# Fast Visual Tracking Using Motion Saliency in Video

*Shan Li, M.C. Lee*

The Department of Computer Science and Engineering, the Chinese University of HK
Email: {sli, mclee}@cse.cuhk.edu.hk

**ABSTRACT**

We present a novel tracking method for surveillance videos where the object and camera motion could be irregular. The tracker is mainly based on the mechanism of motion *saliency* detection, where regions of interest are detected according to the motion saliency of the moving objects. Without estimating global and local motion explicitly, we generate motion saliency by using the rank deficiency of gray scale gradient tensors within the local region neighborhoods of the image. In addition, multiple spatial features of targets are integrated in the system to assist the region-based tracking task. Experiment results on many videos suggest that the proposed method tracks moving targets efficiently even in videos with unstable cameras.

***Index Terms***— Video tracking, motion saliency

## 1. INTRODUCTION

Object tracking is important because it enables several applications such as video surveillance, robotics and image/video analysis. Tracking objects in videos with moving cameras is an important research topic. For example, cameras placed in moving vehicles, lifts and tree braches, or cameras in pan/tilt/zoom operations will all cause global motion in videos. An object tracking system should be able to track the target despite the interference of the background motion. Accurate object tracking under the constraint of low computational complexity, however, presents a challenge for videos with objects/cameras moving irregularly.

Many approaches have been proposed to address surveillance tracking problems. Some [3] [2][5] made use of foreground/ background appearance or motion modeling and track objects accordingly. Their tracking models assume that either the camera is static or the object/camera motion is "regular" enough to be predicted. Other approaches [1][4] [8]use motion estimation methods such as parametric motion estimation or block-based matching to estimate camera and object motions respectively. Objects are detected as regions with motion discontinuities. A common problem of the approaches is that the calculation of motions is highly computation intensive, which makes the real time analysis on videos impossible. Dense unconstrained and non-rigid motion estimation is also highly noisy and unreliable, especially when the dynamic events contain unstructured objects such as running water and flickering fire, etc.

In this paper, we solve the above mentioned problems by proposing a novel method in motion saliency estimation. A tracking system for surveillance videos is presented with two main components. 1) A motion saliency detection model: We observe that in surveillance videos, the tracking targets are usually those with salient motions to the background. A 3D gray scale gradient tensor is used to explore the motion consistency of the local image structure. The estimation of motion saliency is replaced as the rank deficiency measure of the gradient tensor. 2) A region-based tracking model: The features of each target are updated at each frame to generate adaptive templates. The detected regions of interest in the motion saliency maps are labeled as targets if the matching error is small.

The rest of the paper is organized as follows. Section 2 presents the model of motion saliency estimation. The region-based tracking model is introduced in Section 3. Experiment results and discussions are given in Section 4. Finally, Section 5 concludes the paper.

## 2. MOTION SEGMENTATION USING MOTION SALIENCY ESTIMATION

The motion saliency model detects the moving objects whose motion is salient to its background. For example, a moving car in a surveillance video has big motion relative to the background; On the other hand, a walking person in a tracking camera will as well be the target due to its small motion relative to the big motions of the background. Fig. 1 showed the key frames of two shots where foreground objects with salient motions are the targets.

Although we don't compute motion vectors specifically, we explore the 3D structural tensor used in the work [9]. Assuming that optical flow $f$= $(u,v,w)$ is constant within a small neighborhood, the optical flow of the neighborhood can be estimated by solving the following 3D structural tensor:

$$M \cdot \begin{bmatrix} u \\ v \\ w \end{bmatrix} = O_{3\times 1}, \text{ with} \qquad (1)$$

$$M = \begin{bmatrix} \sum_{i\in\Omega} w(c-i)g_x^2 & \sum_{i\in\Omega} w(c-i)g_x g_y & \sum_{i\in\Omega} w(c-i)g_x g_t \\ \sum_{i\in\Omega} w(c-i)g_x \mathrm{g_y} & \sum_{i\in\Omega} w(c-i)g_y^2 & \sum_{i\in\Omega} w(c-i)g_y g_t \\ \sum_{i\in\Omega} w(c-i)g_x \mathrm{g_t} & \sum_{i\in\Omega} w(c-i)g_y \mathrm{g_t} & \sum_{i\in\Omega} w(c-i)g_t^2 \end{bmatrix}$$

(a) a video with a static camera

(b) a video with a moving camera

Fig. 1. Objects with salient motions in surveillance videos. (a) In a surveillance video shot, a moving car under a static camera is the target of tracking task. (b) In a tracking video, due to the panning behavior of the camera, the walking person has small motions relative to the fast motions of the background. The walking person, with its salient motions, is the tracking target.

where the weighting function $w(c)$ selects the size of the neighborhood $\Omega$ centered at the pixel $c$(the selection of $\Omega$ will be introduced at the end of the section). In implementations, the weighting is realized by a Gaussian smoothing kernel. $\nabla g = (g_x, g_y, g_t)$ is the space-time gradients of the intensity at respective pixel within the small neighborhood.

In the above 3D structural tensor, gradients are summed over all pixels within the neighborhood $\Omega$. Eq. (1) is in nature a set of linear equations. Therefore, if the optical flows ($u, v, w$) are constant within the neighborhood, there will be a non-zero solution of Eq. (1). Consequently, the 3×3 coefficient matrix $M$ (i.e., the 3D structural tensor) should be rank deficient : $rank(M) \le 2$.

The rank of the coefficient matrix $M$ can be used to analyze the brightness distribution and motion types within the neighborhood $\Omega$. In the case rank($M$)=3, there are multiple motions within the neighborhood; For rank($M$)=2, a distributed spatial brightness structure moves at a constant motion; In the degenerate image structure (i.e., rank($M$)=0,1), the motion of the image structure can not be concluded by using Eq. (1). However, exploring the fact that the neighborhood with constant brightness or with homogeneous contour most likely belongs to the same object or background, the neighborhood can be assumed to move at a constant motion.

In the presence of noise, $M$ may be always full rank in real videos. Besides, matrix deficiency will be better captured in a continuous measure as mixtures between the types of motion usually occur in a real video sequence. A normalized and continuous measure is therefore needed to quantify the matrix deficiency. Let $\lambda_1 \ge \lambda_2 \ge \lambda_3$ be the eigenvalues of the 3×3 matrix $M$. We define the continuous rank-deficient measure $d_M$ to be:

$$d_M = \begin{cases} 0, & trace(M) < \gamma \\ \dfrac{\lambda^2{}_3}{0.5 \cdot \lambda^2{}_1 + 0.5 \cdot \lambda^2{}_2 + \varepsilon}, & \text{otherwise} \end{cases} \quad (2)$$

where $\varepsilon$ avoids division by 0. The threshold $\gamma$ is used to handle the case rank($M$)=0. It can be chosen adaptively according to the noise level of the image sequence (e.g., in our implementation, $\gamma$ is set as the 0.1 percentile point in the accumulative distribution of the trace values of neighborhood centered at each pixel). The case of $d$=0 is an ideal case of rank deficient (i.e., constant motion), and when $d_M$ =1, the matrix $M$ is clearly full rank (i.e., multiple motions). The continuous definition of $d_M$ allows noisy data handling and provides varying degrees of rank-deficiency for varying degrees of motion-constancies within the neighborhood.

Similar rank-based measures can be used to determine whether two image structures share similar motions. Given the spatiotemporal structural tensors $M_1$ and $M_2$ of a small local neighborhood $\Omega_1$ and a background $\Omega_2$ (the selection of $\Omega$ will be introduced at the end of the section) respectively, we say $\Omega_1$ and $\Omega_2$ share the same motion ($u, v, w$) if the combined coefficient matrix $M_{12}$ is rank deficient:

$$M_{12}=[M_1+M_2] \quad (3)$$

$$= \begin{bmatrix} \sum_{i\in\Omega_1+\Omega_2} w(c-i)g_x^2 & \sum_{i\in\Omega_1+\Omega_2} w(c-i)g_x g_y & \sum_{i\in\Omega_1+\Omega_2} w(c-i)g_x g_t \\ \sum_{i\in\Omega_1+\Omega_2} w(c-i)g_x \mathrm{g}_\mathrm{y} & \sum_{i\in\Omega_1+\Omega_2} w(c-i)g_y^2 & \sum_{i\in\Omega_1+\Omega_2} w(c-i)g_y g_t \\ \sum_{i\in\Omega_1+\Omega_2} w(c-i)g_x g_t & \sum_{i\in\Omega_1+\Omega_2} w(c-i)g_y \mathrm{g}_\mathrm{t} & \sum_{i\in\Omega_1+\Omega_2} w(c-i)g_t^2 \end{bmatrix}$$

The measure of Eq. (2) can be applied to $M_{12}$ to form a rank-deficient value $d_{M12}$. However, the integrated spatiotemporal structural sensor defined in Eq. (3) could be highly sensitive to noise. Multiple motions in either $M_1$ or $M_2$ will both yield high values in $d_{M12}$. $M_{12}$ could be full rank in most real videos, making $d_{M12}$ an impractical measure for motion saliency estimation. This problem is even more serious when the background $\Omega_2$ is composed of multiple neighborhoods in different locations where motions could be more diverse.

The problem can be solved by computing $M_{12}$ based on the averaged motions in $\Omega_1$ and $\Omega_2$ respectively. Recall that in Eq.(1), a single uniform optical flow ($u, v, w$) can be calculated from the linear equation systems if rank($M$)=2. For rank($M$)=3, however, an averaged optical flow $\tilde{f} = (\tilde{u}, \tilde{v}, \tilde{w})$ should satisfy the following linear equation system:

$$\tilde{M}_{2\times 3} \cdot \begin{bmatrix} \tilde{u} \\ \tilde{v} \\ \tilde{w} \end{bmatrix} = O_{3\times 1} \ , \ \text{with} \quad (4)$$

$$\tilde{M} = \begin{bmatrix} \sum_{i\in\Omega} w(c-i)g_x^2 & \sum_{i\in\Omega} w(c-i)g_x g_y & \sum_{i\in\Omega} w(c-i)g_x g_t \\ \sum_{i\in\Omega} w(c-i)g_x \mathrm{g}_\mathrm{y} & \sum_{i\in\Omega} w(c-i)g_y^2 & \sum_{i\in\Omega} w(c-i)g_y g_t \end{bmatrix}$$

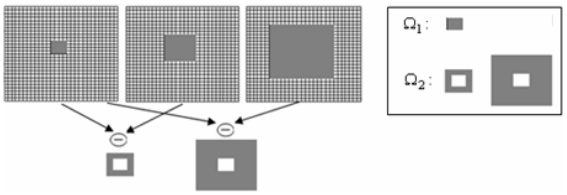


Fig.2. Examples of multi-scale extraction of $\Omega_1$ and $\Omega_2$.The gray areas in the top-row images indicate sub-sampled areas using different scales. The combination of different center fine scales and surround coarse scales forms multi-scale representations of $\Omega_1$and $\Omega_2$. The gray areas in the bottom-row images show the background areas $\Omega_2$ with different scales w.r.t. $\Omega_1$ in the first top-row image.

Replacing motions in $\Omega_1$ and $\Omega_2$ with the averaged motions respectively, we have:

$$\widetilde{M}_1 \cdot \begin{bmatrix} \widetilde{u} \\ \widetilde{v} \\ \widetilde{w} \end{bmatrix} = O_{3\times 1};\ \widetilde{M}_2 \cdot \begin{bmatrix} \widetilde{u} \\ \widetilde{v} \\ \widetilde{w} \end{bmatrix} = O_{3\times 1} \text{ and,}$$

$$\widetilde{M}_{12} = \begin{bmatrix} \widetilde{M}_1 \\ \widetilde{M}_2 \end{bmatrix}_{4\times 3} \cdot \begin{bmatrix} \widetilde{u} \\ \widetilde{v} \\ \widetilde{w} \end{bmatrix} = O_{3\times 1}, \quad (5)$$

Multiplying $\widetilde{M}_{12}^T$ (the transpose of matrix $\widetilde{M}_{12}$) on the left side of $\widetilde{M}_{12}$, we get:

$$M^*_{3\times 3} \cdot \begin{bmatrix} \widetilde{u} \\ \widetilde{v} \\ \widetilde{w} \end{bmatrix} = O_{3\times 1}, \text{ with,} \quad (6)$$

$$M^* = \widetilde{M}_{12}^T \widetilde{M}_{12} = [\widetilde{M}_1^T \widetilde{M}_1 + \widetilde{M}_2^T \widetilde{M}_2]_{3\times 3}$$

The problem of detecting motion saliency between $\Omega_1$ and $\Omega_2$ can therefore be solved by measuring the rank deficiency of the 3×3 matrix *M**. The measure of Eq. (2) can be applied to *M** to form a rank-deficient value $d_{M^*}$. $d_{M^*}$ gives a low value if the averaged motions of $\Omega_1$ and $\Omega_2$ are consistent, and provides a high value if the averaged motion in $\Omega_1$ is much salient to the background $\Omega_2$.

Motion saliency maps $P_m$ can be formed by regarding $d_{M^*}$ as the pixel value in the map. To obtain multi-scaling motion attention maps, $\Omega_1$ and $\Omega_2$ are extracted in a form of center-surround pyramids (Fig.2). Local areas can be sampled from a video frame at different scales $2^{\sigma}$, σ=[0..8]. Center-surround pyramids are formed by regarding areas in a "center" fine scale $2^c$ as $\Omega_1$ and areas in a "surround" coarser scale$2^s$ as $\Omega_1$+$\Omega_2$. A set of six motion attention maps $P_m^{c,s}$ can be formed with $c \in \{2,3,4\}$ and $s = c + \delta$, $\delta \in \{3,4\}$, if the image is big enough for the sampling rate.

Before applying the motion saliency maps in object tracking, a few processing are required to refine the saliency map. To eliminate the modality-dependent amplitude differences, the generated motion saliency maps are first normalized to a fixed range [0, 1]. The maps are then summed across different scales to form an overall motion saliency map. Pixels with saliency values bigger than a threshold (say, 0.1) are considered as region of interest, which are the potential tracking regions in the map.

One example of the motion saliency detection is shown in Fig. 3, where the camera moves rapidly to track a walking person from left to right and then from right to left. In the example, both the camera motion and the object motion are complicated in the sample video. The camera motions change rapidly and constantly during the entire video, which bring much noise to traditional motion estimation methods. Without explicitly calculating the camera and object motions, however, our method effectively detects the salient motions. Pixels corresponding to the walking person have high saliency values in the generated maps.

A few post processing are applied on the overall motion saliency map. First, a binary mask is formed by assigning regions of interest with value 1 and the remaining pixels with value 0. A "close" morphological operation (dilation followed by erosion) is employed to fill the holes in the mask. The neighboring regions with high saliency values are then merged if their color distributions in the original image are similar.

## 3. REGION-BASED OBJECT TRACKING

For tracking purpose, correspondences are required between frames to form object trajectories. Typical tracking methods are divided into four major categories: region-based, active-contour-based, feature-based and model-based tracking [6]. A robust and general tracking system needs less specific models for tracking. Thus a region-based tracking strategy[7] is adopted in our tracking system.

A temporally consistent list of tracked regions is maintained during tracking. Temporal matching is performed based on the support map and bounding box. In particular, a set of spatial features, including object centroid(*c*), bounding box size(*s*), and color distribution (*h*) of each object, is used to construct object template for matching purpose.

The object template in different feature channels can be updated adaptively at each frame. Specifically, the color distribution of the target is modeled for tracking using color histograms. A histogram $H_i(x)$ with $x=(r,g,b)$ provides the discrete probability distribution of the color value $x$ in the target $i$ ($N_i$ is the total number of pixels in target $i$):

$$h(x \mid i) = \frac{H_i(x)}{N_i} \quad (7)$$

Multiple features $f=\{c, s, h\}$ of each target are adaptively updated as:

$$f_{t+1} = \beta \cdot f_t + (1-\beta) \cdot f^*_{t+1} \quad (8)$$

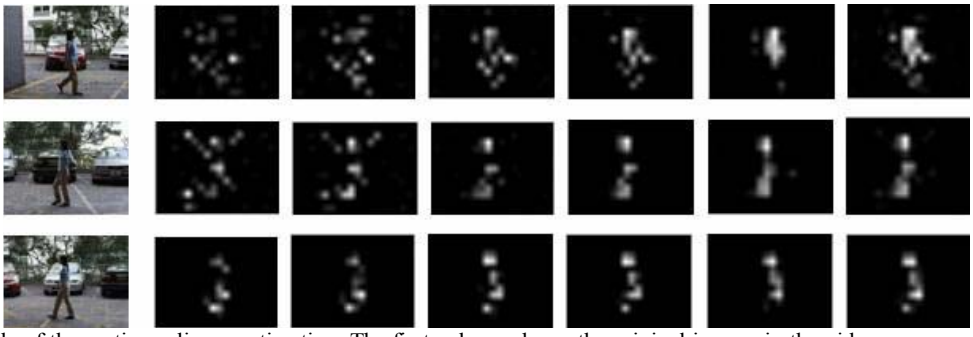

Fig. 3. An example of the motion saliency estimation. The first column shows the original images in the video sequence. Columns from 2 to 6 are the generated motion maps with center-surround scales of 6-3, 7-3, 7-4, 8-4, and 8-5 respectively. The last column shows the combined motion saliency map. In this example, the camera tracks the walking person. While the background has intensive motions, the person's motion is greatly canceled by the camera motion. As shown in the maps, the person has a motion much salient to his intermediate background. The brighter color indicates higher attention values in the attention map.

where $f_{t+1}^{*}$ is the feature template generated from the $t$+1-th frame. $f_{t+1}$ is the updated template. $\beta = 0.8$ is the "forgetting" rate in the updating process.

A cost function $C$ is computed to integrate matching errors from all features and a best match is found for each region of interest detected in the motion saliency map:

$$C = \sqrt{(\frac{c_o - c_b}{\sigma_c})^2 + (\frac{s_o - s_b}{\sigma_s})^2 + (\frac{h_o - h_b}{h_d})^2} \qquad (9)$$

The subscript "$o$" and "$b$" of each feature indicates the target template and the region of interest in saliency map respectively. The deviation $\sigma$ of each feature is calculated from past 50 observations.

## 4. EXPERIMENTS

The proposed tracking method has been applied in different sets of outdoor surveillance videos, where the objects are mostly moving people. The testing videos are used to evaluate the effectiveness of the proposed method in handling complicated motions in videos. Generally, the target and camera motions in these videos are irregular. The targets may change their moving speed and moving directions during the video. Cameras in some videos also have intensive motions. For instance, the camera in the 3rd video of Fig. 4 is jerking since it is hand hold. Some of the tracking results on different scenes are shown in Figure 4. In these images, bounding boxes indicate the target locations.

## 5. CONCLUSION

A novel method for foreground segmentation has been presented, based on the time efficient estimation of motion saliency between objects and backgrounds. A tracking algorithm has also been developed to link the observation templates with the detected objects in the motion saliency map. The experimental results demonstrate that the tracker reliably tracks objects in both stable and unstable cameras.

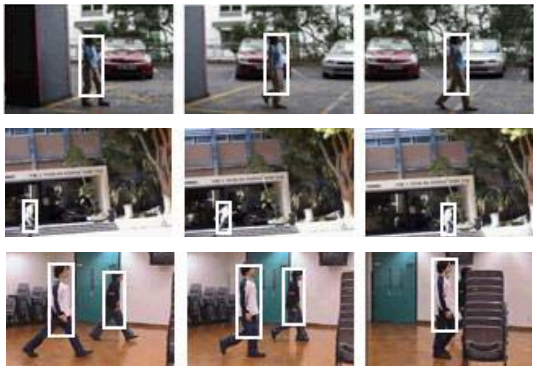

Fig.4 Tracking of moving objects. The bounding box indicates the location of targets.

## 6. REFERENCES


[1] MJ Black and DJ Fleet, "Probabilistic detection and tracking of motion discontinuities," *IEEE Conf. ICCV*, pp.551-558, 1999

[2] M. Isard and A. Blake, "Icondensation: Unifying low-level and high-level tracking in a stochastic framework," *IEEE conf. ECCV*, pp. 893-908, 1998

[3] Ismail Haritaoglu, David Harwood, and Larry S. Davis, "W4: Real-Time Surveillance of People and Their Activities," *IEEE. Trans. PAMI*, vol. 22, No. 8, pp. 809-830, 2000

[4] Karthik Hariharakrishnan and Dan Schonfeld, "Fast object tracking using adaptive block matching," *IEEE trans. Multimedia*, vol.7, no.5, pp. 853-859, 2005

[5] Weiming Hu, Xuejuan Xiao, Zhouyu Fu, Xie, D., Tieniu Tan and Maybank, S., "A system for learning statistical motion patterns," *IEEE Trans. PAMI*, vol.28(9), pp.1450-1464, 2006

[6] Weiming Hu, Tieniu Tan, Liang Wang and Steve Maybank, "A survey on visual surveillance of object motion and behaviors," *IEEE Trans.SMC Part C*, vol.34(3), pp.334-352, 2004

[7] SJ McKenna, S. Jabri, Z. Duric, and A. Rosenfeld, "Tracking groups of people," Computer Vision and Image Understanding, no. 80, pp. 42-56, 2000.

[8] Jie Shao Zhou, K. and Chellappa, R., "Tracking Algorithm Using Background-Foreground Motion Models and Multiple Cues," IEEE Conf. ICASSP, vol.2, pp. 233- 236, 2005

[9] B. Lucas and T. Kanade, "An Iterative Image Registration Technique with an Application to Stereo Vision," International Joint Conference on Artificial Intelligence, pp. 674-679, 1981.