International Journal of Signal Processing, Image Processing and Pattern Recognition
Vol.8, No.10 (2015), pp.233-240
http://dx.doi.org/10.14257/ijsip.2015.8.10.25

Saliency Map for Object Tracking

Dongping Zhang[1], Wenting Li[1], Min Sun[1] and Haibin Yu[2]

[1]College of Information Engineering China Jiliang University, Hangzhou 310018, China

[2]College of Electronic & Information, Hangzhou Dianzi University

silenttree_zju@cjlu.edu.cn

Abstract

Most of the state-of-the-art tracking algorithms rely on either intensity or color information. Therefore, seeking an effective image descriptor to construct a robust tracker is still a challenging problem. This paper investigates the contribution of saliency map for object tracking, and proposes a saliency detection method which is combined with the location information. Our method formulates the tracking framework based on a Bayesian framework, which models the statistical correlation between the target and the surrounding background with saliency values and location information. Additionally, our results show that the proposed method provides superior performance in terms of accuracy, robustness, and speed.

Keywords: *object tracking, saliency detection, real-time*

1. Introduction

As a foundation topic, object tracking plays an important role in the field of computer vision, and it is widely applied in surveillance, activity analysis, classification, intelligent traffic control, and human-computer interfaces, etc. Although many tracking algorithms [2-14] have been proposed, due to numerous factors in real life, when designing an object tracking algorithm, many challenging problems, such as partial occlusions, illumination changes, sophisticated object shape, background clutter and viewpoint variation, still exist.

In general, an object tracking system is composed of target representation scheme, search mechanism and mode update. And object representation is one of dominating components in any visual tracker [1]. According to the representation schemes, tracking algorithms can be categorized into either generative or discriminative model with different definitions. Generative trackers determine the target location by searching for the target which is most similar to the learned appearance model. Therefore the appearance model is significant for the tracking algorithm which is usually modeled by statistical principle [2-4], sparse representation [5-7], incremental learning [8], etc. Discriminative trackers treat the tracking problems as a detection task with the purpose of distinguishing the target from the surrounding background by a binary classification. In [9], a discriminative appearance model proposed by the utilization of structural information with the perspective of mid-level vision. Dinh et al. [10] introduced context information to online tracking which expressed by supporters and distracters. Zhang et al. [11] employed non-adaptive random projections to extract multi-scale image features for the appearance model. Babenko et al. [12] presented a tracking approach by using the online multiple instance learning, which considers the positive and negative bags in the meanwhile. In [13], an online approach has been proposed to learn a kernelized structured support vector machine (SVM) for object tracking.

Most of the state-of-the-art tracking algorithms either rely on intensity or color information as image descriptor [11, 14]. In contrast to visual tracking, saliency map has

ISSN: 2005-4254 IJSIP

shown outstanding performances on object image segmentation [15], object recognition, and video compression [16], etc. This paper investigates the contribution of saliency map for object tracking, and proposes an online tracking algorithm by combining saliency map with spatio-temporal context learning [17]. We exploit saliency information to establish context saliency model, and in order to meet the requirement of tracking, we improve saliency detection method which is proposed in [18] by utilizing the location information.

2. The STC Tracker

Our approach is based on the STC tracker [17], which exploits the spatio-temporal context learning for object tracking. The STC tracker first learned a spatio-temporal context model and calculated the confidence map in the next frame. Then the target location is determined by maximizing the new confidence map. And used Fast Fourier Transform to fast learn and detect.
And in order to learn and detect fast, they adopted Fast Fourier Transform. Here we introduce the algorithm [17] briefly.

In the current frame, we know the target location x^*. And $\Omega_c\left(x^*\right)$ denotes the neighborhood of the location x^*, $c(x)$ is the context feature defined as $c(x)=(I(x),x)$. The STC tracker treats the tracking problem as a process of computing a confidence map. The location of target in $(t+1)$ frame is computed by

$$\begin{aligned} x_{t+1}^* &= \arg\max_{x\in\Omega_c\left(x^*\right)} con_{t+1}(x) \\ &= F^{-1}\left(F\left(H_{t+1}^{stc}(x)\right)\cdot F\left(P\left(c(x)\,|\,o\right)\right)\right) \end{aligned} \tag{1}$$

Where F and F^{-1} denote the Fast Fourier Transform and inverse transform respectively. $H_{t+1}^{stc}(x)$ denotes the saptio-temporal context model. The condition probability $P(c(z)\,|\,o)$ denotes context prior model which models appearance of the local context. During the work, the main task is to learn the saptio-temporal context model in the next frame.

3. Saliency for object track

In [17], they defined the context prior model only via image intensity and a weighted function. The higher context prior probability is obtained from the closer the distance between the context location and the tracking target. In this paper, we redefine the context prior model by saliency map which follows the focus of attention model better.

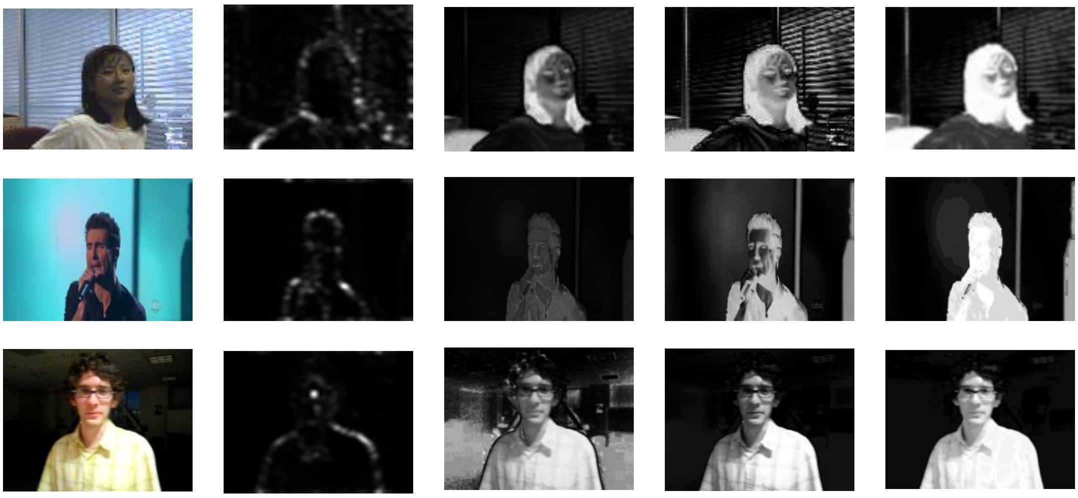
3.1. Saliency Detection

Humans have a remarkable ability to judge the importance of image regions, and focus attention on outstanding parts efficiently. Saliency map originates from visual uniqueness, unpredictability, rarity, surprise, and is often attributed to variations in image attributes like color, gradient, edges, and boundaries [18].

In 1998, Ltti *et al.* [19] first proposed the model of saliency-based visual attention which is inspired by the behavior and the neuronal architecture of the primates. Hou et al. [20] proposed a simple and efficient method for visual saliency detection which detected the saliency map based on spectral residual. They analyzed the input image on the log spectrum, and extracted the residual spectrum from the image in the frequency domain. Finally, they constructed the corresponding saliency map in spatial domain. Achanta et al. [21] proposed a frequency-tuned method which uses features of color and luminance to estimate center-sound contrast.
This method generated a saliency map which uniformly covered the whole object,

rather than the one with low resolution and poorly defined borders which is obtained by [20]. Zhai et al. [22] introduced a hierarchical spatial attention model. By using a color histograms and a pre-computed color difference table to define pixel-level saliency. Cheng et al. [18] proposed a global based method for saliency detection which relies on heterochromia from all other image pixel to produce full resolution saliency maps.

In order to apply saliency for object tracking, we need to obtain the saliency maps with full resolution rather than higher saliency values produced near the edges. Meanwhile, we hope that the algorithm is efficient and robust enough. We compare the above methods on some object tracking databases, and show the experimental results in Figure 1 and Table 1. By analyzing the results, we can conclude that the HC method is superior in terms of both efficiency and effect for object tracking.



（a）original (b) SR (c) FT (d) LC (e)HC

Figure 1. Visual Comparison of Saliency Maps

Table 1. Average Time Taken to Compute a Saliency Map for Images in the Database of David

Method	SR	FT	LC	HC
Time（s）	0.064	0.017	0.023	0.019

3.2. Context Saliency Model

In STC tracker [17], the context prior probability is simply modeled by image intensity and a weighted function, which can not describe the appearance of the target well. The biologic vision system is sensitive to high-contrast, so we can obtain saliency values of image pixels from the color histogram. Sometimes pixels of the same color with the same saliency values are adverse to object tracking. We improve the method of [18] proposed by utilizing the information of target position. The saliency value of pixel at position x is redefined by

$$S(I_x)=\sum_{j=1}^{n} w(x-x^*) f_j D(c_i,c_j) \tag{2}$$

Where c_i is the color value of position x, f_j is the probability of c_j appearance in image I, n is the number of colors, $D(c_i,c_j)$ is the color distance metric between colors c_i and c_j in $L*a*b$ space, $w(x)$ is distance weighting function defined by

$$w(x) = e^{-\frac{|x|^2}{\sigma^2}} \tag{3}$$

Where σ is the size of bounding box.

Therefore, we define the context saliency model as the following:

$$\begin{aligned} P(c(z) \mid o) &= aS(I_x) \\ &= a\sum_{j}^{n} w(x - x^*) f_j D(c_i, c_j) \end{aligned} \tag{4}$$

Where a is normalization coefficient to limit the conditional probability between 0 and 1. It is more suitable for focus of attention model by utilizing the saliency map to model the target appearance. The higher the saliency values calculated at location x, the higher probability can be obtained.

3.3. The Algorithm Process

In this paper we extend the STC tracker with saliency detection. Humans can track the target of interest in complex scenes, since they are good at determine the saliency area of the image accurately and focus on the area of importance for cognitive processes adaptively. In the stage of saliency detection, we improve the proposed method in [18] with location information, and use it for context saliency model. The redefined model is more in line with the principles of the human visual attention. The algorithm is given as follows:

Input: at the t-th frame, target location x^*, $\Omega(x^*)$ denotes the neighborhood of the location x^*, and $con_t(x)$ is the confidence map.

(1) At the $(t+1)$-th frame, the context saliency model can be obtained by formula(5);

(2) At the t-th frame, we can lean the spatial context model by

$$\begin{aligned} h_t^{sc} &= P(x \mid c(z), o) \\ &= F\left(\frac{F(con_t(x))}{F(P(c(z) \mid o))}\right) \end{aligned} \tag{5}$$

(3) Update spatio-temporal context model at the $(t+1)$-th frame by

$$H_{t+1}^{stc} = (1-\rho)H_t^{stc} + \rho h_t^{sc} \tag{6}$$

(4) So at the $(t+1)$-th frame, the location of target is calculated by maximize the confidence map as follow:

$$\begin{aligned} x_{t+1}^* &= \arg\max_{x \in \Omega_c(x_t^*)} con_{t+1}(x) \\ &= \arg\max_{x \in \Omega_c(x_t^*)} F^{-1}\left(F\left(H_{t+1}^{stc}(x)\right) \cdot F\left(aS_{t+1}(I_x)\right)\right) \end{aligned} \tag{7}$$

Especially, in the first frame the confidence map is calculated by

$$con_1(x) = ae^{-\left|\frac{x - x_1^*}{\sigma}\right|} \tag{8}$$

(5) Back to step (1) to process the next frame.

4. Experiments

In this section, we test our tracker on several challenging video sequences, to testify the robustness and efficiency of our proposed algorithm. The common databases we used include illumination change, high brightness, target deformation, target occlusion and outdoor complex scenes and so on. At the same time, our algorithm is compared with other three advanced algorithms (eg. CT[11], STC[17], CSK[23]). All the algorithms run with a MATLAB implementation on i5-2320 Core machine with 3.0GHZ CPU and 4.00GB RAM.

4.2. Experimental Results

In visual tracking, three general indicators are average center location errors, distance precision and success rate. The center location error is definition as the average Euclidean distance between the location of the target center and the ground truth. The distance precision is computed by the center location error with a threshold. And the success rate is identified as $s = \left| \frac{r_t \cap r_g}{r_t \cup r_g} \right|$, where r_t is the tracking results and r_g means the bounding box of ground truth .

Table 2. Average center location errors (pixels). Algorithms compared with CT [11], STC [17] and CSK [23]. The best performing method is shown in bold, and the second one is shown in slant.

Sequence	CT	CSK	STC	our
walking	*7*	*7*	*7*	**5**
lemming	*12*	16	16	**7**
Girl	19	19	22	**12**
football1	*11*	17	73	**10**
david	36	18	*12*	**8**
Boy	22	*20*	26	**4**
singer1	16	12	**6**	*7*
Doll	*5*	**2**	**2**	**2**
Average	16	14	20	7

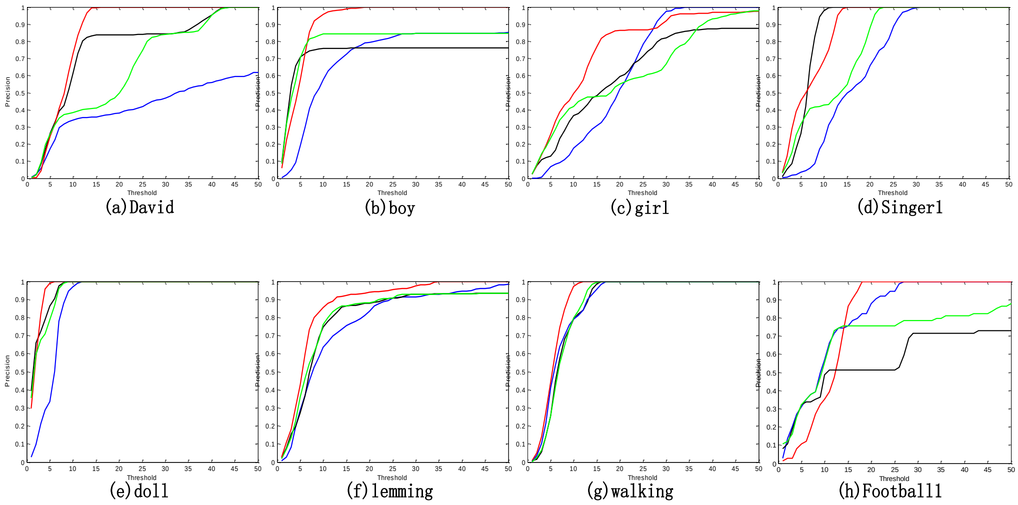
4.2. Quantitative and Qualitative Evaluation

We compare our method with 4 different trackers shown to provide favorable results in literature. Table 2 and Table 3 demonstrate the average center location error and success rate over all 8 video sequences respectively. The experimental results show that our proposed method achieves the best tracking performance on most test sequences except *singer1* sequence due to the lager scale change. By comparing the tests result, we found that our method possesses more robust performance under the condition of abrupt motion, illumination change, and pose variation, etc.

Table 3. Success rate (%). The fonts of bold indicate the best performance while the fonts of slant indicate the second best ones.

Sequence	CT	CSK	STC	our
walking	72	62	*85*	**98**
lemming	79	*87*	51	**93**
girl	36	49	*50*	**83**
football1	*74*	*74*	52	**83**
david	36	68	*80*	**93**
boy	61	*84*	66	**95**
singer1	62	67	**90**	*82*
doll	**100**	*85*	**100**	**100**

average	65	72	72	91

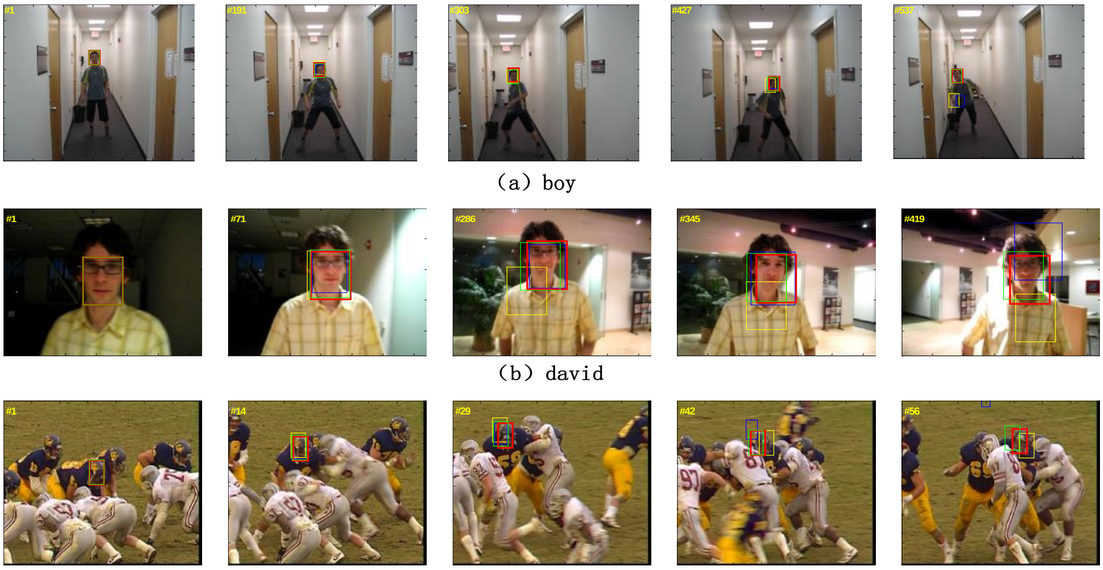


(a)David (b)boy (c)girl (d)Singer1

(e)doll (f)lemming (g)walking (h)Football1

Figure 2. Precision plots for 8 sequences. The black, green and blue lines indicate the STC tracker, CSK tracker and CT method respectively. The line of red represents our method.

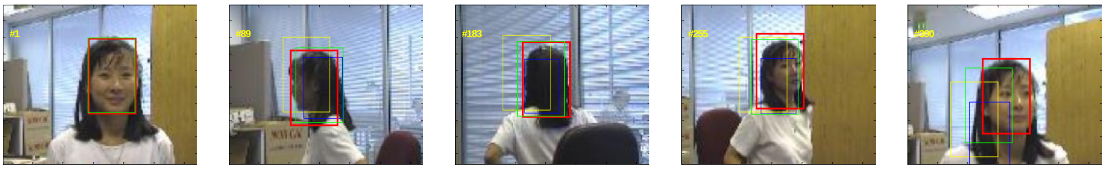
Figure 3 shows the Screenshots of tracking results which include our method and other three methods [11, 17, 23]. In the boy sequence, the major problem of tracking is the target's abrupt motion. After the frame of #479, the STC tracker first lost the target, and then the CT tracker lost the target at #510. The David sequence mainly include pose and illumination variation, the proposed tracker and the CSK tracker are able to track to the end. The football1 sequence contains occlusion, background clutter as well as abrupt motion, only our method can track the target to the end. In the girl sequence, our method obtains accurate result after a series of in-plane rotation.



(a) boy

(b) david

(c)football1



(d) girl

Figure 3. Screenshots of tracking results. The rectangle of the blue, green, yellow and red represent CT [11], STC [17], CSK [23] and our method respectively.

5. Conclusion

In this paper, we present an object tracking method which combines saliency map with spatio-temporal context tracker. By using saliency map and the information of target position, our tracking method is robust for in-plane rotation, abrupt motion, and illumination variation. We show that our method gives superior performance and runs at real-time speeds via a series of experiments.

Acknowledgements

This work was supported by Zhejiang Provincial Science & Technology Research Program (Grant No. 2013C33046, 2014C33075), Zhejiang Provincial NSF (Grant No.Y1110506, LY13H180011), and National Natural Science Foundation of China (Grant No. 61102132).

References

[1] Y. Wu, J. Lim, and M.-H. Yang. Online object tracking: A benchmark. In CVPR, (**2013**).

[2] S. Kwak, W. Nam, B. Han, and J. H. Han. Learning occlusion with likelihoods for visual tracking. In ICCV, (**2011**).

[3] B. Han, Y. Zhu, D. Comaniciu, and L. S. Davis. Visual tracking by continuous density propagation in sequential Bayesian filtering framework. PAMI, 31(5):919–930, (**2009**).

[4] B. Han, L. Davis. On-line density-based appearance modeling for object tracking. In ICCV, (**2005**).

[5] B. Liu, J. Huang, L. Yang, and C. Kulikowsk. Robust Tracking using Local Sparse Appearance Model and K-Selection. In CVPR, (**2011**).

[6] T. Zhang, B. Ghanem, S. Liu, and N. Ahuja, "Robust Visual Tracking via Multi-task Sparse Learning", In CVPR, (**2012**).

[7] X. Mei and H. Ling, "Robust visual tracking and vehicle classification via sparse representation", PAMI, vol. 33, no. 11, (**2011**), pp. 2259-2272.

[8] D. Ross, J. Lim, R.-S. Lin and M.-H. Yang, "Incremental Learning for Robust Visual Tracking", IJCV, vol. 77, no. 1, (**2008**), pp. 125–141.

[9] S. Wang, H. Lu, F. Yang, and M.-H. Yang, "Superpixel tracking", In ICCV, (**2011**).

[10] T. B. Dinh, N. Vo, and G. Medioni, "Context Tracker: Exploring Supporters and Distracters in Unconstrained Environments", In CVPR, (**2011**).

[11] K. Zhang, L. Zhang, and M.-H. Yang, "Real-time Compressive Tracking", In ECCV, (**2012**).

[12] B. Babenko, M.-H. Yang, and S. Belongie. Robust Object Tracking with Online Multiple Instance Learning. PAMI, vol. 33, no. 7, (**2011**), pp. 1619–1632.

[13] S. Hare, A. Saffari, and P. H. S. Torr, "Struck: Structured Output Tracking with Kernels", In ICCV, (**2011**).

[14] M. Danelljan, F. S. Khan, M. Felsberg, and J. V. D. Weijer, "Adaptive Color Attributes for Real-Time Visual Tracking. In CVPR, (**2014**).

[15] Y. Tian, J. Li, S. Yu, and T. Huang, "Learning complementary saliency priors for foreground object segmentation in complex scenes", IJCV, (**2014**), pp. 1-18.

[16] C. Guo and L. Zhang, "A Novel Multiresolution Spatiotemporal Saliency Detection Model and Its Applications in Image and Video Compression", IEEE Trans. Image Processing, vol. 19, no. 1, (**2010**), pp. 185-198.

[17] K. Zhang, L. Zhang and M.-H Yang, "Fast Tracking via Spatio-Temporal Context Learning. arXiv:1311.1939, (**2013**).

[18] M.-M. Cheng, G.-X. Zhang, N. J. Mitra, X. Huang, and S.-M. Hu, "Global contrast based salient region detection", In CVPR, (**2011**).

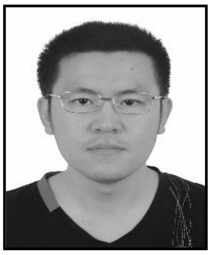
[19] L. Itti, C. Koch, and E. Niebur, “A model of saliency-based visual attention for rapid scene analysis”, *PAMI*, vol. 20, no. 11, (**1998**), pp. 1254-1259.
[20] X. Hou and L. Zhang, “Saliency detection: A spectral residual approach”, In *CVPR*, (**2007**).
[21] R. Achanta, S. Hemami, F. Estrada, and S. Susstrunk, “Frequency-tuned salient region detection”, In CVPR, (**2009**),pp. 1597–1604.
[22] Y. Zhai and M. Shah, “Visual attention detection in video sequences using spatiotemporal cues”, In ACM Multimedia, (**2006**).
[23] J. F. Henriques, R. Caseiro, P. Martins, and J. Batista, “Exploiting the Circulant Structure of Tracking-by-Detection with Kernels”, In ECCV, (**2012**).

Authors

Dongping Zhang, He was born in 1970. He received the Ph.D. in Information & Communication Engineering from Department of Information Science & Electronic Engineering, Zhejiang University, Hangzhou, China, in 2006. Since 2006, He is an associate professor at College of Information Engineering, China Jiliang University in Hangzhou. His research interests include image processing and pattern recognition, computer vision and videos.

Wenting Li, She graduated at the China Jiliang University in Hangzhou in 2013. Currently she is a MSC. She is a student at the College of Information Engineering of China Jiliang University. Her research interests are oriented to image processing, saliency detection, and visual tracking.

Min Sun, she received the B.Sc. degree in Electronic information science and technology from XinZhou Normal University, China, in 2013. She is currently a graduate student in the School of China Jiliang University, China. Her research interest is Video analysis about the crowd behavior recognition and the crowd density estimation.



Haibin Yu, He received Ph.D. in Communication & Information System from the Department of Information Science & Electronic Engineering, Zhejiang University in 2007. Since 2007, he has been a university teacher in the College of Electronic & Information, Hangzhou Dianzi University, Hangzhou, China. In 2010, he was promoted to the associate professor in Electronic Science & Technology. His research interests include image processing, pattern recognition, computer vision, video surveillance and signal processing