Image and Vision Computing 32 (2014) 227–236

Contents lists available at ScienceDirect

Image and Vision Computing

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

Adaptive on-line similarity measure for direct visual tracking ☆

Hadi Firouzi, Homayoun Najjaran

Okanagan School of Engineering, The University of British Columbia, Kelowna, BC, Canada

ARTICLE INFO

Article history:
Received 18 April 2013
Received in revised form 11 November 2013
Accepted 30 January 2014
Available online 8 February 2014

Keywords:
Adaptive metric
Similarity measure
Visual tracking
Template matching

ABSTRACT

This paper presents an on-line adaptive metric to estimate the similarity between the target representation model and new image received at every time instant. The similarity measure, also known as observation likelihood, plays a crucial role in the accuracy and robustness of visual tracking. In this work, an L2-norm is adaptively weighted at every matching step to calculate the similarity between the target model and image descriptors. A histogram-based classifier is learned on-line to categorize the matching errors into three classes namely i) image noise, ii) significant appearance changes, and iii) outliers. A robust weight is assigned to each matching error based on the class label. Therefore, the proposed similarity measure is able to reject outliers and adapt to the target model by discriminating the appearance changes from the undesired outliers. The experimental results show the superiority of the proposed method with respect to accuracy and robustness in the presence of severe and long-term occlusion and image noise in comparison with commonly used robust regressors.

1. Introduction

Visual tracking is a fundamental and essential part of many computer vision, robotic, and video analytic applications including Automatic visual surveillance [10], Behavior analysis [23], Motion capture and animation [20], Vehicle navigation and tracking [1], Traffic monitoring [3], Intelligent preventive safety systems [9], and Industrial robotics. In its simplest form, visual tracking is defined as the problem of locating three-dimensional (3D) target objects (such as a human or car) in a two-dimensional (2D) image plane as they move around a scene [24]. Besides other main parts such as target representation model and localization algorithm, the efficiency and reliability of a tracker are also highly affected by the used similarity measure method. The main goal of a similarity measure is to estimate the distance from the target representation model and the received data or image. Usually a predefined metric such as Euclidean distance is employed to measure the distance. However, these static metrics cannot accurately and robustly estimate the similarity level over time under challenging situations such as long-term occlusion and significant appearance changes.

A primary similarity measure used for the template matching problem is the *Euclidean distance* between the object template and the candidate sub-image. Assume that T is the object template, I is the received image frame, and $W(X; P)$ is the warping function which maps every pixel $X = \{x, y\}$ in the image plane to a pixel $X' = W(X; P)$ in the template based on the transformation parameters $P = \{p_1, \ldots p_k\}$. At every tracking time instant t, the goal of a template-based tracker is to find the best transformation parameter P^t in a way that the distance between the template T^t and the candidate sub-image I^t is minimized. [16] used the *sum of square difference* (SSD) to measure this distance:

$$P^t = \arg\min_P \sum_X \left[T^t(X) - I^t(W(X;P))\right]^2. \quad (1)$$

As illustrated in Eq. (1), the SSD measure can be used in conjunction with a gradient based optimization to estimate the transformation parameter. A least square algorithm is proposed in Ref. [16] to optimize Eq. (1). In general, L2-norm of errors is not robust against outliers, severe appearance variations, illumination changes, and occlusion. As a remedy, a *robust error function*, $\rho(e)$ is used to estimate the error e between the template and the candidate sub-image. Using a robust estimator instead of L2-norm, we obtain:

$$P^t = \arg\min_P \sum_X \rho\left(T^t(X) - I^t(W(X;P))\right). \quad (2)$$

Any function which satisfies the following criteria can be considered as a robust estimator [18]:

1. $\forall e \in \Re \rightarrow \rho(e) > 0$
2. $e_1 > e_2 > 0 \rightarrow \rho(e_1) > \rho(e_2)$
3. $e_1 < e_2 < 0 \rightarrow \rho(e_1) < \rho(e_2)$
4. $\rho(e)$ is piece-wise differentiable.

A wide variety of robust error functions have been used in the literature. The *Geman-McClure* function is commonly used for the task of visual tracking [2,21].

$$\rho(e) = \frac{e^2}{e^2 + \sigma^2} \quad (3)$$

☆ This paper has been recommended for acceptance by Tele Tan.

E-mail addresses: hadi.firouzi@ubc.ca (H. Firouzi), h.najjaran@ubc.ca (H. Najjaran).

0262-8856/$ – see front matter © 2014 Elsevier B.V. All rights reserved.
http://dx.doi.org/10.1016/j.imavis.2014.01.007

Another robust estimator used for tracking [8] is the *Huber* function.

$$\rho(e) = \begin{cases} \frac{1}{2}e^2 & if\ |e| \le \sigma \\ \sigma|e| - \frac{1}{2}\sigma^2 & otherwise \end{cases} \quad (4)$$

where in Eqs. (3) and (4), σ is a scale parameter.

It has been shown that these functions can improve the robustness of a visual tracker against outliers and occlusion [2]. In general, a robust estimator assigns a weight to each error value based on the magnitude of the error. The weight is less when the error is large. Despite the theoretical benefits, there are two practical problems which may significantly damage the efficiency and robustness of these functions. First the robust estimator is application dependent and has to be picked by a designer for different cases. This can be an acceptable limitation for some application, but it is not feasible under general conditions. Also, depending on the distribution of the error a proper scale vector (σ) has to be selected. Moreover, robust regression methods cannot distinguish between outliers and actual significant target appearance changes.

Besides the sum of square differences and robust estimators, other metrics such as *cross cumulative residual entropy* (CCRE) [22], *mutual information* (MI) [6], the *Bhattacharyya coefficient* [5], a convolution of spatial and feature space kernel functions [7], and *sum of conditional variance* (SCV) [19] have been proposed to measure the similarity of the target model and the received images. However, these methods are developed based on static and predefined measures which cannot sufficiently deal with challenging situations in a visual tracking scenario. One challenge is that the most similar candidate sub-image to the target model may not be the best match using a predefined similarity measure. The mentioned problem mainly rises when the target appearance changes over time or it is partially occluded by either itself or other background objects. Another phenomenon which can cause a tracker to fail is the existence of similar background objects known as distracters in a close proximity to the target object. Several works have been introduced to improve the accuracy of trackers in such situations. For instance, Li et al. [15] presented a pyramid-base scale adaptation method for mean-shift tracking. This tracker generates similarity functions at different scales and uses a coarse-to-fine search to avoid trapping in local minimum. Also, Karavasilis et al. [14] used the Gaussian Mixture Model (GMM) as the target representation model and the Differential Earth Mover's Distance (DEMD) as similarity measure for the task of tracking. This method combines DEMD-based tracker and Kalman filer algorithm to handle occlusion. Nevertheless, still the applicability of these predefined similarity measures is limited to specific cases.

Adaptive similarity measures, on the other hand, can be used to find the best match of the target model over time robustly. Collins et al. [4] proposed a dynamic feature selection method for estimating the similarity of the target model and the candidate image. In this method, the total number of features is fixed and the goal is to adaptively rank these features and use a subset of high ranked ones for matching. Although the method proposed in Ref. [4] can select discriminative features properly in some cases, the color features used in this method are not suitable in various applications, and also it is not always feasible to employ a more discriminative feature vector instead of color features due to the used exhaustive search for ranking the features. Recently Jiang et al. [13] proposed a classifier which is learned on-line from the tracking information to find the best match of the target model over time. In this method, an adaptive *Mahalanobis distance* is used to weight each feature in the classification process. According to the experimental results, this adaptive metric performed well in the existence of distracters. However, this method may fail in case of occlusion because of several reasons. First, this method uses proximity based approach to generate positive and negative samples at every time instant. However, in case of occlusion (specifically long term occlusion which has been emphasized in our work) image regions in very close vicinity of the target may not be true positive samples. Therefore, learning from false positive samples may cause the tracker to drift from the target. In addition, there is no specific mechanism in this method for handling occlusion and outliers. Although the method proposed in Ref. [13] is adaptive against target appearance and illumination changes, there is not enough evidence from the experimental results to verify its robustness and accuracy in case of occlusion.

Our proposed adaptive similarity measure differs from the works in the literature in several ways. First, unlike metrics presented in [4] where a subset of the feature vector is adaptively selected for matching, in our method the distance between the target and the image is modeled on-line by an adaptive hybrid model. Also, our method is more robust against severe and long-term occlusion than other relevant methods such as in [13]. Thus, our proposed adaptive metric is designed to reject outliers whereas it deals with appearance changes. Finally, our method requires less predefined parameters in comparison with other methods such as robust regression estimation [18] where a scale vector plays a crucial role in the robustness of the regressors.

In Section 2, first the proposed similarity measure is defined, and then an on-line algorithm to train a histogram-based classifier is described in detail. Next in Section 3, the proposed adaptive metric is used in a template matching problem. The results obtained by our metric is compared with several robust regressors as well as manually labeled ground truth data in Section 4. Lastly, in Section 5 some conclusions and potential future works are discussed.

2. Formulation

From the definition, the goal of a similarity measure is to estimate the distance from a target model and an image. In the proposed adaptive similarity measure, the Euclidean distance of the target model and the image is considered as the matching error. However, unlike a typical SSD method, a histogram-based classifier is learned on-line using the matching error history. Later, this classifier is used to assign a robust weight to each matching error based on its type.

Let $A = \{a_1, \ldots, a_m\}$ and $B = \{b_1, \ldots, b_n\}$ be the features describing the target model and the image, respectively.[1] Assuming that the feature space is metric, the number of features of the target and the image are the same (i.e.,$m = n$), and features have injective relation (i.e., $a_j = b_k \Rightarrow j = k$), we can find the Euclidean error $E = \{e_1, \ldots, e_n\}$ in the feature space as:

$$e_j = a_j - b_j. \quad (5)$$

Inspiring from the work proposed in Ref. [12], we categorize the matching error E based on their history into three classes:

E_i image noise and/or illumination variations,
E_a target appearance changes, and
E_o outliers and occlusion.

The first source of error, E_i, is mainly caused by either small illumination variations or some image noise which is inevitable in image capturing and computer vision. Usually the distribution of this type of error can be modeled by a zero-mean Gaussian function as $E_i \sim N(0, \sigma_i)$. In this work, instead of a Gaussian function a symmetrical range is learned from the previous matching errors. Other source of errors (i.e., E_a and E_o), on the other hand, cannot be easily discriminated from each other. The actual appearance changes may cause significant matching errors which are usually considered as outliers or occlusion by the conventional robust estimators [18]. A proper similarity measure has to reject outliers while it is adapting to the errors because of actual changes in target appearance and pose. Since in a tracking scenario, the target

[1] In this work, the image pixel values are considered as features. However, the proposed method can be suitably integrated with a feature-based tracker.