Pattern Recognition Letters 49 (2014) 250–258

Contents lists available at ScienceDirect

Pattern Recognition Letters

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

Robust scale-adaptive mean-shift for tracking ☆

Tomas Vojir [a,*], Jana Noskova [b], Jiri Matas [a]

[a] The Center for Machine Perception, FEE CTU in Prague, Karlovo Namesti 13, 121 35 Prague 2, Czech Republic
[b] Faculty of Civil Engineering, CTU in Prague, Thakurova 7/2077, 166 29 Prague 6, Czech Republic

ARTICLE INFO

Article history:
Available online 13 April 2014

Keywords:
Object tracking
Mean-shift
Scale estimation
Background weighting

ABSTRACT

The mean-shift procedure is a popular object tracking algorithm since it is fast, easy to implement and performs well in a range of conditions. We address the problem of scale adaptation and present a novel theoretically justified scale estimation mechanism which relies solely on the mean-shift procedure for the Hellinger distance. We also propose two improvements of the mean-shift tracker that make the scale estimation more robust in the presence of background clutter. The first one is a novel histogram color weighting that exploits the object neighborhood to help discriminate the target called background ratio weighting (BRW). We show that the BRW improves performance of MS-like tracking methods in general. The second improvement boost the performance of the tracker with the proposed scale estimation by the introduction of a forward–backward consistency check and by adopting regularization terms that counter two major problems: scale expansion caused by background clutter and scale implosion on self-similar objects. The proposed mean-shift tracker with scale selection and BRW is compared with recent state-of-the-art algorithms on a dataset of 77 public sequences. It outperforms the reference algorithms in average recall, processing speed and it achieves the best score for 30% of the sequences – the highest percentage among the reference algorithms.

1. Introduction

The mean-shift (MS) algorithm by Fukunaga and Hostetler [4] is a non-parametric mode-seeking method for density functions. It was introduced to computer vision by Comaniciu et al. [3] who proposed its use for object tracking. The MS algorithm tracks by minimizing the distance between two probability density functions (pdfs) represented by a target and target candidate histograms. Since the histogram distance (or, equivalently, similarity) does not depend on the spatial structure within the search window, the method is suitable for deformable and articulated objects.

The performance of the mean-shift algorithm suffers from the use of a fixed size window if the scale of the target changes. When the projection of the tracked object becomes larger, localization becomes poor since some pixels on the object are not included in the search window and the similarity function often has many local maxima. If the object become smaller, the kernel window includes background clutter which often leads to tracking failure.

The seminal paper by Comaniciu et al. [3] already considered the problem and proposed changing the window size over multiple runs by a constant factor ($\pm 10\%$). The window size maximizing the similarity to the target histogram was chosen. This approach does not cope well with the increase of the object size since the smaller windows usually have higher similarity and therefore the scale is often underestimated.

Collins [2] exploited image pyramids and used an additional mean-shift procedure for scale selection after estimating the position. The method works well for objects with a fixed aspect ratio, but this often does not hold for non-rigid or a deformable objects. Moreover, the method is significantly slower than the standard MS.

Image moments are used in [1,10] to determine the scale and orientation of the target. The second moments are computed from an image of weights that are proportional to the probability that a pixel belongs to the target model. Yang et al. [13] introduced a new similarity measure that estimates the scale by comparison of second moments of the target model and the target candidate.

Pu and Peng [11] assume target rigidity and restrict motion to scaling and translation. The target is first tracked using the mean-shift both in the forward and backward direction to estimate the translation. Scale is then estimated from feature points matched by an M-estimator with outlier rejection. Similarly, [8,15] rely on ”support features” for scale estimation after the mean-shift algorithm solves for position. Liang et al. [8] search for the target boundary by correlating the image with four

☆ This paper has been recommended for acceptance by J.K. Kämäräinen.
* Corresponding author. Tel.: +420 224 355 767; fax: +420 224 357 385.
E-mail address: vojirtom@cmp.felk.cvut.cz (T. Vojir).

http://dx.doi.org/10.1016/j.patrec.2014.03.025
0167-8655/© 2014 Elsevier B.V. All rights reserved.

templates. Positions of the boundaries directly determine the scale of the target. Zhao et al. [15] exploit affine structure to recover the target relative scale from feature point correspondences between consecutive frames.

Methods depending on feature matching are able to robustly estimate the scale, but they cannot be seamlessly integrated to the mean-shift framework. Moreover, estimating scale from feature correspondences takes times, requires presence of well-localised features that can be detected with high repeatability, and it has difficulties dealing with a non-rigid or a deformable object.

We present a theoretically justified scale estimation mechanism which, unlike the method listed above, relies solely on the mean-shift procedure for the Hellinger distance. Furthermore, we propose a formulation for background weighting that exploits the tracked object's neighborhood to help discriminate the object from the background. Additionally, we present two mechanisms that make the scale estimation more robust in the presence of background clutter and improve tracker performance to level of the state-of-the-art. The performance is compared to state-of-the-art algorithms on a large tracking dataset.

2. Mean-shift tracker with scale estimation

2.1. Standard kernel-based object tracking

In the standard mean-shift tracking of [3], the target is modelled as an m-bin kernel-estimated histogram in a feature space located at the origin:

$$\hat{\mathbf{q}} = \{\hat{q}_u\}_{u=1\dots m} \quad \sum_{u=1}^{m} \hat{q}_u = 1. \tag{1}$$

A target candidate at location $\mathbf{y}$ in the subsequent frame is described by its histogram

$$\hat{\mathbf{p}}(\mathbf{y}) = \{\hat{p}_u(\mathbf{y})\}_{u=1\dots m} \quad \sum_{u=1}^{m} \hat{p}_u = 1; \tag{2}$$

Let $\mathbf{x}_i$ denote pixel locations, n be the number of pixels of the target and let $\{\mathbf{x}_i^*\}_{i=1\dots n}$ be the pixel locations of the target centered at the origin. Spatially, the target covers a unit circle and an isotropic, convex and monotonically decreasing kernel profile $k(x)$ is used. Function $b: R^2 \to 1\dots m$ maps the value of the pixel at location $\mathbf{x}_i$ to the index $b(\mathbf{x}_i)$ of the corresponding bin in the feature space. The probability of the feature $u \in \{1,\dots,m\}$ is estimated by the target histogram as follows:

$$\hat{q}_u = C\sum_{i=1}^{n} k\left(\|\mathbf{x}_i^*\|^2\right)\delta[b(\mathbf{x}_i^*) - u], \tag{3}$$

where δ is the Kronecker delta and C is a normalization constant so that $\sum_{u=1}^{m}\hat{q}_u = 1$.

Let $\{\mathbf{x}_i\}_{i=1\dots n_h}$ be pixel locations in the current frame where the target candidate is centered at location $\mathbf{y}$ and n_h be the number of pixels of the target candidate. Using the same kernel profile $k(x)$, but with a scale parameter h, the probability of the feature $u = 1\dots m$ in the target candidate is

$$\hat{p}_u(\mathbf{y}) = C_h \sum_{i=1}^{n_h} k\left(\left\|\frac{\mathbf{y} - \mathbf{x}_i}{h}\right\|^2\right)\delta[b(\mathbf{x}_i) - u], \tag{4}$$

where C_h is a normalization constant. The difference between probability distributions $\hat{\mathbf{q}} = \{\hat{q}_u\}_{u=1\dots m}$ and $\{\hat{p}_u(\mathbf{y})\}_{u=1\dots m}$ is measured by the Hellinger distance of probability measures, which is known to be a metric:

$$H(\hat{\mathbf{p}}(\mathbf{y}), \hat{\mathbf{q}}) = \sqrt{1 - \rho[\hat{\mathbf{p}}(\mathbf{y}), \hat{\mathbf{q}}]}, \tag{5}$$

where

$$\rho[\hat{\mathbf{p}}(\mathbf{y}), \hat{\mathbf{q}}] = \sum_{u=1}^{m}\sqrt{\hat{p}_u(\mathbf{y})\hat{q}_u} \tag{6}$$

is the Bhattacharyya coefficient of $\hat{\mathbf{q}}$ and $\hat{\mathbf{p}}(\mathbf{y})$. Minimizing the Hellinger distance is equivalent to maximizing the Bhattacharyya coefficient $\rho[\hat{\mathbf{p}}(\mathbf{y}), \hat{\mathbf{q}}]$. The search for the new target location in the current frame starts at location $\hat{\mathbf{y}}_0$ of the target in the previous frame using gradient ascent with a step size equivalent to the mean-shift method. The kernel is repeatedly moved from the current location $\hat{\mathbf{y}}_0$ to the new location

$$\hat{\mathbf{y}}_1 = \frac{\sum_{i=1}^{n_h}\mathbf{x}_i w_i g\left(\left\|\frac{(\hat{\mathbf{y}}_0 - \mathbf{x}_i)}{h}\right\|^2\right)}{\sum_{i=1}^{n_h} w_i g\left(\left\|\frac{(\hat{\mathbf{y}}_0 - \mathbf{x}_i)}{h}\right\|^2\right)}, \tag{7}$$

where

$$w_i = \sum_{u=1}^{m}\sqrt{\frac{\hat{q}_u}{\hat{p}_u(\hat{\mathbf{y}}_0)}}\delta[b(\mathbf{x}_i) - u] \tag{8}$$

and $g(x) = -k'(x)$ is the derivative of $k(x)$, which is assumed to exist for all $x \geqslant 0$, except for a finite set of points.

2.2. Scale estimation

Let us assume that the scale changes frame to frame in an isotropic manner[1]. Let $\mathbf{y} = (y^1, y^2)^T$, $\mathbf{x}_i = (x_i^1, x_i^2)^T$ denote pixel locations and N be the number of pixels in the image. A target is represented by an ellipsoidal region $\frac{(x_i^{*1})^2}{a^2} + \frac{(x_i^{*2})^2}{b^2} < 1$ in the image and an isotropic kernel with profile $k(x)$ as in [3], restricted by a condition $k(x) = 0$ for $x \geqslant 1$, is used. The probability of the feature $u \in \{1,..,m\}$ is estimated by the target histogram as

$$\hat{q}_u = C\sum_{i=1}^{N} k\left(\frac{(x_i^{*1})^2}{a^2} + \frac{(x_i^{*2})^2}{b^2}\right)\delta[b(\mathbf{x}_i^*) - u], \tag{9}$$

where C is a normalization constant. Let $\{\mathbf{x}_i\}_{i=1\dots N}$ be the pixel locations of the current frame in which the target candidate is centered at location $\mathbf{y}$. Using the same kernel profile $k(x)$, the probability of the feature $u = 1\dots m$ in the target candidate is given by

$$\hat{p}_u(\mathbf{y}, h) = C_h \sum_{i=1}^{N} k\left(\frac{(y^1 - x_i^1)^2}{a^2h^2} + \frac{(y^2 - x_i^2)^2}{b^2h^2}\right)\delta[b(\mathbf{x}_i) - u], \tag{10}$$

where

$$C_h = \frac{1}{\sum_{i=1}^{N} k\left(\frac{(y^1 - x_i^1)^2}{a^2h^2} + \frac{(y^2 - x_i^2)^2}{b^2h^2}\right)}. \tag{11}$$

The parameter h defines the scale of the target candidate and thus the number of pixels with non-zero values of the kernel function.

For a given kernel and variable h, C_h can be approximated in the following way: Let n_1 be the number of pixels in the ellipsoidal region of the target model, and let n_h be the number of pixels in the ellipsoidal region of the target candidate with a scale h; then $n_h \doteq h^2 n_1$. Using the definition of Riemann integral we obtain:

$$\begin{aligned}\sum_{i=1}^{N} k\left(\frac{(x_i^1)^2}{a^2h^2} + \frac{(x_i^2)^2}{b^2h^2}\right)\frac{\pi abh^2}{n_h} &\approx \int\int_{\left\{\frac{(x^1)^2}{a^2h^2} + \frac{(x^2)^2}{b^2h^2} < 1\right\}} k\left(\frac{(x^1)^2}{a^2h^2} + \frac{(x^2)^2}{b^2h^2}\right)dx^1dx^2 \\ &= h^2ab\int\int_{\|\mathbf{x}\|<1} k(\|\mathbf{x}\|^2).\end{aligned} \tag{12}$$

[1] Generalization to the anisotropic where $\mathbf{h} = (h^1, h^2)^T$ is straightforward.