Computer Vision and Image Understanding 113 (2009) 544–555

Contents lists available at ScienceDirect

Computer Vision and Image Understanding

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

Accurate appearance-based Bayesian tracking for maneuvering targets ☆

Emilio Maggio *, Andrea Cavallaro

Multimedia and Vision Group, Queen Mary University of London, Mile End Road, London E1 4NS, UK

ARTICLE INFO

Article history:
Received 1 May 2007
Accepted 9 December 2008
Available online 24 December 2008

Keywords:
Object representation
Color histogram
Tracking
Mean Shift
Particle Filter

ABSTRACT

We propose a tracking algorithm that combines the Mean Shift search in a Particle Filtering framework and a target representation that uses multiple semi-overlapping color histograms. The target representation introduces spatial information that accounts for rotation and anisotropic scaling without compromising the flexibility typical of color histograms. Moreover, the proposed tracker can generate a smaller number of samples than Particle Filter as it increases the particle efficiency by moving the samples toward close local maxima of the likelihood using Mean Shift. Experimental results show that the proposed representation improves the robustness to clutter and that, especially on highly maneuvering targets, the combined tracker outperforms Particle Filter and Mean Shift in terms of accuracy in estimating the target size and position while generating only 25% of the samples used by Particle Filter.

1. Introduction

Image-based tracking is an important component in many applications, such as video surveillance, medical image sequence analysis, augmented reality, smart rooms and object-based video compression. The goal of image-based tracking is to estimate the position and the shape of an object or a region over time. This requires the definition of a target model and of a process that first generates candidate targets and then evaluates the similarity between the model and a candidate.

A simple and widely used *target model* is the template [1], which stores luminance or color values, and their location. Although template computation is simple and fast, the values stored in the template may become non-representative of the object appearance in presence of noise, partial occlusions, pose or scale changes. Solutions have been proposed to update the template over time [2–4] and to cope with occlusions [5] and pose changes [6]. However the complete pixel information may be unnecessary for the tracking task: a target representation should be descriptive enough to disambiguate the object from the background, while allowing a certain degree of flexibility to cope with changes of target scale, pose, scene illumination and partial occlusions. To this extent, color histograms have been used as target models for their invariance to scaling and rotation, robustness to partial occlusions, data reduction and efficient computation [7–10]. However, the descriptiveness of color histograms is limited by the lack of spatial information, which makes it difficult to discriminate targets with similar color properties. To overcome this problem, the information of the first two spatial moments associated to the location of the related color can be added to each bin of the histogram [11]. Alternatively, multiple histograms on different parts of the target can be used [8,12,13], although there is no widely accepted solution. The multi-part representation in [12] divides a target into two non-overlapping areas (top and bottom parts). This solution is effective for the specific application (i.e., tracking ice-hockey players), as it generally corresponds to the shirt and the trousers, but it is not necessarily effective on a generic target. An alternative to improve the distinctiveness of the target model is the use of multiple features. For example, gradient information can be used to complement color information [9,14,15]. However, computing several features for each candidate target may be computationally expensive for real-time applications.

After the definition of a target model, a *search method* is needed to select the candidate target locations to be evaluated against the model. To this extent, Particle Filters (PF) have been widely used in image-based tracking [16,8,17,2]. PF is a probabilistic method based on Monte Carlo sampling that can deal with multi-modal probability density functions (*pdf*s). PF-based trackers use the multiple hypotheses associated with the samples (i.e., the particles) to cope with occlusions and to recover from lost tracks. As the number of particles required to model the underlying *pdf* increases exponentially with the dimensionality of the state space, efficient proposal distributions for particle sampling are desirable. A popular choice is to draw the samples according to the target dynamic model, thus resorting to an algorithm known as CONDENSATION [18,16] (here referred to as PF-C). However, sampling in PF-C does not account for information from the most recent measurement. As a consequence, when the dynamic model is not accurate, the area

☆ The authors acknowledge the support of the UK Engineering and Physical Sciences Research Council (EPSRC), under Grant EP/D033772/1.

* Corresponding author. Fax: +44 2078827997.
E-mail addresses: emilio.maggio@ieee.org (E. Maggio), andrea.cavallaro@elec.qmul.ac.uk (A. Cavallaro).

1077-3142/$ - see front matter © 2008 Elsevier Inc. All rights reserved.
doi:10.1016/j.cviu.2008.12.003

of the state space around the target is not densely sampled. To account for the latest measurement many sampling strategies have been proposed [19–23]. Markov Chain Monte Carlo (MCMC) samplers have been used to sample the particles in high-dimensional state spaces (e.g., 3D body tracking) [20,24,25]. However, due to the relatively large number of steps necessary for MCMC to converge, no improvements in terms of efficiency are reported on low-dimensional state spaces [20]. Simulated Annealing is an alternative approach for particle sampling [26]: first particles are randomly spread over the state space, then a layered procedure re-draws the samples proportionally to their likelihood. When the relationship between state and measurement can be linearized, an alternative is to sample from the Gaussian estimate computed by an Extended Kalman Filter associated to each particle [19]. Similarly, EKF can be substituted with an unscented transform that does not require linearization [19,27]. Both methods assume that the modes of the *pdf* are well represented by their first and second order moments. Furthermore, while PF-C requires explicit definition of the likelihood only, the last two methods also require the explicit formulation of the measurement equation.

A different approach to particle sampling is to drive the particles according to point estimates of the gradient of either the posterior or the likelihood [21,20,24,25,28,29]. When the appearance model is a template, optical flow can be used to drive the particles towards peaks of the likelihood. However, as motion blur can affect the accuracy of optical flow, the particle shifting procedure is enabled only when the momentum of the object is small [21]. A more principled solution, known as Kernel Particle Filter, uses kernel density estimation to produce from the particle set a continuous approximation of the posterior *pdf*. Then, sample-based Mean Shift (MS), a kernel-based iterative procedure, is used to approximate the gradient of the *pdf* and to climb its modes [22,23]. However, as the accuracy of density estimate and of its gradient depends on the sampling rate, a reduction of the number of samples may affect the quality of the final approximation. An alternative to sample-based MS is color-based MS [7], a very popular tracking algorithm that uses color histograms. Color-based MS performs a gradient descent of the model-candidate distance using the kernel-weighted color density (and not the sample density) estimate. Unlike PF, which requires the costly computation of one candidate model (e.g., a color histogram) per particle, color-based MS has a low computational cost. In fact the gradient estimate requires the computation of one histogram per iteration only. However, while sample-based approaches such as PF are flexible in terms of search region, the search of color-based MS is limited by the kernel size. For this reason, color-based MS fails to track small and fast moving targets as well as to recover the position of a target after a total occlusion. Given the complementarity of the two algorithms, combinations of MS and PF have been studied [28,29]. However, the convergence of the MS procedure with the used flat kernel is not demonstrated [28]. Moreover, as the particles are re-displaced by MS, it is not clear how the new sampling distribution can be approximated to correctly compute their weights [28,29].

In this paper we propose an algorithm that improves the robustness of the target representation and increases tracking flexibility and effectiveness. The target representation uses semi-overlapping color histograms that improve the sensitivity to rotations and anisotropic scale changes, while maintaining the robustness and flexibility typical of single color histograms. Related to this partition, an extension of MS is proposed, which is used to find local minima of the model-candidate distance in the state space. Particles generated by PF are shifted toward these minima, which represent positions with high probability of locating a target, thus increasing the efficiency of the particles. The increased efficiency is obtained on low-dimensional state spaces (3D to 5D), where other gradient-based sampling methods are ineffective [20].

The paper is organized as follows. Section 2 discusses the use of spatial information in target representation and introduces the color histogram representation. The proposed tracker is presented in Section 3. Section 4 describes the evaluation procedure used in Section 5 to assess the experimental results. Finally, in Section 6 we draw conclusions and discuss future research directions.

2. Target representation

Let us approximate the target shape with an ellipse and represent the target state with $\mathbf{x} = [\mathbf{y}, \mathbf{s}]$, where $\mathbf{y} = [x, y]$ is the center of the ellipse and $\mathbf{s} = [e, \theta, h_1]$. The variable e is the ellipse eccentricity, θ is its clockwise rotation and h_1 is the length of the semi-axis used as reference for the rotation. In the following we will also use h_2 as the length of the second semi-axis (Fig. 1).

Let the target representation of the pixels inside the ellipse be their weighted color distribution approximated by a normalized color histogram. Given an image $\mathbf{z}$ the normalized color histogram $\mathbf{r}(\mathbf{x}, \mathbf{z}) = \{r_u(\mathbf{x}, \mathbf{z})\}_{u=1,\dots,U}$ with U bins of a target candidate $\mathbf{x}$ can be calculated by selecting for each pixel in the ellipse the bin index u corresponding to its color and then cumulating on the bin the values obtained with a weighting kernel $k(.)$. The kernel $k(.)$ usually gives higher weight to pixels near the center of the ellipse as they are less likely to be occluded by other objects [7]. Given the coordinates of the $n(\mathbf{x})$ pixels inside the ellipse $\{\mathbf{w}_i\}_{i=1,\dots,n(\mathbf{x})}$, the Dirac's delta function $\delta(.)$, a function $b(\mathbf{w}_i, \mathbf{z})$ that associates a pixel of the image $\mathbf{z}$ with position $\mathbf{w}_i$ to the histogram bin, and

$$A(\mathbf{s}) = \begin{bmatrix} \frac{\cos\theta}{h_2} & -\frac{\sin\theta}{h_2} \\ \frac{\sin\theta}{h_1} & \frac{\cos\theta}{h_1} \end{bmatrix},$$

the matrix used to scale and rotate the kernel, the computation of the bin value $r_u(\mathbf{x}, \mathbf{z})$ can be formalized as

$$r_u(\mathbf{x}, \mathbf{z}) = C(\mathbf{x}) \sum_{i=1}^{n(\mathbf{x})} k(\|A(\mathbf{s})(\mathbf{y} - \mathbf{w}_i)\|^2)\delta[b(\mathbf{w}_i, \mathbf{z}) - u], \quad u = 1, \dots, U, \tag{1}$$

where $C(\mathbf{x})$ is a normalization function defined as

$$C(\mathbf{x}) = \frac{1}{\sum_{i=1}^{n(\mathbf{x})} k(\|A(\mathbf{s})(\mathbf{y} - \mathbf{w}_i)|^2)}. \tag{2}$$

Then, we define the target model as the color distribution of the object at track initialization, i.e., $\mathbf{o} = \mathbf{r}(\mathbf{x}_I, \mathbf{z}_I)$, where $\mathbf{x}_I$ and $\mathbf{z}_I$ are the state and the image frame at initialization.

The matching quality of a candidate is defined by the candidate-model distance, d, between the normalized histograms $\mathbf{r}(\mathbf{x}, \mathbf{z})$ and $\mathbf{o}$:

$$d[\mathbf{r}(\mathbf{x}, \mathbf{z}), \mathbf{o}] = \sqrt{1 - \rho[\mathbf{r}(\mathbf{x}, \mathbf{z}), \mathbf{o}]}, \tag{3}$$

where ρ is the Bhattacharyya coefficient [30]

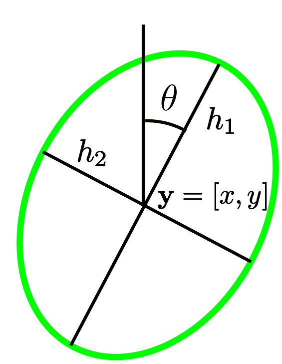

Fig. 1. Parameters defining the ellipse bounding the target area.