

ONLINE FEATURE EVALUATION FOR OBJECT TRACKING USING KALMAN FILTER

Zhenjun Han, Qixiang Ye, Jianbin Jiao+

Graduate University of Chinese Academy of Sciences, Beijing, 100049, China

+Corresponding Author: Fax: +86-10-88256278, Email: jiaojb@gucas.ac.cn

Abstract

An online feature evaluation method for visual object tracking is put forward in this paper. Firstly, a combined feature set is built using color histogram (HC) bins and gradient orientation histogram (HOG) bins considering the color and contour representation of an object respectively. Then a novel method is proposed to evaluate the features' weights in a tracking process using Kalman Filter, which is used to comprise the inter-frame predication and single-frame measurement of features' discriminative power. In this way, we extend the traditional filter framework from modeling motion states to modeling feature evaluation. Experiments show this method can greatly improve the tracking stabilization when objects go across complex backgrounds.

1. Introduction

Visual tracking is significant for computer vision systems. It is a prerequisite for understanding video contents, a crucial step for visual surveillance applications, human computer interaction systems and robotics [2], etc. It has been extensively studied in the past decade, mainly on object representation, motion modeling and searching methods [3] etc.

Object representation is the cornerstone of tracking. A good representation should be robust to object rotation, scale variation, partial occlusion etc, in a tracking process. In the past years, color histogram (HC) features are widely used in tracking [2][4] for its effectiveness and efficiency. The disadvantage of HC is that it fails when the tracked object has similar color with its background. Recently, histogram of orientations gradient (HOG) [6] is widely applied for object detection and tracking. HOG captures edges or gradient structure, which are the intrinsic characteristics of local contours and shapes. In [6], Dalal et al. has justified that the representation ability of HOG is

almost as strong as SIFT [8] descriptor given a fixed scale. However, HOG also have some disadvantages. For example, its efficiency is lower than HC and can not represent object with large smooth regions effectively since the contours of these objects are indistinctive.

A problem of object tracking which needs to be solved is that foreground and background appearances are changing, during which the discriminative power of feature will vary. For example, green color components of HC have a weak discriminative power when both the object and the background have large green components. Therefore, features should be evaluated online to improve the tracking robustness. Colloins et al. [4] define the discriminative power of a feature according to the two-class variance ratio measures of the object and its background, then the top N discriminative features are selected for tracking. Liang et al. [2] propose a similar approach, in which the discriminative power of a feature is calculated based on Bayes error rate between the object and its background. Bayes error rate of a feature is calculated by the intersection of the likelihood function of the object and its background on the feature. Wang et al. [5] propose a method to online select a subset of Haar wavelet features by Adaboost learning. Chen et al. [3] propose a hierarchical Monte Carlo algorithm to evaluate region confidences for object tracking online. However, these existing methods usually evaluated features by considering the discriminative power in one or multiple frames, while the prediction of the features' discriminative power from one frame to another is seldom considered.

In this paper, we use a combined feature set of HC and HOG, called HOGC, for object tracking. The combined feature set is the evolvement of color, edge orientation histograms [6] and SIFT descriptors [8]. For the feature set, we put forward an online method to evaluate the weight value of each feature bin based on the foreground-background discriminative power. The evaluation adjusts the weight value of each feature bin

by a Kalman filter by comprehensively considering the discriminative powers of the feature bins in current frame and the weigh value of each feature bin in the previous frames. Evaluated features will be recombined after multiply with new weights for tracking.

The rest of the paper is organized as follows: the tracking algorithm is presented in section 2, the experiments in section 3 and conclusion in section 4.

2. Object tracking algorithm

The flow chart of the tracking algorithm is shown in Fig.1. Features of both the object and its background will be firstly extracted and evaluated. Then a fast image block match algorithm on HOGC is carried out to track the object.

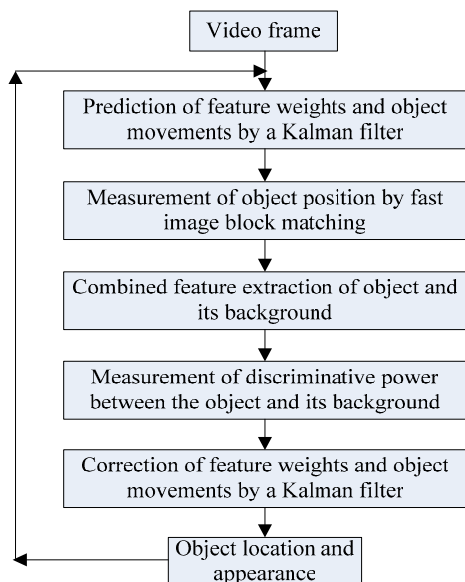


Fig.1. Flow chart of the proposed tracking algorithm.

2.1. Combined features for object representation

In a video frame, we need first to define an object region and its background region in order to obtain the discriminative power between the object and its background. If an object is represented by a rectangle area of $h \times w$ pixels, then its background is defined as the areas around it, as showed in Fig.2. The size of the outline rectangle of the background area is defined as $\sqrt{2}h \times \sqrt{2}w$ pixels in this paper empirically. The object features $\{F_i(x, y)\}, i = 0, 1, \dots, N$ are extracted from the pixels in (x, y) centered object area, and background features $\{B_i(x, y)\}, i = 0, 1, \dots, N$ are from pixels in background area whose center is (x, y) , where N is the feature dimension.

Color histogram: We define a color histogram (HC) of 48 dimensions for both the object and its background. In each color component in RGB color space, 16 dimensions of histogram features are calculated.

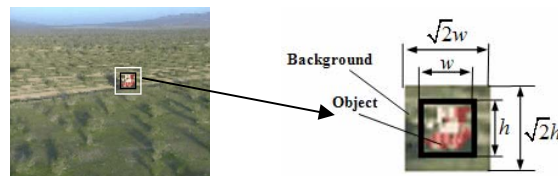


Fig.2. Object and background regions.

Gradient orientation histogram: We follow the idea of [6] to extract HOG features on gray value image regions. On each window, a histogram of 72 dimensions is extracted to describe the gradient orientation of an object.

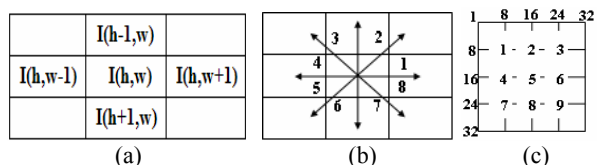


Fig.3. HOG extraction. (a) Mask for pixel gradient calculation, (b) Orientation bin for voting, (c) 9 blocks in the image window.

Detail steps of HOG features extraction are described as follows. Grayscale space with no gamma correction; resizing the object rectangle region into an image window of fixed size, say, 32×32 pixels; then we divide the image window into small spatial regions (“cells”) with the size of 8×8 and each group of 2×2 cells is integrated into a block in a sliding fashion, therefore block overlaps each other; different from the method in [6], each block is only consisted of a 8-bin HOG without local normalization. Each pixel in the block calculates $ori(h, w)$ for orientation histogram channel based on the orientation of the gradient element centered on it, the mask for the calculation of $ori(h, w)$ is showed in Fig.3a. In our approach, $ori(h, w)$ of each pixel is calculated using the Eq. (1).

$$\begin{aligned}
 I &= G(\sigma, 0) * I_0 \\
 dy &= I(h+1, w) - I(h-1, w) \\
 dx &= I(h, w+1) - I(h, w-1) \\
 ori(h, w) &= a \tan 2(dy, dx) \quad ori \in [-\pi, \pi]
 \end{aligned} \quad (1)$$

where $G(\sigma, 0)$ is a Gaussian scale function and σ is determined by experiment. Votes of $ori(h, w)$ in a block are accumulated into 8 orientation histogram bins, showed in Fig.3b, to form a HOG feature of the block. Then we combine the HOG of each block to obtain a 72 dimension features (9 blocks in the image window

showed in Fig.3c) for the whole object.

2.2. Feature evaluation and movement updating in Kalman filter framework

Extracted HOGC features will be evaluated during the tracking process, which is under the following constrains: (1) Weight value or discriminative power of a feature is reflected by a float value fallen into (0.0, 1.0). (2) After the evaluation, features of higher discriminative power should have larger weight values, and vice versa. Simply, weight value of a feature is a linear function of the discriminative power. (3) Weight value and discriminative power of a feature are both with Gaussian distribution;

We define the feature's discriminative power S_i^t between object and its background in the t th frame as

$$S_i^t = \frac{|B_i^t(x, y) - F_i^t(x, y)|}{B_i^t(x, y) + F_i^t(x, y)}, i = 1 \dots N, \quad (2)$$

where $F_i^t(x, y)$ is the i th feature bin of the object at frame t , $B_i^t(x, y)$ is the i th feature bin of the background at time t and N is the feature dimensions. Then we use the Eq. (2) to normalize the discriminative power of each feature bin. Eq. (2) indicates that the larger S_i^t , the more discriminative between the object and its background.

$$S_i^t = \frac{S_i^t}{\sum_{i=1}^N S_i^t}, i = 1 \dots N. \quad (3)$$

In this paper, we put forward a method to integrate our feature evaluation into the Kalman filter [1]. Let's define the state of the Kalman filter, including the feature weights and the movement of the tracked object, while the measurements including the discriminative powers of the combined features and the position of the tracked object. This can induce a Kalman filter as

$$\begin{cases} \begin{pmatrix} W^{t+1} \\ \Delta W^{t+1} \\ Pos^{t+1} \\ \Delta Pos^{t+1} \end{pmatrix} = \begin{pmatrix} I_{N \times N} & I_{N \times N} & 0 & 0 \\ 0 & I_{N \times N} & 0 & 0 \\ 0 & 0 & I_{M \times M} & I_{M \times M} \\ 0 & 0 & 0 & I_{M \times M} \end{pmatrix} \begin{pmatrix} W^t \\ \Delta W^t \\ Pos^t \\ \Delta Pos^t \end{pmatrix} + u_t \\ \begin{pmatrix} S^t \\ mPos^t \end{pmatrix} = \begin{pmatrix} I_{N \times N} & 0 & 0 & 0 \\ 0 & 0 & I_{M \times M} & 0 \end{pmatrix} \begin{pmatrix} W^t \\ \Delta W^t \\ Pos^t \\ \Delta Pos^t \end{pmatrix} + v_t \end{cases} \quad (4)$$

where $W^t = \{W_1^t, W_2^t, \dots, W_N^t\}$ is the weight vector of the feature set at time t , and W_i^t is the weight value of the i th feature bin, and $\Delta W^t = W^t - W^{t-1}$. While $S^t = \{S_1^t, S_2^t, \dots, S_N^t\}$ is the discriminative power

vector at time t , and S_i^t is the discriminative power of the i th feature bin and can be calculated by Eq. (2) and (3), and $\Delta S^t = S^t - S^{t-1}$. $Pos^t = \{x, y\}^t$ is the location of the tracked object we predict and $\Delta Pos^t = Pos^t - Pos^{t-1}$. $mPos^t = \{mx, my\}^t$, the location obtained during the matching procedure. Therefore M is 2 in our approach. u_t and v_t are both Gaussian white noises empirically.

The following are the steps of our tracking approach:

1. Initialization ($t=0$). In this step, we initiate weight value of each feature bin $(W_i^0, \Delta W_i^0)^T = \left(\frac{1}{N}, 0\right)^T$; Pos^0 is the position where the tracked object is initialized in the first frame, and $\Delta Pos^0 = 0$.

2. Prediction ($t>0$). Using the Kalman filter to predict the prior weight of each feature bin and the position of the tracked object. Then we use the prior HOGC and position to guide the object position searching (the matching procedure) in next frame.

3. Correction ($t>0$). After we obtain the best match of the object in the searching region in the next frame by Eq. (5), we use the discriminative power and the best candidate position to carry out the state correction using the Kalman filter finding the posterior weight of each feature bin and the posterior position of the object.

The object's best matching position $(x, y)^t$ at t frame can be searched by

$$\text{Min}_c \left(\sum_{i=1}^N \left(W_i^t \times |F_i^{t-1}((x, y)^{t-1}) - F_i^t((x, y)^t)| \right) \right). \quad (5)$$

where $(x, y)_c^t$ denotes the c th candidate position in the t frame. In each weight updating iteration, we use

$W_i^t = \frac{W_i^t}{\sum_{i=1}^N W_i^t}$ to normalize the weight of each feature bin to ensure $\sum_i W_i^t = 1$.

3. Experiments

Our algorithms had been tested on several video datasets, most of which are publicly available [7]. The selected videos are challenging. The appearances of the object and its background keep changing during almost all the time. Nevertheless, exciting results are obtained in most experiments.



Frame 50

Frame 300

Frame 700

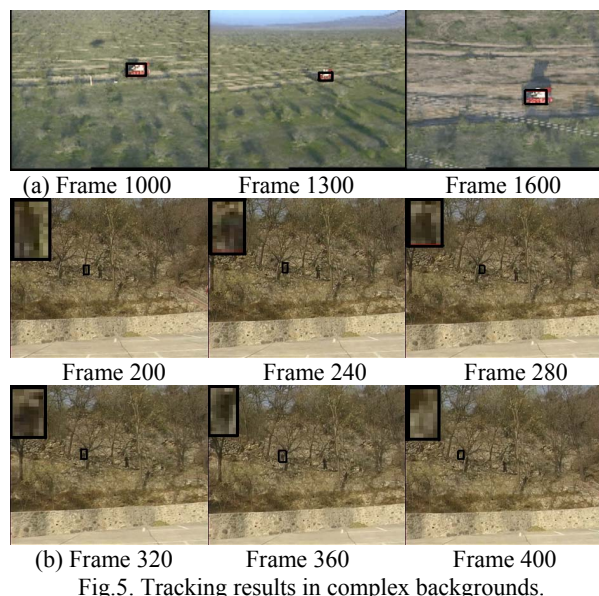


Fig.5. Tracking results in complex backgrounds.

Fig.5 shows the tracking results of our approach. The background are quite complex, especially Fig.5b illustrates the most challenging video in our experiments. The top-left rectangle represents the person being tracked. The background has the similar

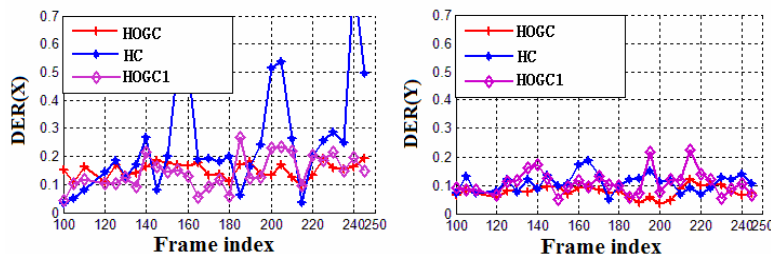


Fig.6. Displacement Error Rates (DER) in X and Y direction of our method, color histogram and HOGC1 (HOGC-based tracking without weight evaluation) based tracking

4. Conclusions and future works

We presented an approach to evaluate combined features for robust tracking of objects against complex background. Experiments proved that HOGC are effective combined features for object tracking. Experiments also validate the feature evaluation approach. In the future, we will test more features in the proposed feature evaluation method.

5. Acknowledgment

This work is supported in part by “Science 100 Plan” of Chinese Academy of Sciences and Chinese National Science Foundation under Grant No. 60672147.

6. References

[1] E. Cuevas, D. Zaldivar and R. Rojas. Kalman Filter for Vision Tracking. *Technical Report B*, 2005.

color to the tracking object (a person in small size) and there are some small trees which is quite similar to the object in shape. Our proposed approach can still track the object robustly even in such a complex circumstance.

To quantitatively evaluate the proposed method, we define relative displacement error rates (DER). The DER is calculated by

$$DER = \frac{\text{displacement error between tracked object position and groundtruth}}{\text{Size of the object}}$$

In our experiments we use the DER in X and Y directions of 30 video clips to evaluate the tracking algorithm, which are showed in Fig.6. It can be seen on the figure that the DER (about 0.1 to 0.2) is quite small in the whole tracking process. We also compare DER of our method with the method that only use single feature set (color histogram). Results show that the proposed combined feature set (HOGC) with online feature evaluation have a much lower tracking DER than the tracking results on color histogram feature set (Fig.6). It can also be seen on the figure that the evaluated HOGC works steadily than un-evaluated one (HOGC1). As for tracking speed, our method can work in real time on a computer with Pentium IV CPU (2.4G).

[2] D.W. Liang, Q.M. Huang, W. Gao, and H.X. Yao. Online Selection of Discriminative Features Using Bayes Error Rate for Visual Tracking. *7th Pacific-Rim Conference on Multimedia*, 547–555, 2006.

[3] D. Chen, and J. Yang. Robust Object Tracking via Online Spatial Bias Appearance Model Learning. *IEEE Trans. PAMI*, Vol 29, 2157–2169, 2007.

[4] R. Collins, and Y. Liu. Online Selection of Discriminative Tracking Features. *Proceedings of Ninth IEEE ICCV. Vol 1*, 346–352, 2003.

[5] J. Wang, X. Chen, and W. Gao. Online Selecting Discriminative Tracking Features Using Particle Filter. *Proceedings of IEEE CVPR. Vol. 2*, 1037–1042, 2005.

[6] N. Dalal, and B. Triggs. Histograms of Oriented Gradients for Human Detection. *CVPR*, 1063–6919, 2005.

[7] VIVID Tracking Evaluation Web Site at: <http://www.vividevaluation.ri.cmu.edu/datasets/datasets.html>.

[8] D.G. Lowe. Distinctive Image Features from Scale-Invariant Keypoints. *IJCV*, 91–110, 2004.