Target tracking with composite linear filters on noisy scenes

V. Kober[a], V. Kuznetsov[a]

[a] *Chelyabinsk State University, 454001, 129 Br. Kashirinikh, Kashirinikh, Chelyabinsk, Russia*

Abstract

A tracking system using a bank of adaptive linear filters is proposed. Tracking is carried out by means of multiple target detections. The linear filters are designed from multiple views of a target using synthetic discriminant functions. For each view an optimum filter is derived from noisy reference image and disjoint background model. An iterative algorithm is used to improve the performance of the synthesized filters. The number of filters in the bank can be controlled to guarantee a prescribed tracking accuracy. Computer simulation results show that the proposed algorithm is able to precisely track a target.

Keywords: object tracking; correlation filters; composite filters

1. Introduction

Object tracking systems are used for applications such as video surveillance, motion based recognition, and vehicle navigation [1]. Tracking requires processing large amounts of data. Two approaches can be taken: to reduce the amount of information to be processed and to carry out the processing faster. In the former approach, features are usually computed. A feature extractor ideally outputs a small number of features. Matching these features across frames yields the displacement information. When the camera rate is high, preprocessing might be done by subtracting the background of a given frame from the next one, so that the only information is left in the frame in the area where movement took place. On the other hand, tracking can be done on the original image without extracting features, by using appearance-based models [1, 2]. This approach requires that either the appearance of a target does not change much along the frame sequence or possible views of the target should be known a priori. This results in a high computational cost. Optical processors were used extensively for object detection [3]. They perform a fast detection by exploiting the parallelism inherent in optical systems. In the case when tracking requires only knowing the position of the object at a given time, the problem may be solved by using a consecutive detection approach [4]. The use of optical correlators for fast object detection allows real-time tracking applications. The approach of tracking by successive detection has several advantages. Correlation filters can be designed to analytically minimize the probability detection errors, thus detection in each scene is optimum with respect to a detection criterion. Additionally, correlation filters can be designed to minimize errors in the estimation of the target location. Furthermore, since the target is being detected in each scene, there is no problem with situations when the target is being temporarily occluded in the scene because it can be correctly detected upon reentering the scene.

Composite correlation filters were proposed for taking into account multiple views of a target in a single correlation operation [5]. An optimum correlation filter can be designed for each view of the target, and filters for multiple views can be combined into a single composite filter [6-9].

If information about a background where detection will be carried out is available, the discrimination capability (DC) of composite filters can be improved using an adaptive approach [10]. It has been shown that the performance of composing filters degrades with an increasing number of views. This problem can be solved by using a bank of composite filters when each of them is designed with a subset of known views of the target. Filters in the bank are then applied in rapid successive correlations, and the maximum value over all correlation output planes is chosen as the estimation of the location of the target. The number of filters in the bank is chosen to ensure a required accuracy. In other words, the parameter space of possible distortions is divided in such a way to always get an error of the position estimation less than a prescribed value with the minimum number of correlations.

The presentation is organized as follows. In Section 2, design of a correlation filter-based optoelectronic tracking system is presented. Computer simulations are given and discussed in Section 3. Finally, our conclusions are summarized in Section 4.

2. Design of tracking system

2.1. Problem formulation

One-dimensional notation is used for simplicity. Let us consider a discrete sequence of images where $s_i(x)$ denotes an i-th image. Let $t(x;\theta)$ denote a target. θ is a vector of parameters that determine the appearance (distortion) of the target, such as rotation or scaling. Let x_i and θ_i denote the position and the appearance of the target in the i-th image, respectively. The problem consists of calculating an estimation $\hat{x}_i$ of the target position in the i-th image. No assumption is made on the relation between x_i and x_{i+1}. Location estimation is performed using a bank of composite filters described by transfer functions $\{H_j(\omega), j=1,2,...N_h\}$ where N_h is the number of filters in the bank. Each filter $H_j(\omega)$ is composed using training images of different views of the target.

2.2. Optimum filter for single-frame detection

In this section we derive an optimum filter for detecting a target in a single frame using a noisy view image. Let $r(x)$ denote a reference image showing one view of the target and $w(x)$ denote the shape of the target in $r(x)$. $w(x)$ takes a value of unity inside the target area and zero otherwise. The suffix is omitted in this section since only one frame is used for the design of optimum filters. So, $s(x)$ is an observed scene represented by a nonoverlapping signal model. The target is opaque and appears over a background that is spatially disjoint. Additive noise is also considered to be present the frame. Filters will be designed from a signal model in which a reference image is corrupted by additive noise. This can model the case when the reference image is captured in controlled environment with low quality equipment and no processing is done on the captured images. The model is formally defined as

$$r(x) = t(x - x_r) + n_r(x) \tag{1}$$

$$s(x) = t(x - x_s) + b(x)w(x - x_s) + n_s(x), \tag{2}$$

where $b(x)$ denotes a nonoverlapping background in the scene; $n_r(x)$ and $n_s(x)$ represent additive noises in the reference image and the input scene, respectively. x_r and x_s are the coordinates of the target in the images. $b(x)$ is a stationary random process with the mean value μ_b. The zero-mean process $b_0(x) = b(x) - \mu_b$ has the power spectral density $B_0(\omega)$. Additive noise is modeled as a zero-mean stationary process. The view parameter θ is not present in this section because a filter is designed for one particular view given in the reference image; therefore, there is no need to distinguish between different views. All random processes and random variables in the model are considered as statistically independent of each other.

The scene model given in (2) implies that pattern recognition must be performed by matching the input scene with a new target formed by the target and the weighted inverse support function $\mu_b w(x)$. Let $t_s(x)$ denote the composite target

$$t_s(x) = t(x) + \mu_b w(x), \tag{3}$$

and $\tilde{n}_s(x, x_s)$ denote the nondeterministic part of the scene

$$\tilde{n}_s(x, x_s) = b_s^0(x)w(x - x_s) + n_s(x). \tag{4}$$

If the target and its support function are explicitly known, the transfer function of the generalized optimum filter (GOF) is the best with respect to the peak-to-output energy (POE) [7-9]:

$$GOF(\omega) = \frac{T_s^*(\omega)}{|T_s(\omega)|^2 + (2\pi)^{-1} B_s^0(\omega) \bullet |W(\omega)|^2 + N_s(\omega)}, \tag{5}$$

where $T_s(\omega)$ and $W(\omega)$ are the Fourier transforms of $t_s(x)$ and $w(x)$, respectively, $N_s(\omega)$ is the power spectral density of $n_s(x)$, and $\bullet$ denotes the convolution operation. In (1) the target signal is given implicitly in the reference image. Therefore, it is required to approximate the GOF using the information available in the reference image. Let $\hat{w}(x)$ denote an estimation of $w(x)$ that can be obtained from the reference image. An estimation of $T_s(\omega)$ can be obtained by applying a linear filter to the reference image and then to adding the weighted estimation of the inverse support function, that is

$$\hat{t}(x - x_r) = g(x) \bullet r(x) + \mu_b \hat{w}(x - x_r), \tag{6}$$

where $g(x)$ is the impulse response of the linear filter used for image filtering. Under the minimum mean square error (MMSE) criterion, the optimum filter for eliminating additive noise is the smoothing Wiener filter. Let $\hat{T}(\omega)$ and $\hat{W}(\omega)$ denote the Fourier transforms of $\hat{t}(x)$ and $\hat{w}(x)$, respectively. Substituting $\hat{T}(\omega)$ and $\hat{W}(\omega)$ into (5), an approximation of the optimum filter $\hat{GOF}$ can be obtained.

2.3. Composite filter design for distortion invariant recognition

An optimum filter, with respect to the POE criterion, can be designed for each view of the target available for training. A modified synthetic discriminant function (SDF) algorithm is used for designing a composite filter from the impulse responses of multiple optimum filters [6, 10]. Let $\hat{gof}_i(x)$ denote the inverse Fourier transform of the conjugate of $GOF_i(\omega)$. $GOF_i(\omega)$ is the optimum filter for the i-th view of the target, $i \in \{1, 2, \ldots, N_v\}$ where N_v is the number of views available for training. Let $\hat{t}_s^i(x)$ denote the estimation given in (6) used in the design of $\hat{gof}_i(x)$, and let $\{\alpha_1, \alpha_2, \ldots, \alpha_{N_h}\}$ be N_h subsets taken from the

set of training images. Here $|\alpha_j|$ denotes the size of the j-th subset. Each subset is used to compose the impulse response of the j-th composite filter $h_j(x)$. $h_j(x)$ is a SDF filter for distortion invariant detection obtained as a linear combination of the vectors $g\hat{o}f_i(x)$, $i \in \alpha_j$ and undesired objects to be rejected. The coefficients in the linear combination are chosen to satisfy constraints on the filter output. $h_j(x)$ is expressed as

$$h_j(x) = \sum_{i \in \alpha_j} a_i g\hat{o} f_i(x) + \sum_{i=1}^{N_f} b_i p_i(x), \tag{7}$$

where a_i and b_i are coefficients, N_f is the number of patterns to be rejected, and $p_i(x)$ is the appearance of the i-th patterns. Now suppose that all signals are discrete. Let $\mathbf{M}_j$ be a matrix with $|\alpha_j| + N_f$ columns, where the columns are the vectors $g\hat{o}f_i(x)$ and $p_i(x)$, respectively. Let $\mathbf{a}$ denote a column vector formed by $|\alpha_j|$ coefficients a_i and N_f coefficients b_i. Using vector-matrix notation, the vector form of $h_j(x)$ is given by

$$\mathbf{h}_j = \mathbf{M}_j \mathbf{a}. \tag{8}$$

The equality constraints are given by a column vector $\mathbf{u}$ formed by $|\alpha_j|$ ones followed by N_f zeros. Let $\mathbf{T}_j$ denote $|\alpha_j| + N_f$ column matrix formed by the vectors $t_s^i(x)$ and $p_i(x)$. The weighing coefficients are chosen to satisfy the following condition:

$$\mathbf{u} = \mathbf{T}_j^{+} \mathbf{h}_j, \tag{9}$$

where $^+$ denotes conjugate transpose. From (8) and (9) the resulting expression for $\mathbf{h}_j$ is given by

$$\mathbf{h}_j = \mathbf{M}_j \left(\mathbf{T}_j^{+} \mathbf{M}_j \right)^{-1} \mathbf{u}. \tag{10}$$

When the background image is available, an adaptive algorithm can be used to improve the DC of composite filters [14]. The background can be described either stochastically, treated as a realization of a stochastic process, or deterministically in the form of a typical background picture. The background used in the adaptive algorithm can also contain false objects or structures similar to the target. The DC is used to characterize the ability of a filter to distinguish the target from other objects in the scene that may have similar appearance. The DC is defined as

$$DC = 1 - \frac{\left| y_{max}^B \right|^2}{\left| y_{max}^T \right|^2}, \tag{11}$$

where y_{max}^B is the maximum value in the correlation plane over the area that is occupied by the background, while y_{max}^T is the maximum value in the correlation plane over the area occupied by the target. The background area and the target area are complementary. Ideally, the values of the DC should be close to unity, which indicates a good capacity to discriminate the target against unwanted objects. Negative values of the DC indicate a failure to detect the target.

The patterns to be rejected in (7) can be obtained with an iterative algorithm. Initially a filter is composed using only the images of the target. At each step a filter is applied to the sample background. A pattern is then synthesized from the background at the location of the highest value in the filter output. This pattern is then added to the training set to be rejected, and a new filter is created. The new filter ensures zero output at the location of the rejected pattern. In this manner, the value of the DC for the new filter is monotonically increasing.

2.4. Optoelectronic tracking system

Hybrid optoelectronic systems can be used to implement correlation filter-based processors [3]. Hybrid systems have two basic architectures: 4f correlator (4FC) and joint transform correlator (JTC). An advantage of the JTC over the 4FC is that the former is less sensitive to misalignments of an optical setup such as scale, horizontal, vertical, and azimuthal differences between the input and frequency planes. The input plane (joint image) consists of a scene image alongside the template used for filtering. A block diagram of the tracking system is shown in Fig. 1. For pure digital implementation prediction and fragmentation segmentation stages are included into the tracking system [11-14]. The bank of filters considers different subsets of views. The *i*-th input image is processed using all filters in the bank in rapid succession. The system output at a fixed time is the plane with maximum value of the DC. It is computed as follows: first, the highest peak in each correlation plane is found, then using the region of support of the target in each plane the highest sidelobe and the DC are calculated, and finally, the filter output with the maximum DC among all filters is taken as the system output. The target trajectory as a function of time is formed from positions of the system output peaks. For certain applications, such as aerial surveillance, the bank of filters can also account for changes in the background due to different types of terrain. The adaptive algorithm can be used with different backgrounds, representative of each type of terrain. The resulting composite filters are then included in the bank of filters.

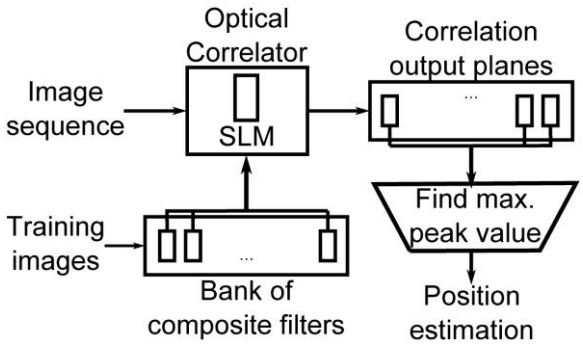

Fig. 1. Block diagram of the optoelectronic tracking system.

A simpler implementation can be implemented using a single composite correlation filter. While this presents a performance advantage in terms of processing time, location accuracy may be degraded owing to a large number of views included into a single filter. Conversely, a larger filter bank provides higher accuracy at the cost of processing time.

3. Results and discussion

In this section we analyze the performance of the proposed system for target tracking in a nonoverlapping background. We consider location accuracy as the performance criterion. Location accuracy can be characterized by the location errors (LE) [21] defined as

$$\mathrm{LE} = \sqrt{\left(x_T - \hat{x}_T\right)^2 + \left(y_T - \hat{y}_T\right)^2}, \tag{12}$$

where $\left(x_T, y_T\right)$ and $\left(\hat{x}_T, \hat{y}_T\right)$ are the coordinates of the target exact position and the position estimation, respectively, when using the coordinates of the correlation peak as the location of the target. Using filters derived from the model in (1) and (2) the estimated location of the target will not be at the exact position of the target in the input scene; instead, it will be displaced by the location of the target in the reference image. The size of all images used in the experiments is 256×256 pixels. A sample noisy reference image is shown in Fig. 2 (a); the target is a butterfly at the center of the reference image. The target has size 28×44 pixels, and has a mean value of 100 with a standard deviation of 25. Image intensities are in the range $\left[0,255\right]$. A test input scene is shown in Fig. 2(b).

(a) (b)

Fig. 2. Images used in experiments. (a) Sample reference image, (b) sample scene.

Image sequences are generated by the following algorithm:

Generate a random initial position x_s, rotation τ, and scale σ for the target.

Generate a random direction vector in polar form $v = \left(\rho, \theta\right)$.

Generate a random sequence $\Delta_i\tau$ with the correlation coefficient 0.95 used as perturbation for τ.

Generate uncorrelated zero-mean random sequences $\Delta_i\rho$, $\Delta_i\theta$, and $\Delta_i\sigma$ used as perturbations for ρ, θ, σ, respectively.

For each frame i:

Rotate the target on τ degrees, scale it by a factor of σ, and place it at the location x_s in the background.

Add white noise to the scene.

Update the target location $x_s \leftarrow x_s + v$.

Update $\tau \leftarrow \tau + \Delta_i\tau$, $\rho \leftarrow \rho + \Delta_i\rho$, $\theta \leftarrow \rho + \Delta_i\theta$, and $\sigma \leftarrow \Delta_i\sigma$.

The target is allowed to rotate freely, while its scale is kept in the range $\left[0.8,1.2\right]$. The training images contain distorted versions of the target scaled by factors 0.9, 1.0, 1.1 and rotated by -9, -3, 3, and 9 degrees. A background used in training with the adaptive algorithm. Note that the background is not the same with that of used in tracking experiments. One hundred image sequences generated each with 100 frames. The composite filters are designed using noisy reference images corrupted by

additive noise with a standard deviation of 10. Two system variants are tested. Let us A denote a system variant that uses one composite filter including all views available for training. Let us B4, B6, and B9 denote systems that use banks of 4, 6, and 9 composite filters, respectively. The performance of the filters is shown in Fig. 3

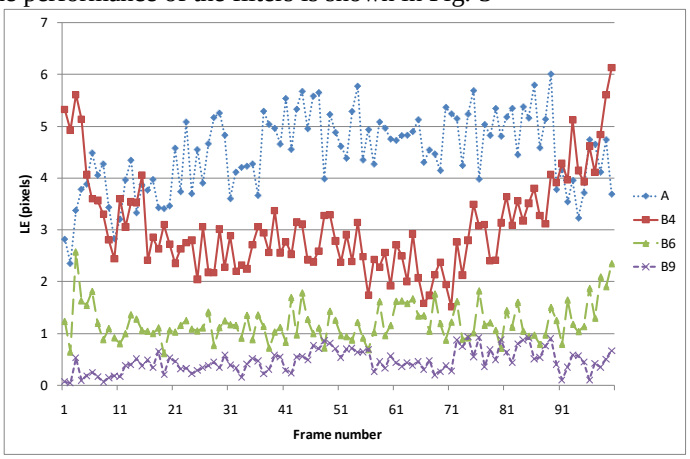

Fig. 3. Location errors for each frame in image sequence for system using one, four, six, and nine composite filters.

The plot shows the location errors per frame averaged over all the generated sequences. It can be seen that the system A is able to track the position of the target with an average error of 4.5 pixels. It can also be seen how the errors decrease when the number of filters in the bank is increased. For the test images, nine composite filters are required to achieve the location accuracy of one pixel or less.

4. Conclusion

In this paper we proposed an optoelectronic system for a real-time object tracking. Correlation filters for locating a target in nonoverlapping background noise were used for detecting the target in each frame in the image sequence. A bank of filters adapted to different possible views and typical backgrounds was proposed to achieve accurate tracking. Optimum filters were designed from training images that do not need to be manually processed. With the help of computer simulations, we showed that the proposed system is able to estimate the target's trajectory through the image sequence. Two variants of the proposed system showed that there exists a trade-off between simplicity (processing time) and tracking accuracy. It was shown that increasing the number of correlations per frame leads to accurate estimation of the target trajectory.

Acknowledgements

This work was supported the Russian Science Foundation grant №15-19-10010.

References

[1] Yilmaz, A., Javed, O., Shah, M. Object tracking: A survey // ACM Computing Surveys. – 2006. – Vol. 38(4). –Art. ID. 13.
[2] Davey, S.J., Rutten, M.G., Cheung, B. A comparison of detectionperformance for several track-before-detect algorithms // EURASIP J. Adv. Signal Process. – 2008. – Vol. 2008. – Art. ID 428036.
[3] Diaz-Ramirez, V.H., Kober, V.Adaptive phase-input joint transform correlator // Applied Optics.- 2007.-Vol. 46(26).- P. 6543-6551.
[4] Bruno, M.G.S., Araújo, R.V., Pavlov, A.G. Sequential Monte Carlo Methods for Joint Detection and Tracking of Multiaspect Targets in Infrared Radar Images // EURASIP J. Adv. Signal Process.- 2007. – Vol. 2008. – Art. ID 217373.
[5] Casasent, D. Unified synthetic discriminant function computational formulation // Appl. Opt.-1984.-Vol. 23.-P.1620-1627.
[6] Ramos Michel, M., Kober, V. Design of correlation filters for recognition of linearly distorted objects in linearly degraded scenes // Journal OSA A.- 2007.- Vol. 24(11).-P. 3403-3417.
[7] Aguilar-Gonzalez, P.M., Kober, V. Design of correlation filters for pattern recognition using a noisy reference // Optics Communications.- 2012.- Vol. 285.-P. 574–583.
[8] Aguilar-Gonzalez, P.M., Kober, V. Design of correlation filters for pattern recognition with disjoint reference image // Optical Engineering.- 2011.- Vol. 50(11).-P. 117201-8.
[9] Kober, V., Aguilar-Gonzalez, P.M., Karnaukhov, V. Automated object detection with a correlation filter designed from a noisy image // Journal of Communications Technology and Electronics.- 2014.-Vol. 59(6).-P. 571–575.
[10] Aguilar-Gonzalez, P.M., Kober, V, Diaz-Ramirez, V.H. Adaptive composite filters for pattern recognition in nonoverlapping scenes using noisy training images // Pattern Recognition Letters.- 2014.- Vol. 41.-P. 83–92.
[11] Diaz-Ramirez, V.H., Picos, K., Kober, V. Target tracking in nonuniform illumination conditions using locally adaptive correlation filters // Optics Communications.- 2014.- Vol. 323.-P. 32–43.
[12] Diaz-Ramirez, V.H., Contreras, V., Kober, V., Picos, K. Real-time tracking of multiple objects using adaptive correlation filters with complex constraints // Optics Communications.- 2013.- Vol. 309.-P. 265-278.
[13] Ontiveros-Gallardo, S.E., Kober, V. Correlation-based tracking using tunable training and Kalman prediction/ Andrew G. Tescher // Proc. SPIE's 61 Annual Meeting.-San Diego: SPIE, 2016.- V. 9971.- P. 997129-9.
[14] Ruchay,A., Kober, V. A correlation-based algorithm for recognition and tracking of partially occluded objects/ Andrew G. Tescher // Proc. SPIE's 61 Annual Meeting.-San Diego: SPIE, 2016.- V. 9971.- P. 99712R -9.