Image and Vision Computing 29 (2011) 295–305

Contents lists available at ScienceDirect

Image and Vision Computing

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

Visual tracking using the Earth Mover's Distance between Gaussian mixtures and Kalman filtering

Vasileios Karavasilis, Christophoros Nikou *, Aristidis Likas

Department of Computer Science, University of Ioannina, PO Box 1186, 45110 Ioannina, Greece

ARTICLE INFO

Article history:
Received 1 February 2010
Received in revised form 29 November 2010
Accepted 10 December 2010

Keywords:
Visual tracking
Gaussian mixture model (GMM)
Expectation-Maximization (EM) algorithm
Differential Earth Mover's Distance (Differential EMD)
Kalman filter

ABSTRACT

In this paper, we demonstrate how the differential Earth Mover's Distance (EMD) may be used for visual tracking in synergy with Gaussian mixtures models (GMM). According to our model, motion between adjacent frames results in variations of the mixing proportions of the Gaussian components representing the object to be tracked. These variations are computed in closed form by minimizing the differential EMD between Gaussian mixtures, yielding a very fast algorithm with high accuracy, without recurring to the EM algorithm in each frame. Moreover, we also propose a framework to handle occlusions, where the prediction for the object's location is forwarded to an adaptive Kalman filter whose parameters are estimated on line by the motion model already observed. Experimental results show significant improvement in tracking performance in the presence of occlusion.

1. Introduction

One important field in computer vision is tracking. Tracking is the procedure of generating an inference about motion given a sequence of images. Solutions to this problem have a variety of applications, some of them being: surveillance, targeting, recognition from motion, motion-based video compression, teleconferencing, video indexing and traffic monitoring. In tracking problems, it is assumed that the model of the object is known, that is how the object looks like, or its appearance. Based on a set of measurements in image frames the object's position should be estimated. In that context, the Differential Earth Mover's Distance (DEMD) tracking algorithm [40,41] was recently presented. In this method, the object is represented by a histogram (called a signature) and the distance between signatures in consecutive frames to be minimized is the Earth Mover's Distance [25]. The computational complexity of the EMD prevents a direct implementation in many real time applications. To overcome this drawback, the DEMD algorithm based on sensitivity analysis of the simplex method provides an acceleration compared with its standard counterpart [40,41].

Motivated by the efficiency of the differential EMD tracking algorithm [40,41] and the compactness of the representation of probability densities using Gaussian mixture models [5], we propose in this paper to first model the appearance of the target by a Gaussian mixture model trained on a weighted likelihood and then to employ the differential EMD approach for tracking. According to our model, motion between adjacent frames results in variations of the mixing proportions of the Gaussian components representing the object. These variations affect the distance between the mixtures, at the same image location, representing the object in consecutive frames. By these means, the gradient of the EMD, namely the differential EMD [40,41], between Gaussian mixtures shows the direction of the minimum and consequently the target location.

Moreover, in a second phase of this work, we propose to consider the estimated location of the target as a measurement (observation) of a time-varying Kalman filter in order to address cases presenting occlusions. Hence, the prediction for the object's location is forwarded to a Kalman filter whose state matrix parameters are not constant but they are updated on-line based on recent history of the estimated motion.

The contribution of the presented work is twofold. At first, the proposed approach leads to a significant improvement in terms of execution time with respect to the differential EMD tracking algorithm [40,41] without compromising the accuracy of the method. At second, based on the motion model already observed, occlusions are successfully handled by modifying on-line the state matrix of a Kalman filter.

The remainder of the paper is organized as follows: A review of tracking algorithms is presented in Section 2. In Section 3, the modeling of the object to be tracked by a Gaussian mixture is presented. The tracking algorithm relying on the minimization of the Earth Mover's Distance between Gaussian mixtures is presented in Section 4. In Section 5, the extension of the algorithm in order to

* Corresponding author. Tel.: +30 265 100 8802.
E-mail addresses: vkaravas@cs.uoi.gr (V. Karavasilis), cnikou@cs.uoi.gr (C. Nikou), arly@cs.uoi.gr (A. Likas).

0262-8856/$ – see front matter © 2010 Elsevier B.V. All rights reserved.
doi:10.1016/j.imavis.2010.12.002

address the problem of occlusion is described. Experimental results are shown in Section 6 which is followed by our conclusion in Section 7.

2. Related work

Tracking algorithms may be classified in two categories [9]. The first category is based on filtering and data association, while the second family of methods relies on target representation and localization. The algorithms based on filtering assume that the moving object has an internal state which may be measured and, by combining the measurements with the model of state evolution, the object's position is estimated. The first method of that category is the Kalman filter [28] which successfully tracks objects even in the case of occlusion if the assumed type of motion is correctly modeled [11]. Another approach in this category is the particle filters [2,37]. This category also includes Condensation [16] and ICondensation [17] algorithms which are more general than Kalman filters, as they do not assume specific type of densities and, using factored sampling, have the ability to predict an object's location under occlusion as well. Also, in this category, methods based on feature extraction and tracking were also proposed [32]. The object is represented by a set of scale invariant landmarks [18] which are tracked using optical flow [19,27]. These methods have the drawback that the type of object's movement should be correctly modeled.

On the other hand, tracking algorithms relying on target representation and localization employ a probabilistic model of the object appearance and try to detect this model in consecutive frames of the image sequence. More specifically, color or texture features of the object, masked by an isotropic kernel, are used to create a histogram. Then, the object's position is estimated by minimizing a cost function between the model's histogram and candidate histograms in the next image. A representative method in this category is the mean shift algorithm [9] where the object is supposed to be inside an ellipse and the histogram is constructed from pixel values inside that ellipse. Extensions of the main algorithm are proposed in [34] where the mean shift is combined with particle filters, in [42] where scale invariant features are used and in [36] where various distance measures are associated with the mean shift algorithm. Other approaches using multiple kernels [12] and a Newton style optimization procedure [14] were also proposed. The DEMD tracking algorithm [40,41] also belongs to the category of methods relying on target representation and localization.

The above methods track only one object at a time. Other works track many objects simultaneously [1,4,7,30] and in these cases occlusions may be detected more efficiently. Moreover, the object to be tracked is usually represented by its color histogram, but this is not always necessary. A Gaussian mixture model (GMM) was used in [33] to represent the object in a joint spatial-color space and in [29] for background subtraction and the contour of the object was tracked in [26,39]. Moreover, it is well-known that a level set representation also addresses the problem of multiple objects [10,20,22–24]. In any case, combining multiple object representations could make the tracking procedure more robust [15,21,31,35]. Finally, in [6], multiple views of an object are learnt through Principal Component Analysis (PCA) and a Support Vector Machine (SVM) classifier was also used in [3]. A review of tracking methods can be found in [38].

3. Target appearance modeling

In this section we present the basic idea of minimizing the Earth Mover's Distance between Gaussian mixture models for tracking. We describe the GMM as a way of representing an object's appearance and define the EMD as a distance between two GMM.

3.1. Background on weighted Gaussian mixture models

A one dimensional Gaussian distribution has a probability density function given by

$$\mathcal{N}(I_n|\mu,\sigma) = \frac{1}{\sqrt{2\pi}\sigma}\exp\left(-\frac{(I_n-\mu)^2}{2\sigma^2}\right) \tag{1}$$

where I_n is the intensity of the n^{th} pixel, μ is the mean value and σ^2 is the variance of the distribution.

Let two Gaussian distributions be:

$$f_1(I_n) = \mathcal{N}(I_n|\mu_1,\sigma_1), \qquad f_2(I_n) = \mathcal{N}(I_n|\mu_2,\sigma_2). \tag{2}$$

The Gaussian mixture model (GMM) is a convex combination of Gaussian components [5]. A single component is given by Eq. (1) and the GMM with m components is expressed by

$$f(I_n|\mu,\sigma,\pi) = \sum_{i=1}^{m}\pi_i\mathcal{N}(I_n|\mu_i,\sigma_i) \tag{3}$$

where $\mu=\{\mu_i\}_{i=1,\dots,m}$, $\sigma=\{\sigma_i\}_{i=1,\dots,m}$ and $\pi=\{\pi_i\}_{i=1,\dots,m}$, are the model parameters. The parameters π_i represent the importance of each component and satisfy the constraints $\sum_{i=1}^{m}\pi_i=1$ and $\pi_i\geq 0$, $\forall i=1,\dots,m$.

We assume that we have grayscale images, and each object may be described by the intensities of its pixels. An object is represented by an ellipsoidal region, and the object's pixels are those lying inside that region. The usual way to represent an object is by histograms of m_h bins. This approach has the disadvantage that the number of the bins must be specified *a priori*. However, it is a very common and efficient way of modeling the object to be tracked in the majority of the state of the art trackers [9].

In this work we propose the representation of an object using a GMM. The parameters of the GMM are estimated by clustering the density values of the object's pixels using the EM algorithm [5]. An advantage of the GMM representation is that the number of components m is significantly smaller than the number of distinct intensities.

Every object may be represented by an ellipsoidal region with finite precision. As an effect, inside the ellipse, there will be regions not belonging to the object. Usually, these regions exist at the edges of the ellipse. To eliminate the influence of regions not belonging to the object, the ellipse is weighed by a kernel as will be explained below.

At first, we assume that the center of the ellipse is in the spatial location $(0,0)$. Then, the ellipse is normalized to a unit circle by dividing each pixel coordinates by h_x and h_y, which are the sizes of the ellipse in the horizontal and vertical directions respectively. Let the normalized pixel locations be (x_n,y_n). An isotropic kernel, with profile $k(x)$, is applied to pixels inside the unit circle to attribute corresponding weights at every pixel. The weight for a pixel indexed by n is defined by

$$w_n = \frac{k\left(x_n^2+y_n^2\right)}{\sum_{i=1}^{N}k(x_i^2+y_i^2)}. \tag{4}$$

Notice that $x_n^2+y_n^2\leq 1$ because the point (x_n,y_n) is inside the unit sphere and $\sum_{n=1}^{N}w_n=1$. The kernel profile $k(x)$ is a convex monotonic decreasing function such that $k:[0,\infty)\rightarrow\Re$ and g is the negative derivative of the kernel function, $g(x)=-k'(x)$. We use a kernel with Epanechnikov profile [9]

$$k(x) = \begin{cases}\frac{1}{2}(1-x) & \text{if} \quad x\leq 1\\ 0 & \text{otherwise}\end{cases} \tag{5}$$