Pattern Recognition Letters 29 (2008) 2165–2173

Contents lists available at ScienceDirect

Pattern Recognition Letters

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

A spatial-color mean-shift object tracking algorithm with scale and orientation estimation

Jwu-Sheng Hu *, Chung-Wei Juan, Jyun-Ji Wang

Department of Electrical and Control Engineering, National Chiao Tung University, Hsinchu 300, Taiwan, ROC

ARTICLE INFO

Article history:
Received 30 October 2007
Received in revised form 26 May 2008
Available online 22 August 2008

Communicated by Y. Ma

Keywords:
Mean-shift
Object tracking
Principle component analysis
Object deformation

ABSTRACT

In this paper, an enhanced mean-shift tracking algorithm using joint spatial-color feature and a novel similarity measure function is proposed. The target image is modeled with the kernel density estimation and new similarity measure functions are developed using the expectation of the estimated kernel density. With these new similarity measure functions, two similarity-based mean-shift tracking algorithms are derived. To enhance the robustness, the weighted-background information is added into the proposed tracking algorithm. Further, to cope with the object deformation problem, the principal components of the variance matrix are computed to update the orientation of the tracking object, and corresponding eigenvalues are used to monitor the scale of the object. The experimental results show that the new similarity-based tracking algorithms can be implemented in real-time and are able to track the moving object with an automatic update of the orientation and scale changes.

1. Introduction

In visual tracking, object representation is an important issue, because it can describe the correlation between the appearance and the state of the object. An appropriate object representation makes the target model more distinguishable from the background, and achieves a better tracking result. Comaniciu et al. (2003) used the spatial kernels weighted by a radially symmetric normalized distance from the object center to represent blob objects. This representation makes mean-shift tracking more efficient. Radially symmetric kernel preserves representation of the distance of a pixel from the center even the object has a large set of transformations, but this approach only contains the color information of the target and the spatial information is discarded. Parameswaran et al. (2006) proposed the tunable kernels for tracking, which simultaneously encodes appearance and geometry that enable the use of mean-shift iterations. A method was presented to modulate the feature histogram of the target that uses a set of spatial kernels with different bandwidths to encode the spatial information. Under certain conditions, this approach can solve the problem of similar color distribution blocks with different spatial configurations.

Another problem in the visual tracking is how to track the scale of object. In the work by Comaniciu et al. (2003), the mean-shift algorithm is run several times, and for each different window size, the similarity measure Bhattacharyya coefficient is computed for comparison. The window size yielding the largest Bhattacharyya coefficient, i.e. the most similar distribution, is chosen as the updated scale. Parameswaran et al. (2006), Birchfield and Rangarajan (2005) and Porikli and Tuzel (2005) use the similar variation method to solve the scale problem. But this method is not always stable, and easily makes the tracker lose the target. Collins (2003) extended the mean-shift tracker by adapting Lindeberg's theory (Lindeberg, 1998) of feature scale selection based on local maxima of differential scale-space filters. It uses blob tracking and a scale kernel to accurately capture the target's variation in scale. But the detailed iteration method was not described in the paper. An EM-like algorithm (Zivkovic and Krose, 2004) was proposed to estimate simultaneously the position of the local mode and used the covariance matrix to describe the approximate shape of the object. However, implementation details such as deciding the scale size from the covariance matrix were not given. Other attempts were made to study different representation methods. Zhang et al. (2004) represented the object by a kernel-based model, which offers more accurate spatial–spectral description than general blob models. Later, they further extend the work to cope with the scaling and rotation problem under the assumption of affine transformation (Zhang et al., 2005). Zhao and Tao (2005) proposed the color correlogram to use the correlation of colors to solve the related problem. But these methods did not consider the influence of complex background.

This work extends the traditional mean-shift tracking algorithm to improve the performance of arbitrary object tracking. At the same time, the proposed method tries to estimate the scale and orientation of the target. This idea is similar to the CAMSHIFT

* Corresponding author. Tel.: +886 3 5712121x54318; fax: +886 3 5715998.
E-mail address: jshu@cn.nctu.edu.tw (J.-S. Hu).

0167-8655/$ - see front matter © 2008 Elsevier B.V. All rights reserved.
doi:10.1016/j.patrec.2008.08.007

algorithm (Bradski, 1998) except spatial probability information as well as background influence are considered. The subject of this paper is divided into two parts. The first part is to develop the new spatial-color mean-shift trackers for the purpose of capturing the target more accurately than the traditional mean-shift tracker. The second part is to develop a method for solving the scale and orientation problem mentioned above. The solution, though seems straightforward, has never been proposed in literature. The effectiveness proved by experiments shows a further enhancement on the mean-shift algorithm.

2. Model definition

Birchfield and Rangarajan (2005) proposed the concept of spatial histogram, or spatiogram, in which each histogram bin contains the mean and covariance information of the locations of the pixels belonging to that bin. This idea involves the spatially weighted mean and covariance of the locations of the pixels. The spatiogram captures the spatial information of the general histogram bins. However, as shown in Fig. 1, if cyan and blue belong to the same bin, these two blocks have the same spatiogram, even though they have different color patterns.

Let the image of interest have M pixels and the associated color space can be classified into B bins. For example, in RGB color space, if each color is divided into 8 intervals, the total number of bins is 512. The image can be described as $I_x = \{\boldsymbol{x}_i, \boldsymbol{c}_{\boldsymbol{x}_i}, b_{\boldsymbol{x}_i}\}_{i=1,\dots,M}$ where $\boldsymbol{x}_i$ is the location of pixel i with color feature vector $\boldsymbol{c}_{\boldsymbol{x}_i}$ which belongs to the $b_{\boldsymbol{x}_i}$th bin. The color feature vector has the dimension d which is the color channels for the pixel (for example, in RGB color space, $d = 3$ and $\boldsymbol{c}_{\boldsymbol{x}_i} = (R_{\boldsymbol{x}_i}, G_{\boldsymbol{x}_i}, B_{\boldsymbol{x}_i})$). To keep the robustness of color description of the spatiogram, we extend the spatiogram and define a new joint spatial-color model of the image I_x as

$$h_{I_x}(b) = \left\langle n_b, \boldsymbol{\mu}_{P,b}, \sum_{P,b}, \boldsymbol{\mu}_{C,b}, \sum_{C,b} \right\rangle, \quad b = 1, \dots, B \tag{1}$$

where n_b, $\mu_{P,b}$, and $\sum_{P,b}$ are the same as the spatiogram proposed by Birchfield and Rangarajan (2005). Namely, n_b is the number of pixels, $\mu_{P,b}$ the mean vector of pixel locations, and $\sum_{P,b}$ the covariance matrix of pixel locations belonging to the bth bin. In (1), we add two additional elements. $\mu_{C,b}$ is the mean vector of the color feature vectors and $\sum_{C,b}$ is the associated covariance matrix.

3. Spacial-colour mean-shift object tracking algorithm

Using the spatial-color feature and the concept of expectation (Yang et al., 2005), two different tracking algorithms are proposed as the following.

3.1. Spatial-color mean-shift tracking algorithm (tracker 1)

The p.d.f. of the selected pixel $\boldsymbol{x}, \boldsymbol{c}_{\boldsymbol{x}}, b_{\boldsymbol{x}}$ in the image model $h_{I_x}(b)$ (see (1)) can be estimated using kernel density function.

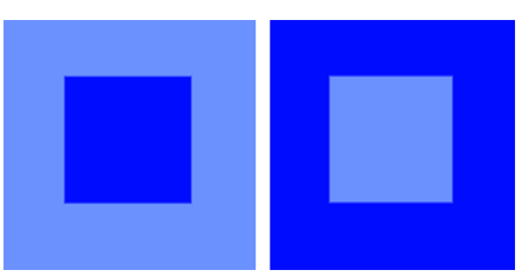

Fig. 1. Illustration of the same spatial information with different color distribution for one bin.

$$p(\boldsymbol{x}, \boldsymbol{c}_{\boldsymbol{x}}, b_{\boldsymbol{x}}) = \frac{1}{B}\sum_{b=1}^{B} K_P\left(\boldsymbol{x} - \boldsymbol{\mu}_{P,b}, \sum_{P,b}\right) K_C\left(\boldsymbol{c}_{\boldsymbol{x}} - \boldsymbol{\mu}_{C,b}, \sum_{C,b}\right)\delta(b_{\boldsymbol{x}} - b) \tag{2}$$

where K_P and K_C are multivariate Gaussian kernel functions and can be regarded as the spatially weighted and color-feature weighted function respectively. It is also possible to use a smooth kernel such as Gaussian (Yang et al., 2005). Using the concept of the expectation of the estimated kernel density, we can define a new similarity measure function between the model $h_{I_x}(b)$ and candidate $I_y = \{\boldsymbol{y}_j, \boldsymbol{c}_{\boldsymbol{y}_j}, b_{\boldsymbol{y}_j}\}_{j=1,\dots,N}$ as

$$\begin{aligned} J(I_x, I_y) = J(\boldsymbol{y}) &= \frac{1}{N}\sum_{j=1}^{N} p(\boldsymbol{y}_j, \boldsymbol{c}_{\boldsymbol{y}_j}, b_{\boldsymbol{y}_j}) \\ &= \frac{1}{NB}\sum_{j=1}^{N}\sum_{b=1}^{B} K_P\left(\boldsymbol{y}_j - \boldsymbol{\mu}_{P,b}, \sum_{P,b}\right) K_C\left(\boldsymbol{c}_{\boldsymbol{y}_j} - \boldsymbol{\mu}_{C,b}, \sum_{C,b}\right)\delta(b_{\boldsymbol{y}_j} - b) \end{aligned} \tag{3}$$

The spatial-color model in (2) under measures like (3) might be sensitive to small spatial changes. This problem was discussed by O'Conaire et al. (2007) and Birchfield and Rangarajan (2007). However, this model also gives advantages of orientation estimation. As shown in Fig. 2, if there is no deformation between candidate and target, and the distance of motion is not excessively large between two adjacent frames, we can consider the motion of object of two frames as a pure translation. Under these assumptions, the center of target model $\boldsymbol{x}_0$ is proportional to the center of candidate $\boldsymbol{y}$ in the candidate image. As a result, we can normalize the pixels location and then obtain the new similarity measure function as the following:

$$\begin{aligned} J(\boldsymbol{y}) = \frac{1}{NB}\sum_{j=1}^{N}\sum_{b=1}^{B} & K_P\left(\boldsymbol{y}_j - \boldsymbol{y} - (\boldsymbol{\mu}_{P,b} - \boldsymbol{x}_0), \sum_{P,b}\right) \\ & \times K_C\left(\boldsymbol{c}_{\boldsymbol{y}_j} - \boldsymbol{\mu}_{C,b}, \sum_{C,b}\right)\delta(b_{\boldsymbol{y}_j} - b) \end{aligned} \tag{4}$$

The best candidate for matching can be found by computing the maximum value of the similarity measure. Let the gradient of the similarity function with respect to the vector $\boldsymbol{y}$ equal to $\boldsymbol{0}$, i.e., $\nabla J(\boldsymbol{y}) = \boldsymbol{0}$, we can obtain the new position $\boldsymbol{y}_{\text{new}}$ of the target to be tracked,

$$\begin{aligned} & \nabla J(\boldsymbol{y}) = \boldsymbol{0} \\ & \Rightarrow \frac{1}{NB}\sum_{j=1}^{N}\sum_{b=1}^{B}\left(\sum_{P,b}\right)^{-1}(\boldsymbol{y}_j - \boldsymbol{y} - \boldsymbol{\mu}_{P,b} + \boldsymbol{x}_0)K_P K_C \delta(b_{\boldsymbol{y}_j} - b) = \boldsymbol{0} \\ & \Rightarrow \left\{\sum_{j=1}^{N}\sum_{b=1}^{B}\left(\sum_{P,b}\right)^{-1} K_P K_C \delta(b_{\boldsymbol{y}_j} - b)\right\}(\boldsymbol{y} - \boldsymbol{x}_0) \\ & = \sum_{j=1}^{N}\sum_{b=1}^{B}\left(\sum_{P,b}\right)^{-1}(\boldsymbol{y}_j - \boldsymbol{\mu}_{P,b)} K_P K_C \delta(b_{\boldsymbol{y}_j} - b) \end{aligned}$$

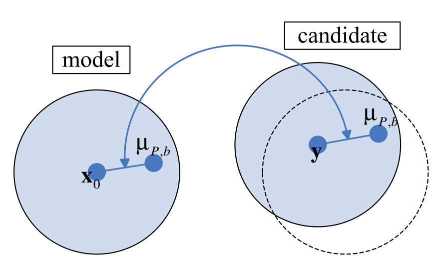

Fig. 2. Illustration of pure translation.