KERNEL-BASED SPATIAL-COLOR MODELING FOR FAST MOVING OBJECT TRACKING

R. Venkatesh Babu

Department of Electrical Engineering,
Indian Institute of Science,
Bangalore, India

Anamitra Makur

School of EEE,
Nanyang Technological University,
Singapore

ABSTRACT

Visual tracking has been a challenging problem in computer vision over the decades. The applications of Visual Tracking are far-reaching, ranging from surveillance and monitoring to smart rooms. Mean-shift (MS) tracker, which gained more attention recently, is known for tracking objects in a cluttered environment and its low computational complexity. The major problem encountered in histogram-based MS is its inability to track rapidly moving objects. In order to track fast moving objects, we propose a new robust mean-shift tracker that uses both spatial similarity measure and color histogram-based similarity measure. The inability of MS tracker to handle large displacements is circumvented by the spatial similarity-based tracking module, which lacks robustness to object's appearance change. The performance of the proposed tracker is better than the individual trackers for tracking fast-moving objects with better accuracy.

Index Terms— Visual Tracking, Mean-Shift, Object Tracking, Kernel Tracking

1. INTRODUCTION

The objective of object tracking is to faithfully locate the targets in successive video frames. The major challenges encountered in visual tracking are, cluttered background, noise, change in illumination, occlusion and scale/appearance change of the objects.

Visual tracking in a cluttered environment remains one of the challenging problems in computer vision for the past few decades. Various applications like surveillance and monitoring, video indexing and retrieval require the ability to faithfully track objects in a complex scene involving appearance and scale change. Though there exist many techniques for tracking objects, color-based tracking with kernel density estimation, introduced in [1, 2], has recently gained more attention among research community due to its low computational complexity and robustness to appearance change. The former [1] is due to the use of a deterministic gradient ascent (the "mean shift" iteration) starting at location in previous frame. The latter [2] relies on the use of a global appearance model, usually in terms of colors, as opposed to very precise appearance models such as pixel-wise intensity templates. The mean-shift algorithm was originally proposed by Fukunaga and Hostetler [3] for clustering data. It was introduced to image processing community by Cheng [4] a decade ago. This theory became popular among vision community after its successful application to image segmentation and tracking by Comaniciu and Meer [5, 6]. Later, many variants of the mean-shift algorithms were proposed for various applications [7, 8, 9, 10, 11, 12].

Though mean-shift tracker performs well on sequences with relatively small object displacement, its performance is not guaranteed when the objects move fast. Here, we attempt to improve the performance of mean-shift tracker when the object undergoes large displacements (when the object regions do not overlap between the consecutive frames). The problem of large displacements is tackled by cascading a spatial similarity-based mean-shift module with the traditional color histogram-based mean-shift tracker. The performance of the proposed tracker is better than the individual trackers.

The paper is organized as follows. Section 2 explains the proposed hybrid MS tracker to track fast moving objects and the results illustrating the better performance of proposed tracker is given in section 3. Concluding remarks are given in section 4.

This work was supported under $I^2R - NTU$ joint R&D (2) project.

2. PROPOSED HYBRID TRACKER

The object to be tracked is specified by location of its center and scale (for a fixed aspect ratio) in the image plane. The objective of the tracking algorithm is to find the object location in successive frames. MS tracker has problems with large displacements or when the object regions do not overlap between consecutive frames. In this paper, we would like to solve this problem within mean-shift domain itself. In the proposed tracker, the new spatial similarity measure-based tracker proposed by Yang et al., [9] is coupled with the conventional MS tracker to overcome the aforementioned difficulty. The spatial-similarity based MS tracker could be able to track objects with large displacements, but, the performance of this tracker degrades when the appearance of the object changes. On the other hand, the MS trackers which rely on persistent global object properties such as color, can be much more robust to detailed appearance changes due to shape and pose changes, but fails for large displacement. These complementary properties of these trackers can be used for developing a roust hybrid tracker that can handle both object's appearance change and large displacement. The main modules of proposed tracker is shown in Fig. 1. The first module separates the foreground object from its surrounding background. This preprocessing step helps the spatial similarity based MS tracker to rely only on the object pixels for tracking. The spatial-similarity based MS module provides the object displacement between two consecutive (current and next) frames. The color histogram based MS tracker is initialized by the displacement given by spatial similarity based MS module in the previous step. The final object displacement is given by the color-histogram based MS tracker which uses the fixed color model for the object. Each of these modules are explained in the following section.

2.1. Object-background separation

Tracking an object will be efficient if we can separate the object region from the background at each time instant. The object-background

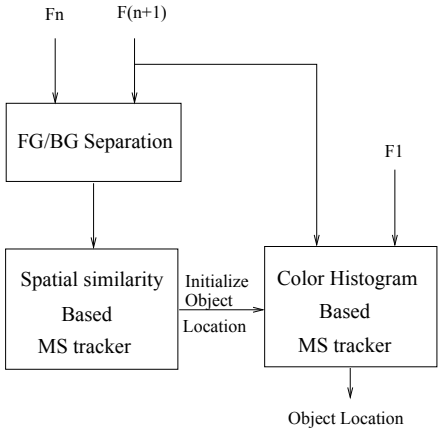

Fig. 1. Overview of the proposed tracking system

separation is useful in weighting the pixels for similarity-based MS tracker. To achieve this, the R-G-B based joint *pdf* of the object region and that of a neighborhood surrounding the object is obtained. This process is illustrated in Fig. 2. The region within the red rectangle is used to obtain the object *pdf* and the region between the green and red rectangles is used for obtaining the background *pdf*. The resulting log-likelihood ratio of foreground/background region is used to determine object pixels. The log-likelihood of a pixel considered within the outer bounding rectangle is (green rectangle in Fig. 2) obtained as

$$L_i = \log \frac{\max\{h_o(i), \epsilon\}}{\max\{h_b(i), \epsilon\}} \tag{1}$$

where $h_o(i)$ and $h_b(i)$ are the probabilities of ith pixel belonging to the object and background respectively; and ϵ is a small non-zero value to avoid numerical instability. The non-linear log-likelihood maps the multi-modal object/background distribution as positive values for colors associated with foreground and negative values for background. Only reliable object pixels are used in spatial-similarity-based MS tracker. The binary weighting factor T if ith pixel is obtained as:

$$T_i = \begin{cases} 1 & \text{if} \quad L_i > th_o \\ 0 & \text{otherwise} \end{cases} \tag{2}$$

where, th_o is the threshold to decide on the most reliable object pixels. Once the object is localized, by user interaction or detection in the first frame, the likelihood map of the object/background is obtained using (2). Typical value of th_o is set at 0.8.

2.2. Spatial Similarity-based Mean-shift tracking

Let $I_x = \{\mathbf{x}_i, \mathbf{u}_i\}_{i=1}^{N}$ and $I_y = \{\mathbf{y}_j, \mathbf{v}_j\}_{j=1}^{M}$ represent the target model and candidate model image. Here, $\mathbf{x}_i$ and $\mathbf{y}_j$ indicate the pixel locations with respect to model centers $\mathbf{x}_*$ and $\mathbf{y}$; $\mathbf{u}_i$ and $\mathbf{v}_j$ represent the feature vector of target and candidate model at locations $\mathbf{x}_i$ and $\mathbf{y}_j$ respectively (e.g., the RGB color values or gray scale intensity values). The target model I_x contains only the pixels for which the value of $T_i = 1$. The similarity between target (I_x) and candidate (I_y) in the joint feature-spatial space is

$$\rho_s(I_x, I_y) = \frac{1}{M.N} \sum_{i=1}^{N} \sum_{j=1}^{M} w\left(\left|\frac{\mathbf{x}_i - \mathbf{y}_i}{h_s}\right|^2\right) .k\left(\left|\frac{u_i - \mathbf{v}_i}{h_r}\right|^2\right) \tag{3}$$

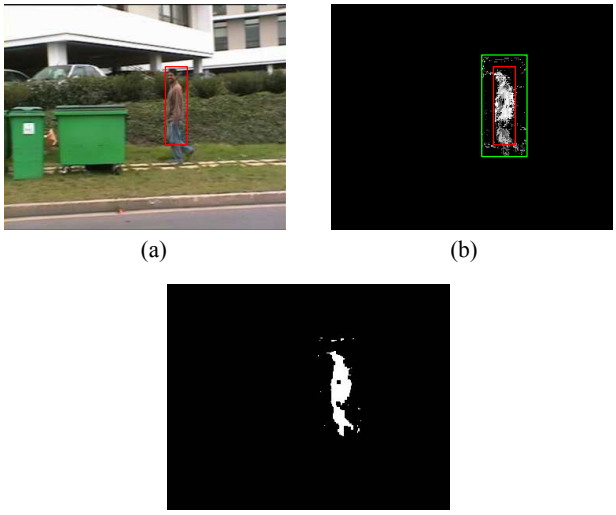

Fig. 2. (a) Initial frame with object boundary (b) likelihood map L (c) Mask obtained after morphological operations (T).

where, $k(x)$ and $w(x)$ are convex and monotonically decreasing kernel profiles. h_s and h_r are spatial and feature-space bandwidths.

Under the assumption that the motion between the frames is pure translation, the equation (3) becomes,

$$\rho_s(I_x, I_y) = \frac{1}{MN} \sum_{i=1}^{N} \sum_{j=1}^{M} w\left(\left|\frac{(\mathbf{x}_i - \mathbf{x}_*) - (\mathbf{y}_i - \mathbf{y})}{h_s}\right|^2\right) . \quad k\left(\left|\frac{\mathbf{u}_i - \mathbf{v}_i}{h_r}\right|^2\right) \tag{4}$$

Tracking is carried out by maximizing the above given similarity measure, or equivalently, by minimizing the following equation with respect to y.

$$L(I_x, I_y) = -log \;\; \rho_s(I_x, I_y) \tag{5}$$

where, $L(I_x, I_y)$ is a metric.

Let $\hat{\mathbf{y}}_0$ be the current position of the target model. The mean-shift algorithm recursively moves the current position $\hat{\mathbf{y}}_0$ to the new position $\hat{\mathbf{y}}_1$, until reaching the density mode according to

$$\hat{\mathbf{y}}_1 = \frac{\sum_i \sum_j \mathbf{y}_j k_{ij} g_{ij}}{\sum_i \sum_j k_{ij} g_{ij}} - \frac{\sum_i \sum_j \mathbf{x}_i k_{ij} g_{ij}}{\sum_i \sum_j k_{ij} g_{ij}} + \mathbf{x}_* \tag{6}$$

where, $g(x) = -w'(x)$ is the shadow kernel of kernel $w(x)$, $g_{ij} = g\left(|\frac{(\mathbf{x}_i - \mathbf{x}_*) - (\mathbf{y}_j - \hat{\mathbf{y}}_0)}{h_s}|^2\right)$

In our work, the target model (I_x) is not fixed at the first frame, instead it is freshly obtained from recently localized object in current frame. The candidate model (I_y) is initialized from the future frame with the model center ($\hat{\mathbf{y}}_0$) same as target model center ($\mathbf{x}_*$). The iteration is terminated when the mean shift between ($\hat{\mathbf{y}}_n - \hat{\mathbf{y}}_{n-1}$) consecutive iteration is less than a threshold (τ_s). The kernels k and w used in this module are Epanechnikov Kernel, hence the shadow kernel profile of w ($g = -w'$) is uniform kernel. The new location ($\hat{\mathbf{y}}_n$) is used for initializing the color-histogram based MS tracker.

2.3. Color Histogram-based MS Tracker

The target color model $\mathbf{q} = (q_i)_{i=1\cdots m}$, with $\sum_{i=1}^{m} q_i = 1$, is composed of m bins in some appropriate color space (e.g., RGB or Hue-Saturation). It is gathered at the initialization of the overall tracking. The candidate histogram $\mathbf{p}(\mathbf{x})$, at location $\mathbf{x}$ in the current frame is given by:

$$p_i(\mathbf{x}) = \frac{\sum_{\mathbf{d}\in D} k(|\mathbf{d}|^2)\delta[b(\mathbf{x}+\mathbf{d}) - i]}{\sum_{\mathbf{d}\in D} k(|\mathbf{d}|^2)} \quad (7)$$

where $k(x)$ is a convex and monotonic decreasing kernel profile, almost everywhere differentiable and with support D, which assigns smaller weights to pixels far away from the center, δ is the Kronecker delta function, and function $b(\mathbf{x}) \in \{1...m\}$ is the color bin number at pixel $\mathbf{x}$ in the current frame. One seeks the location whose associated candidate histogram is as similar as possible to the target one. When similarity is measured by Bhattacharyya coefficient, $\rho(\mathbf{p}, \mathbf{q}) = \sum_i \sqrt{p_i q_i}$, convergence towards the nearest local minima is obtained by the iterative mean-shift procedure [1]. In our case, this gradient ascent at time t is initialized at $\mathbf{y}_0 = \hat{\mathbf{y}}_n$ (where, $\hat{\mathbf{y}}_n$ is the location obtained from the spatial similarity-based MS tracker) and proceeds as follows:

1. Given current location $\mathbf{y}_0$, compute candidate histogram $\mathbf{p}(\mathbf{y}_0)$ and Bhattacharyya coefficient $\rho[\mathbf{p}(\mathbf{y}_0), \mathbf{q}]$.
2. Compute candidate position

$$\mathbf{y}_1 = \frac{\sum_{\mathbf{d}\in D} w(\mathbf{y}_0+\mathbf{d})k'(|\mathbf{d}|^2)(\mathbf{y}_0+\mathbf{d})}{\sum_{\mathbf{d}\in D} w(\mathbf{y}_0+\mathbf{d})k'(|\mathbf{d}|^2)}$$

   with weights at location $\mathbf{x}$

$$w(\mathbf{x}) = \sum_{i=1}^{m} \sqrt{\frac{q_i}{p_i(\mathbf{y}_0)}}\delta[b(\mathbf{x}) - i].$$

3. While $\rho[\mathbf{p}(\mathbf{y}_1), \mathbf{q}] < \rho[\mathbf{p}(\mathbf{y}_0), \mathbf{q}]$ do $\mathbf{y}_1 \leftarrow \frac{1}{2}(\mathbf{y}_1 + \mathbf{y}_0)$
4. If $\|\mathbf{y}_1 - \mathbf{y}_0\| < \varepsilon$ stop, otherwise set $\mathbf{y}_0 \leftarrow \mathbf{y}_1$ and repeat Step 2.

The final estimate provides the location of the object in current frame ($\mathbf{y}_1$).

3. PERFORMANCE OF THE COMBINED TRACKER

The proposed algorithm has been tested on several videos. The proposed tracking system, which uses both spatial similarity and color histogram of the object, works better than either of them used individually. Tracking results for PETS surveillance video is presented in Figs. 3 and 6. In order to simulate fast motion, all the videos are temporally subsampled by 4 before tracking. Figure 3 shows every 10th frame of the video (subsampled) of a walking person. The proposed tracker could track the person more accurately compared to spatial similarity MS tracker. The color based MS tracker fails to track the person in the initial phase itself due to the narrow bandwidth along the direction (horizontal) of object motion. Similar performance is observed for tracking a 'cycling person' in a PETS sequence (Fig. 6). In our experiments, the bandwidth values h_s and h_r are set at 5 and 9 respectively for spatial similarity MS module. Figs 4 and 5 show the tracking result for 'rugby' sequence. This sequence is a challenging dynamic one where the player move fast and undergo appearance change. In these sequences, the color-based

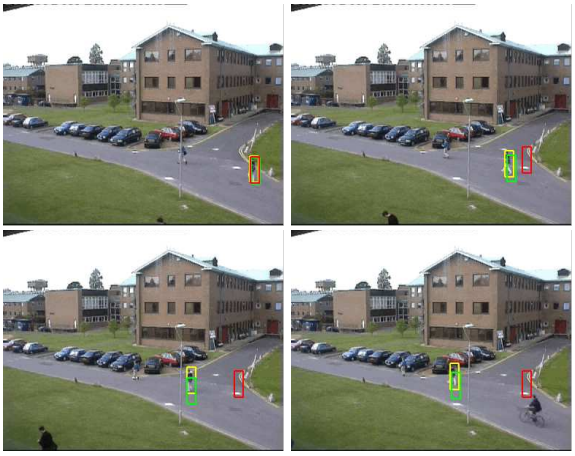

Fig. 3. Tracking result of proposed system (yellow) against Spatial Similarity based MS (green) and Color based MS (red) tracker for PETS surveillance sequence.

MS tracker fails quickly compared to spatial similarity-based MS tracker. The spatial similarity-based tracker could not track the objects accurately, while the proposed tracker could be able to track the objets more accurately. The spatial bandwidth (h_s) parameter fixes the spatial search range for localizing the object in the next frame. In these video sequences, the scale change of the object is not considered for tracking.

4. CONCLUSION

In this paper, we have proposed an efficient visual tracker for fast moving objects by coupling spatial similarity based MS tracker along with color based mean-shift tracker, which have complementary properties. The spatial similarity based tracker uses the local properties of the object for tracking while the color based tracker utilizes the global color information of the object. The proposed tracker uses both these properties for tracking rapidly moving objects. The performance of the proposed tracker is observed to be better than the individual ones. Since both trackers have real-time computational complexity, the proposed compound tracker is suitable for real time tracking of objects.

5. REFERENCES

[1] D. Comaniciu, V. Ramesh, and P. Meer, "Real-time tracking of non-rigid objects using mean shift," in *Proc. Conf. Comp. Vision and Pattern Recog.*, Hilton Head, SC, 2000.

[2] G. Bradski, "Computer vision face tracking as a component of a perceptual user interface," in *Workshop on Application of Computer Vision*, Princeton, NJ, Oct. 1998, pp. 214–219.

[3] K. Fukunaga and L. D. Hostetler, "The estimation of the gradient of a density function, with applications in pattern recognition," *IEEE Transactions on Information Theory*, vol. 21, no. 1, pp. 32–40, 1975.

[4] Y. Cheng, "Mean shift, mode seeking, and clustering," *IEEE Trans. Pattern Analysis Machine Intelligence*, vol. 17, no. 8, pp. 790–799, 1995.

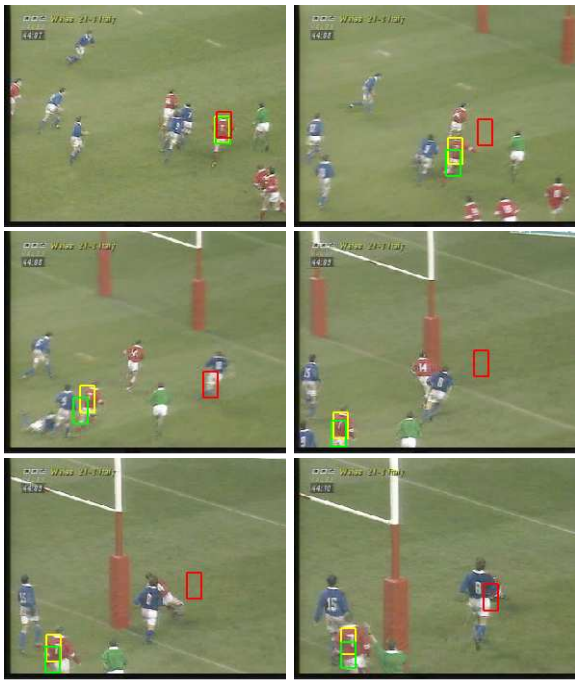

Fig. 4. Tracking result of proposed system (yellow) against Spatial Similarity based MS (green) and Color based MS (red) tracker for rugby sequence.

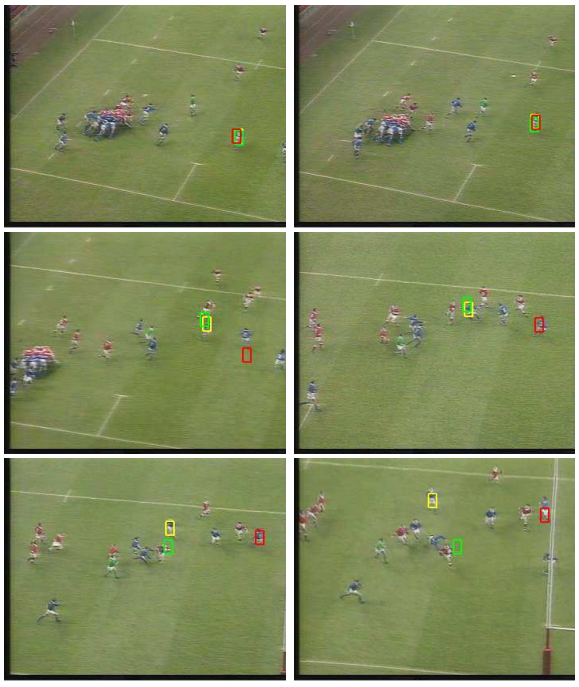

Fig. 5. Tracking result of proposed system (yellow) against Spatial Similarity based MS (green) and Color based MS (red) tracker for rugby sequence.

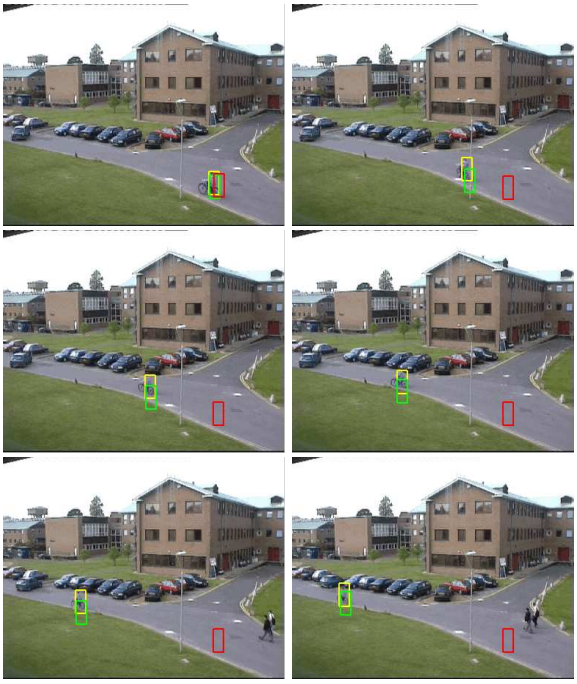

Fig. 6. Tracking result of proposed system (yellow) against Spatial Similarity based MS (green) and Color based MS (red) tracker for PETS surveillance sequence.

[5] D. Comaniciu and P. Meer, "Mean shift: A robust approach toward feature space analysis," *IEEE Trans. Pattern Analysis and Machine Intelligence*, vol. 24, no. 5, pp. 603–619, May 2002.

[6] D. Comaniciu, V. Ramesh, and P. Meer, "Kernel-based object tracking," *IEEE Trans. Pattern Anal. Machine Intell.*, vol. 25, no. 5, pp. 564–577, 2003.

[7] F. Porikli and O. Tuzel, "Multi-kernel object tracking," in *IEEE International Conference on Multimedia and Expo*, 2005, pp. 1234–1237.

[8] Chunhua Shen, M. J. Brooks, and A. van den Hengel, "Fast global kernel density mode seeking with application to localisation and tracking," in *IEEE International Conference on Computer Vision*, 2005, vol. 2, pp. 1516–1523.

[9] Changjiang Yang, R. Duraiswami, and L. Davis, "Efficient mean-shift tracking via a new similarity measure," in *IEEE Conf. on Computer Vision and Pattern Recognition*, 2005, vol. 1, pp. 176–183.

[10] Caifeng Shan, Yucheng Wei, Tieniu Tan, and F. Ojardias, "Real time hand tracking by combining particle filtering and mean shift," in *IEEE International Conference on Automatic Face and Gesture Recognition*, 2004, pp. 669–674.

[11] G. Hager, M. Dewan, and C. Stewart, "Multiple kernel tracking with SSD," in *IEEE Conf. Computer Vision and Pattern Recognition*, 2004, vol. 1, pp. 790–797.

[12] K. Deguchi, O. Kawanaka, and T. Okatani, "Object tracking by the mean-shift of regional color distribution combined with the particle-filter algorithms," in *International Conference on Pattern Recognition*, Aug. 2004, vol. 3, pp. 506–509.