Pattern Recognition 45 (2012) 4494–4509

Contents lists available at SciVerse ScienceDirect

Pattern Recognition

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

Robust decentralized multi-model adaptive template tracking

Hadi Firouzi *, Homayoun Najjaran

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

ARTICLE INFO

Article history:
Received 24 September 2011
Received in revised form
10 March 2012
Accepted 3 May 2012
Available online 19 May 2012

Keywords:
Non-rigid object
Robust fusion
EM algorithm
Mixture of Gaussian
Decentralized object localization

ABSTRACT

In this paper, a robust and efficient visual tracking method through the fusion of several distributed adaptive templates is proposed. It is assumed that the target object is initially localized either manually or by an object detector at the first frame. The object region is then partitioned into several non-overlapping subregions. The new location of each subregion is found by an EM[1]-like gradient-based optimization algorithm. The proposed localization algorithm is capable of simultaneously optimizing several possible solutions in a probabilistic framework. Each possible solution is an initializing point for the optimization algorithm which improves the accuracy and reliability of the proposed gradient-based localization method to the local extrema. Moreover, each subregion is defined by two adaptive templates named immediate and delayed templates to solve the "drift" problem.[2] The immediate template is updated by short-term appearance changes whereas the delayed template models the long-term appearance variations. Therefore, the combination of short-term and long-term appearance modeling can solve the template tracking drift problem. At each tracking step, the new location of an object is estimated by fusing the tracking result of each subregion. This fusion method is based on the local and global properties of the object motion to increase the robustness of the proposed tracking method against outliers, shape variations, and scale changes. The accuracy and robustness of the proposed tracking method is verified by several experimental results. The results also show the superior efficiency of the proposed method by comparing it to several state-of-the-art trackers as well as the manually labeled "ground truth" data.

1. Introduction

The ability of accurately and efficiently tracking real-world (especially non-rigid) objects in dynamic and cluttered environments plays an essential role for most computer vision and video analytic applications such as automated visual surveillance systems [29,33,14,25,37], face and activity recognition [19,71,64], motion capture and animation [59,21,2], video games (e.g., EyeToy [54]), vehicle navigation and tracking [28,5], traffic monitoring [13], intelligent preventive safety systems [53,36,27], human computer interaction [9], industrial robotics [49], and medical diagnosis [3,68,67]. Although visual tracking has been studied for many years [10,40], it is still challenging due to the inevitable object appearance variations, scale changes, occlusion, illumination changes, image noise, unpredictable and complex motion, and cluttered and dynamic background. For instance, a target moving far from the camera can be occluded partially or fully for a short-term by some closer objects. The location and shape of targets may significantly change during the tracking task. Specifically, the main difficulty in tracking non-rigid objects is related to the high dimensional complexity and uncertainty in the real applications [65]. As a result, developing an efficient and robust non-rigid object tracking capable of attacking the mentioned problems is necessary to fulfill the demands for the current and future real-world machine vision applications.

In this paper a robust and distributed object tracking method based on a multi-initializing points EM-like optimization algorithm and multiple heterogeneous adaptive templates is proposed. The method is capable of tracking non-rigid objects with variable appearance, shape, scale, and unpredicted motion in cluttered environments. It is assumed that the target object has been detected manually or automatically (by any existing object detection method) at the first frame and the goal is to adaptively track the target object without any prior knowledge about the object appearance and motion. Also the target object is the whole or part(s) of a real-world object (e.g., a human, face, or car) and cannot have a chaotic or huge movement between two consecutive images. However, the camera may not be stationary and different parts of the target object can move independently in any direction whereas the whole object is approaching to a specific location.

* Corresponding author. Tel.: +1 250 5750281.
E-mail addresses: hadi.firouzi@ubc.ca (H. Firouzi), h.najjaran@ubc.ca (H. Najjaran).

[1] Expectation Maximization, see [45] for more information.

[2] The problem of gradually updating the object appearance model with irrelevant information such as background pixel values [43].

0031-3203/$ - see front matter © 2012 Elsevier Ltd. All rights reserved.
http://dx.doi.org/10.1016/j.patcog.2012.05.005

2. Visual tracking

Roughly speaking, a typical visual target tracking method consists of two main components:

1. *Representation model*, which describes the target appearance and can be used to evaluate the likelihood of the target object being at a particular location in 2-D images.
2. *Localization method*, by which the most likely location of the object in the current frame can be found. The target location is estimated either by a search strategy or filtering algorithm. The difference between them is that, in the former the object is located by searching for the similar image region to the representation model within a close neighbor around the previous location. However, in the latter, first several possible object locations are estimated based on the object motion model and then the best fit to the representation model is chosen as the object location in the current image.

In addition to the above tracking components, a typical visual tracking task can be involved in two different processes: (1) called Target Representation and Localization (TRL), a bottom-up process, which has to estimate the target location in the current image considering the target appearance changes (2) Filtering and Data Association (FDA), a top-down process, which evaluates different hypotheses based on the target dynamics to find the most likely location of the target in the current image. Depending on the application, these two processes can be combined with different importance factors. For example in the case of face tracking [19] in a crowded environment, the tracking method is more based on TRL than FDA, because modeling the target appearance is more reliable than predicting the target dynamics. On the other hand, for the applications such as aerial video surveillance where the target motion is more important and can be estimated [37], the FDA process is mostly used for target tracking.

In general, the FDA is a process to solve the discrete state space problem [6]. The state space dynamic equation can be defined by a nonlinear, time-varying, vector-valued function f_k such that $x_k = f_k(x_{k-1}, v_k)$ where $\{x_k\}_{k=0,1,..}$ and $\{v_k\}_{k=1,2,..}$ are the state and noise sequences over time. The measurement equation is also defined by function h_k with the same properties as that of f_k: $z_k = h_k(z_{k-1}, n_k)$ where $\{z_k\}_{k=1,..}$ and $\{n_k\}_{k=1,2,..}$ are the observation and noise sequences over time. It is noted that both z_k and n_k are independent and identically distributed (i.i.d.). Given all observations $\{z_k\}_{k=1,..}$ up to the current time instant, the main goal is to estimate the state vector x_k. In theory the optimal solution can be found by the Bayesian Filtering (BF) that is a recursive two-step (i.e., prediction and update steps) process to approximate the probability density function (pdf) $p(x_k|z_{1:k})$. In the prediction step, given the previous state pdf $p(x_{k-1}|z_{1:k-1})$ at time instant $k-1$, the prior state pdf $p(x_k|z_{1:k-1})$ at time instant k is predicted based on the state dynamic equation and the state transition function $p(x_k|x_{1:k-1})$. Accordingly in the second step, the likelihood function $p(z_k|x_k)$ of the current observation z_k is used to update the posterior state pdf $p(x_k|z_{1:k})$ at time instant k.

Based on the functions f_k,h_k and the noise model, different filtering methods have been proposed. In its simplest form, when the dynamic and measurement equations are linear and the noise is white noise (i.e., zero mean Gaussian) the optimal solution can be found by the Kalman Filter (KF) [6]. In KF the posterior pdf is a Gaussian distribution. If the functions f_k,h_k are nonlinear but the noise is Gaussian, the Extended Kalman Filter (EKM) [6] and Unscented Kalman Filter (UKF) [32], a more recent alternative, can be used to estimate the posterior pdf which is still modeled as Gaussian. In contrast to EKM, UKF uses a parametric model to estimate the mean and covariance of the posterior pdf using a set of discretely sampled points. In the case that the state space is composed of a discrete and finite set of states, the tracking problem can be solved by the Hidden Markov Model (HMM) [52]. In the most general case, the functions f_k,h_k can be nonlinear and there is no assumption on probability density functions, the problem can be solved based on a sequential Monte Carlo method such as the Particle Filter (PF) [35], also called Bootstrap Filter [24]. In PF, the prior pdf is modeled by a set of random samples with different importance weights and the posterior pdf is approximated based on these samples and associated weights (refer to [4,20] for reviews).

In the case of cluttered scenes where more than one target can be present [41], the problem of multi-target tracking will arise. This problem can be viewed as the validation and association of the observations to the targets [6]. In the validation step, noise and outliers are removed by only considering the observations whose predicted probability of appearance is high. Then a data association algorithm is needed to relate the validated observations to the existing targets. In addition to some intuitive strategies such as the Nearest Neighbor Filter which chose the closest observation to the targets, the Probabilistic Data Association Filter (PDAF) can be used to track single target when multiple observations are available. PDAF assumes that only one observation can be related to a specific target and other observations are because of noise or outliers and modeled by a uniform distribution. For the case of multiple targets data association, different strategies such as Joint Data Association Filter (JDAF) [6] and Multiple Hypothesis Filter (MHF) [6,18,58] can be used. JDAF calculates the joint probability for all observation-to-target associations, whereas MHF estimates the probability of a certain observation sequence produced by a specific target. MHF can also be used to track a multi-modal density function [11]. Pinho and Tavares [50] used KF algorithm for corresponding multiple object features by minimizing a global cost function which is defined based on Mahalanobis distance between features in sequential images. Also several data association methods have been proposed for particle filter-based multi-target tracking [56,30].

In this work, we will focus more on the second process i.e., Target Representation and Localization (TRL) rather than Filtering and Data Association (FDA) process. In fact, for tracking of real-world and specially non-rigid objects, target localization based on shape and appearance adaptation can be more reliable and informative than the target dynamic modeling and motion estimation. In the case of tracking with no information about motion dynamics (although target motion dynamics can be estimated during the time, it is not reliable due to the unpredicted and complex target and camera motion), localization based on the target appearance and shape plays a crucial role in developing a robust tracking method. Despite the FDA process which has its root in control theory, TRL process is specifically related to the visual tracking and image registration methods [69,48]. Both target tracking and image registration can be viewed as a likelihood maximization problem, however in tracking only small variations in target appearance and location are assumed over two sequential images, therefore here an efficient and accurate gradient-based optimization algorithm is proposed to localize the target object based on the object appearance and shape changes.

In the following sections, first in Section 3 several related methods are reviewed. The details of the proposed robust decentralize template-based tracking method are explained in Section 4 where the object representation model and decentralized localization are proposed. In Section 5 the proposed tracker has been applied on several challenging videos and the results have been compared with four state-of-the-art methods as well as the ground truth data. Some conclusions and potential extensions for future work are provided in Section 6.