based on the mean and variance of class discriminant determined by $M > E(M) + \alpha \times Var(M)$, where α is an empirical parameter. The object detection (in Section 4) will be triggered to estimate new object layers.

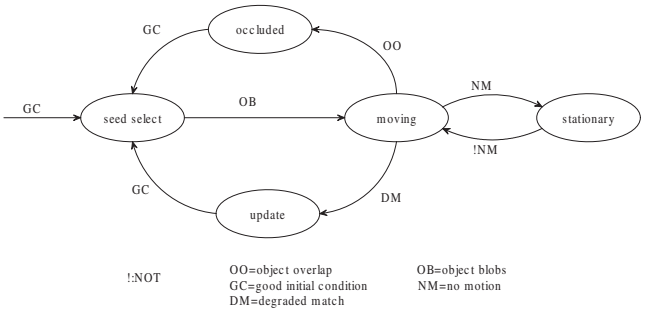


Figure 1. State transition diagram

Object moving. This state is transited from *seed select* whenever object blobs are detected. The degree of object occlusion and matching (see Section 5) will be calculated to determine the change of state.

Object occluded. Object occlusion is detected if any two and more object blobs have large degree of overlap. The trackers of those occluded objects are deleted temporarily, and will be re-emerged by backward tracking from the next seed point.

Object update. The target candidate template will change gradually over time during tracking. If the matching score (defined as distance function in Section 5) degrades, the corresponding tracker will be removed and recovered by backward tracking from the next seed point.

Object stationary. An object remains in stationary state if no motion is detected.

4. Object Detection

EM algorithm is employed for the detection of objects at the selected seed points. The segmentation priors and motion parameters (computed as in Section 2.2) associated with seed points are used to initialize EM. Our EM algorithm is similar to [6] by Sawhney and Ayer, except the segmentation prior which is a by-product of motion discrimination is incorporated directly in E-step as conditional expectation. Let the vector $\Psi = [\Pi^T, \Sigma^T, \Theta^T]^T$ represents all unknown parameters, where $\Pi = [\pi_1, \cdots, \pi_g]^T$, $\Sigma = [\sigma_1, \cdots, \sigma_g]^T$, $\Theta = [\theta_i, \cdots, \theta_g]^T$ are population proportions, variances and motion parameters respectively. Furthermore, let $\mathbf{z}_j = [z_{1j}, \cdots, z_{gj}]^T$ (as [4, p. 48]) represents the vector of ownership indicator, our EM algorithm is expressed as follows.

E Step. The expectation τ_{ij} of the binary ownership at j^{th} pixel location $\mathbf{p}_j$ for i^{th} population , at m^{th} iteration, is given by

$$\tau_{ij} = E(z_{ij}^m | I(\mathbf{p}_j), z_{ij}^{m-1})$$

$$
\begin{aligned}
&= \mathrm{pr}(z_{ij}^m = 1 | I(\mathbf{p}_j), z_{ij}^{m-1}) \\
&= \frac{p(I(\mathbf{p}_j)|z_{ij}^m = 1) p(z_{ij}^m = 1 | z_{ij}^{m-1})}{\sum_{k=1}^g p(I(\mathbf{p}_j)|z_{kj}^m = 1) p(z_{kj}^m = 1 | z_{ij}^{m-1})} \\
&= \frac{\rho_i^m f(I(\mathbf{p}_j), \theta_i, \sigma_i)}{\sum_{k=1}^g \rho_k^m f(I(\mathbf{p}_j), \theta_k, \sigma_k)} \qquad (7)
\end{aligned}
$$

where $\rho_k^m = p(z_{kj}^m = 1 | z_{ij}^{m-1})$ indicates the reliability of ownership by given a segmentation prior at $(m-1)^{th}$ iteration. $f(I(\mathbf{p}_j), \theta_i, \sigma_i)$ is a density function for the random variate of image intensity according to different motion parameters. It usually takes a form of normal distribution.

M Step. Given ownership expectation $\{\tau_{ij}\}$, the maximum likelihood estimates of parameter Ψ, $\hat{\Psi}$, satisfy the following equations:

$$
\hat{\pi}_i = \sum_{j=1}^n \tau_{ij}/n \quad i = \{1, \cdots, g\} \qquad (8)
$$

$$
\sum_{i=1}^g \sum_{j=1}^n \tau_{ij} \frac{\partial \log f(I(\mathbf{p}_j); \theta_i, \sigma_i)}{\partial \hat{\sigma}_i} = 0 \qquad (9)
$$

$$
\sum_{i=1}^g \sum_{j=1}^n \tau_{ij} \frac{\partial \log f(I(\mathbf{p}_j); \theta_i, \sigma_i)}{\partial \hat{\theta}_i} = 0 \qquad (10)
$$

Eqn (10) is solved by Gaussian-Newton algorithm as in [6].

5. Object Tracking

Mean shift [1, 2] is adopted for object tracking due to its efficiency and robustness to non-rigid motion. The tracking algorithm is appearance-based and mean shift procedure is utilized to match a target candidate which is most similar to the target model. The similarity measure is based on Bhattacharyya coefficient metric between the color density distributions of a target model $\hat{\mathbf{q}} = \{\hat{q}_u\}_{u=1\cdots m}$ (with $\sum_{u=1}^m \hat{q}_u = 1$) and a target candidate $\hat{\mathbf{p}}(\mathbf{y}) = \{\hat{p}_u(\mathbf{y})\}_{u=1\cdots m}$ (with $\sum_{u=1}^m \hat{p}_u = 1$). Bhattacharyya coefficient is given as

$$
\hat{\rho}(\mathbf{y}) \equiv \rho[\hat{\mathbf{p}}(\mathbf{y}), \hat{\mathbf{q}}] = \sum_{u=1}^m \sqrt{\hat{p}_u(\mathbf{y}), \hat{q}_u} \qquad (11)
$$

where m is the quantization level of a color histogram. By Eqn (11), the distance between two distributions is

$$
d(\mathbf{y}) = \sqrt{1 - \rho[\hat{\mathbf{p}}(\mathbf{y}), \hat{\mathbf{q}}]} \qquad (12)
$$

The target color distribution can be represented as follows. Denote $\{\mathbf{x}_i\}_{i=1,\cdots,n}$ as the pixel locations of a target candidate centered at $\mathbf{y}$. A convex and monotonic decreasing kernel profile k is used to assign smaller weights to the locations that are farther from $\mathbf{y}$. Let $b(\mathbf{x}_i)$ as a function which indexes the histogram bin of a given color, the normalized probability of a color u in a target candidate is

$$
\hat{p}_u(\mathbf{y}) = \frac{\sum_{i=1}^n k\left(\left\|\frac{\mathbf{y}-\mathbf{x}}{h}\right\|^2\right) \delta[b(\mathbf{x}_i) - u]}{\sum_{i=1}^n \left(\left\|\frac{\mathbf{y}-\mathbf{x}}{h}\right\|^2\right)} \qquad (13)
$$

where δ is the Kronecker delta function and h is the radius of the kernel profile. The distribution of target model, $\hat{\mathbf{q}}$, can be derived in a similar fashion.

Given an initial target location $\mathbf{y}_0$, the new location $\mathbf{y}_1$ of a target candidate is achieved by maximizing Eqn (11) based on the mean shift iteration given by

$$
\mathbf{y}_1 = \frac{\sum_{i=1}^n \mathbf{x}_i \omega_i g\left(\left\|\frac{\mathbf{y}_0-\mathbf{x}}{h}\right\|^2\right)}{\sum_{i=1}^n \omega_i g\left(\left\|\frac{\mathbf{y}_0-\mathbf{x}}{h}\right\|^2\right)} \qquad (14)
$$

where $g = -k'$ and

$$
\omega_i = \sum_{u=1}^m \delta[b(\mathbf{x}_i) - u] \sqrt{\frac{\hat{q}_u}{\hat{p}_u(\hat{\mathbf{y}}_0)}}
$$

The mean shift tracking algorithm, in principle, searches for local maximum in the neighborhood of initial location by exploiting the gradient of surface.

6. Experiments

To demonstrate the effectiveness of our approach, a soccer sequence with multiple players is used for testing. In this sequence, a camera is moved to track four players that run randomly on soccer field. The target models are represented by RGB histograms with $32 \times 32 \times 32$ bins.

Figure 2 shows the discriminant curve calculated by Eqn (5) at every two adjacent frames. The dashed horizontal line indicates the class separability M where $M > E(M) + \alpha \times Var(M)$ (see Section 3). Only those frames whose M values are above the horizontal line will be considered for seed selection. There are two seeds selected as the starting tracking positions. Initially, the first seed which consists of three distinctive clusters is picked up at 11^{th} frame. Another seed is adaptively selected at 145^{th} frame when the target models are occluded at 139^{th} frame.

Figure 3 shows the results of object detection by our EM algorithm at the two selected seed frames. The target windows are initialized based on the segmentation results by EM. Initially, there are three clusters found at the 11^{th} frame by motion clustering (as shown in Figure 4): one cluster corresponds to the background, one corresponds to the two players on the left who run in the same direction, and one corresponds to another two players on the right with same motion direction. By simultaneous estimation of motion models and segmented regions, EM successfully detect the blobs of four players, as shown in Fig 3(c).

Figure 5 shows two events when the candidate matching based on mean shift tracking degrades. The top row is an event that a player moves out of the camera view. The corresponding tracker is removed at the 65^{th} frame. The bottom row shows an event of occlusion. The three players at the 139^{th} frame, shown on the left, are tracked according to the target models estimated at the first seed (11^{th} frame). Because one of the player is occluded, the three players at

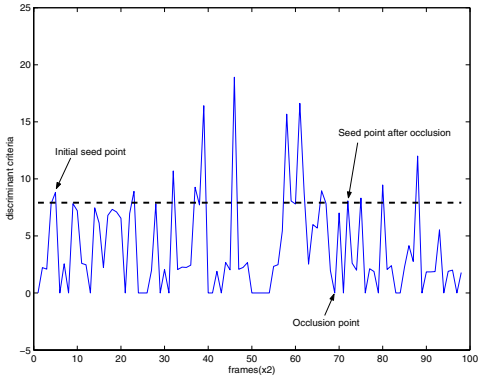
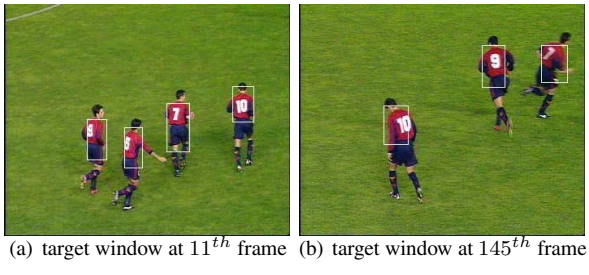


Figure 2. Discriminant Curve



(a) target window at 11^{th} frame (b) target window at 145^{th} frame

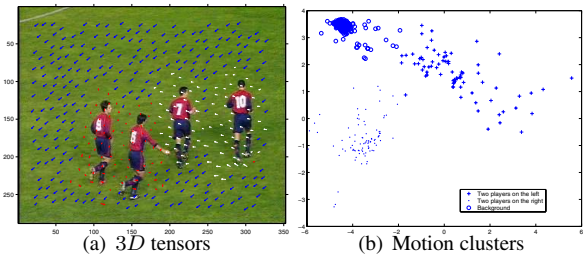
(c) detected objects at 11^{th} (d) detected objects at 145^{th}

Figure 3. Object initialization and detection.

the 140^{th} frame, shown on the right, are tracked based on the target models detected at the second seed frame (145^{th} frame), instead of the first seed. In this sequence, all players are tracked effectively and correctly in spite of occlusion, object disappearance and moving camera.

7. Conclusion

We have presented a unified approach for multiple objects tracking. The novelty lies on the novel seed selection, which finds the frames with good initial conditions for object segmentation. This has indeed led to robust object detectors based on EM algorithm. With the accurate target models, the objects can be tracked rapidly, and most importantly more effectively, by mean shift algorithm. Notice that object occlusion, which is regarded as a difficult task, can also be circumvented in our approach.



(a) $3D$ tensors (b) Motion clusters

Figure 4. Motion analysis on 11^{th} frame.

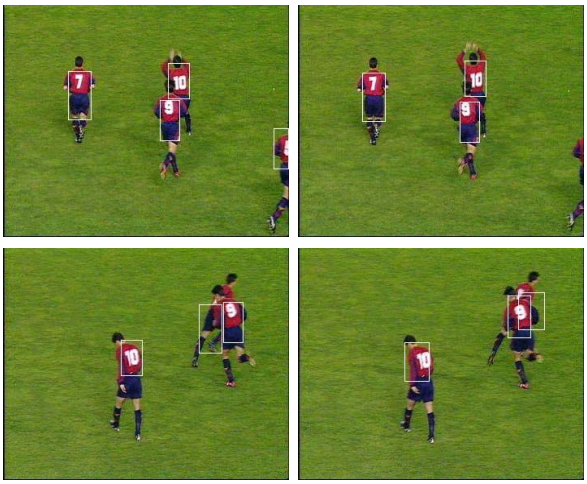


Figure 5. Object tracking at frames 64, 65, 139 and 140 (from left to right and top to bottom).

Acknowledgements

The work described in this paper was fully supported by a RGC grant CityU1072/02E (project No. 9040693) from the Research Grants Council of the Hong Kong SAR.

References

[1] D. Comaniciu, V. Ramesh, and P. Meer. Real-time tracking of non-rigid objects using mean shift. In *IEEE Conf. on CVPR*, volume 2, pages 142–149, June 2000.

[2] D. Comaniciu, V. Ramesh, and P. Meer. Kernel-based object tracking. *IEEE Trans. PAMI*, 25(5):564–577, May 2003.

[3] M. Isard and A. Blake. Condensation - conditional density propagation for visual tracking. *IJCV*, 29(1):5–28, 1998.

[4] G. McLachlan and D.Peel. *Finite Mixture Models*. Wiley, New York, 2000.

[5] N. Paragios and R. Deriche. Geodesic active regions for motion estimation and tracking. In *IEEE ICCV*, volume 1, pages 688–694, 1999.

[6] H. S. Sawhney and S. Ayer. Compact representation of videos through dominant and multiple motion estimation. *IEEE Trans. PAMI*, 18(8):814–830, Aug. 1996.