Optics Communications 365 (2016) 140–149

Contents lists available at ScienceDirect

Optics Communications

journal homepage: www.elsevier.com/locate/optcom

Invited Paper

Target tracking with dynamically adaptive correlation

Leopoldo N. Gaxiola [a], Victor H. Diaz-Ramirez [a,*], Juan J. Tapia [a], Pascuala García-Martínez [b]

[a] *Instituto Politécnico Nacional - CITEDI, Ave. Instituto Politécnico Nacional 1310, Nueva Tijuana, Tijuana, B.C. 22435, Mexico*
[b] *Departamento de Óptica, Universitat de València, c/Dr. Moliner 50, 46100 Burjassot, Spain*

ARTICLE INFO

Article history:
Received 27 October 2015
Received in revised form
26 November 2015
Accepted 30 November 2015
Available online 17 December 2015

Keywords:
Target tracking
Adaptive correlation filters
Online training
Opto-digital correlators

ABSTRACT

A reliable algorithm for target tracking based on dynamically adaptive correlation filtering is presented. The algorithm is capable of tracking with high accuracy the location of a target in an input video sequence without using an offline training process. The target is selected at the beginning of the algorithm. Afterwards, a composite correlation filter optimized for distortion tolerant pattern recognition is designed to recognize the target in the next frame. The filter is dynamically adapted to each frame using information of current and past scene observations. Results obtained with the proposed algorithm in synthetic and real-life video sequences, are analyzed and compared with those obtained with recent state-of-the-art tracking algorithms in terms of objective metrics.

1. Introduction

Nowadays, target tracking is a widely investigated topic in engineering and computer vision. Video surveillance, vehicle navigation, human–computer interaction, and robotics are examples of tracking applications. Target tracking consists in estimating the trajectory of a target from a sequence of observed images while the target moves through a detection zone. A main challenge in target tracking is when the observed scene is degraded with additive and disjoint noise, nonuniform illumination, blurring, and by appearance modifications of the target such as pose changes, rotations, and scaling. Additionally, eventual occlusions of the target and the need of a fast algorithm execution are important issues that a tracking algorithm must solve [1–3].

In the last few years, several algorithms have been proposed with the aim of solving real-life tracking problems [1–3]. Some of these algorithms require a strong a priori knowledge of the target for *offline* training purposes before tracking begins [4]. Other algorithms only require basic information of the target because they use learning and adaptation mechanisms to train the system *online* during tracking operation [2,5]. The latter algorithms are called *online* training algorithms and are preferred over *offline* training algorithms because they have higher flexibility. One well-known *online* training algorithm is the multiple instance learning (MIL) [6]. This algorithm operates by classifying image templates in sets of true- and false-class images collected during tracking operation. The true-class set contains image templates of the expected views of the target having a similar appearance to that of the actual view of the target in the scene. The false-class set contains scene fragments to be rejected having similar contents to those of the target. A successful state-of-the-art tracking algorithm is the structured output tracking with kernels (Struck) [7]. The Struck algorithm employs a structured support vector machine (SVM) to directly link the target's location space with the training samples collected during tracking operation. Struck has exhibited excellent results in many tracking benchmarks [2,3]; thus, it is often used as a reference for comparison with new tracking algorithms. Another important tracker is the tracking learning detection (TLD) [5]. The TLD algorithm uses a set of structural constraints with a sampling strategy that exploits a boosting classifier. The TLD algorithm is able to track the position of a target in a video sequence with tolerance to scene perturbations.

An attractive alternative to existing tracking algorithms is given by correlation filtering. Correlation filters have a good formal basis and they can be implemented for real-time applications either in hybrid opto-digital correlators [8,9] or in digital programmable devices such as graphics processing units (GPUs) [10] by exploiting massive parallelism. A correlation filter is a linear system in which the coordinates of the maximum intensity value in the output correlation function are estimates of the target's coordinates within the observed scene [11]. These filters can be designed to recognize targets in cluttering and noisy environments [12–17]. Also, they are able to estimate with high accuracy the location of a target in a scene with tolerance to nonuniform illumination and to geometrical modifications of the target [18–22].

Our hypothesis is that the performance of target tracking can be significantly improved in terms of efficiency of target detection and accuracy of location estimation of the target by applying a

[*] Corresponding author.
E-mail address: vdiazr@ipn.mx (V.H. Diaz-Ramirez).

http://dx.doi.org/10.1016/j.optcom.2015.11.077
0030-4018/© 2015 Elsevier B.V. All rights reserved.

dynamically adaptive correlation filtering to multiple frames. Various proposals for performing target tracking based on correlation filtering have been suggested [9,25–27]. The majority of these methods utilize a bank of correlation filters which are constructed *offline* before tracking process begins using available views of the target [9,25].

Recently, Bolme et al. [28] suggested the use of an adaptive correlation filtering for target tracking by exploiting an *online* training approach. This algorithm is competitive with respect to standard tracking algorithms, but with lower complexity [29,2]. In this approach, a correlation filter is used to detect and locate the target in the observed scene in each frame. The filter is updated *online* according to current and past scene observations, and by taking into account intraclass distortions of the target. The used filter in this algorithm is the minimum output sum of squared error (MOSSE) [28]. The MOSSE filter produces a prespecified output correlation plane in response to a given set of training images. The main limitation of the MOSSE filter for *online* training tracking is that since the target is in constant motion the actual coordinates of the target required for constructing the filter cannot be precisely known. This situation can introduce a considerable bias in the location estimation of the target [30].

In this work, we propose a reliable algorithm for target tracking in noisy scenes using an *online* training approach. The proposed algorithm employs a dynamically adaptive correlation filtering designed as a combination of filter templates optimized for detection and location estimation of the target in an observed scene. First, a set of geometrically distorted versions of the reference image of the target is constructed. Afterwards, an optimal filter template is designed for each created image. The filter templates are combined to form a single composite filter. One can note, that by using the suggested approach we avoid the need to specify beforehand the expected location of the target in the next frame, as in the case of MOSSE filter. The proposed tracking algorithm incorporates a prediction mechanism that exploits the temporal relationship of input frames in order to improve the tracking accuracy by taking into account the kinematics of the target. Furthermore, the proposed algorithm integrates an efficient re-initialization mechanism that automatically reestablishes the tracking if the system fails.

The paper is organized as follows. Section 2 explains the proposed filter design for target tracking. Section 3 describes the suggested algorithm for robust target tracking. Computer simulation results obtained with the proposed approach are presented and discussed in Section 4. This results are compared with those obtained with successful state-of-the-art tracking algorithms in terms of detection efficiency and tracking accuracy. Finally, Section 5 presents our conclusions.

2. Design of composite correlation filters using optimized templates

We are interested in the design of a correlation filter able to recognize a target from an observed scene when it is corrupted with additive and disjoint noise. In addition, the filter needs to be robust in recognizing different views of the target. Let $T = \{t_i(x, y);\ i = 1, \ldots, N\}$ be a set of training images given by different views of the target to be recognized. The input scene is assumed that is formed by a target $t(x, y)$ embedded into a disjoint background $b(x, y)$ at unknown coordinates (τ_x, τ_y), and the scene is corrupted with zero-mean additive noise $n(x, y)$, as follows:

$$f(x, y) = t(x - \tau_x, y - \tau_y) + b(x, y)\bar{w}(x - \tau_x, y - \tau_y) + n(x, y), \tag{1}$$

where $\bar{w}(x, y)$ is a binary function defined as zero inside the target area and unity elsewhere. The optimum filter for detecting the target from Eq. (1) in terms of the signal-to-noise ratio (SNR) [12] and the minimum variance of measurements of location errors (LE) [18], is the generalized matched filter (GMF) [18,15], whose frequency response is given by

$$H^*(u, v) = \frac{T(u, v) + \mu_b \bar{W}(u, v)}{P_b(u, v) \otimes \left|\bar{W}(u, v)\right|^2 + P_n(u, v)}, \tag{2}$$

where $T(u, v)$ and $\bar{W}(u, v)$ are the Fourier transforms of $t(x, y)$ and $\bar{w}(x, y)$, respectively; μ_b is the mean value of the background image $b(x, y)$; and $P_b(u, v)$ and $P_n(u, v)$ denote spectral density functions of $b_0(x, y) = b(x, y) - \mu_b$ and $n(x, y)$, respectively. The symbol "⊗" denotes convolution.

The use of correlation filters for target tracking is usually performed by focusing the processing on a small fragment of the input scene in each frame, where it is assumed that the target is contained. It is important to consider that the appearance of the target can be different in each observed frame.

Thus, two important issues must be addressed in order to design a reliable correlation filter for target tracking based on the filter model of Eq. (2). First, the support region function of the target $\bar{w}(x, y)$ is explicitly unknown. Also, the statistical properties of the background and additive noise processes in the scene fragment can be time-variant. As a result, the scene parameters required to synthesize the GMF in Eq. (2) must be locally estimated for each observed frame. Second, the filter must be able to recognize the target and its intraclass distortions with a single correlation operation.

It can be seen that the size of the scene fragment to be processed is small compared with the size of the whole scene image. Also, the area of the scene fragment is almost fully occupied by the area of the target within the fragment. In this case, by assuming that the region of support of the fragment is equivalent to the region of support of the target, we only have a small detection error. Hence, the GMF to detect a target from a scene fragment with an explicitly unknown support function can be approximated by

$$H^*(u, v) \approx \frac{T(u, v)}{P_b(u, v) + P_n(u, v)}. \tag{3}$$

Next, in order to correctly estimate the functions $P_b(u, v)$ and $P_n(u, v)$ for filter synthesis in Eq. (3), suppose that the background in the fragment has a separable exponential covariance function [31]; thus $P_b(u, v)$ can be computed by

$$P_b(u, v) = \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} \sigma_{b0}^2 \rho_x^{|x|} \rho_y^{|y|} \exp[-i(ux + vy)]dxdy, \tag{4}$$

where σ_{b0}^2 is the variance of $b_0(x, y)$ and ρ_x and ρ_y are correlation coefficients in the x and y directions respectively. These coefficients can be a priori known; otherwise, they can be estimated from the observed scene [4].

Now, consider that the noise-free image $r(x, y) = t_i(x - \tau_x, y - \tau_y) + b(x, y)\bar{w}(x - \tau_x, y - \tau_y)$ in the fragment, and the additive noise $n(x, y)$ are independent. Thus, the covariance function of the observed fragment is $C_f(x, y) = C_r(x, y) + C_n(x, y)$, where $C_r(x, y)$ is the covariance function of $r(x, y)$ and $C_n(x, y) = \sigma_n^2\delta(x, y)$ is the covariance function of white noise. Note that the noise variance can be estimated as $\sigma_n^2 = C_r(0, 0) - C_f(0, 0)$; however, $C_r(0, 0)$ is unknown. One can note that $C_n(x, y) = 0,\ \forall (x, y) \neq 0$. Hence, the values of $C_f(x, y);\ (x, y) \neq 0$ can be used to estimate $C_r(0, 0)$. This can be done by using linear extrapolation, as

$$C_r(0, 0) = 2C_f(0, 1) - C_f(0, 2). \tag{5}$$

Moreover, let $h_i(x, y)$ be the impulse response of a GMF