A Two-phase Tracking Method based on Structural Entropy Appearance Model

Dawei YANG, Fan ZHOU

School of Information Technology and Engineering, Shenyang ligong University, Liaoning Shenyang 110159, China

*Corresponding author(Fan ZHOU, zhoufan@sylu.edu.cn)

Abstract

Aiming at the problem caused by partial occlusion, illumination and pose change, a two-phase tracking method base on structural entropy appearance model is proposed in the paper. The structural entropy appearance model is constructed firstly, then it is combined with the naïve Bayesian classifier. Object tracking is carried out within the Bayesian inference framework: in the initial tracking phase, the preliminary estimation of object location is achieved by using the principle of particle filter; in the calibrating phase, the final target location is firmed. Experimental results on some video sequences demonstrate the effectiveness and robustness of the proposed method.

Keywords: Object tracking, Structural entropy appearance model, Two-phase searchin

1. INTRODUCTION

Object tracking has a wide range of applications in image understanding, human-computer interaction, intelligent monitoring, robotics, and others. In recent years, although a lot of tracking algorithms have been proposed, to design a robust tracking algorithm is still a challenging task due to factors such as illumination change, occlusion, pose change and complex background. The proposed algorithms are divided into two categories according to the appearance model: the generative model tracking algorithm and the discriminative model tracking algorithm.

The generative tracking methods first learn an appearance model to represent the object, and then search the target area with the largest similarity as the tracking result. Ross et al. proposed an incremental visual tracking (IVT), which represents the object with a low dimensional PCA subspace and assumes that the error term is Gaussian distribution using the least square method to solve the objective function. Although IVT is effective for appearance change caused by illumination change and pose change, it still does not track well for some challenging factors (such as partial occlusion and background disturbance). Adam et al. designed an appearance model using multiple pieces, which is robust to partial occlusion. Recently, sparse representations have been introduced to the tracking task, and show good results for partial occlusion, illumination change and pose change. However, the generative methods discard useful information surrounding the target area, these information can better distinguish the object from background.

The discriminative tracking methods view tracking as a binary classification problem, and estimate decision boundary using the local search within the blocks of target image and background. These methods are also called detection-tracking, namely viewing tracking as a detection task. Grabner et al. proposed online Boosting feature selection. However, this method uses only one positive sample and multiple negative samples to update the classifier. If the object location detected by the current classifier is not accurate, extracted samples will also be inaccurate, leading to a suboptimal update classifier. The cumulative errors will seriously degrade the performance of the classifier, eventually leading to failure (drift). Multiple instance learning (MIL) tracker puts the positive and negative samples respectively into the positive and negative bag, using bag similarity function to train classifiers online. This method has been proved to deal with drift problem very well. However, the noise or model used in this method does not take into account the importance of positive sample. Therefore, the tracker may choose a few less effective positive samples.

In order to deal with the drift problem caused by illumination change, partial occlusion and pose change, an object tracking method based on structural entropy appearance model is proposed in this paper. The object image is divided into some blocks which keeping in spatial structure, the entropy of each block is calculated, then the feature vector of the appearance is formed using these entropies. The tracking is carried out in the frame of Bayes inference of two-phase. In the initial tracking phase, the preliminary estimation of object location is achieved by using the principle of particle filter. In the calibrating phase, the adjustment is made to determine the final target location. Experiments on several challenging video sequences and comparison with existing two-phase particle filter methods verified that the proposed method has good tracking performance.

2. STRUCTURAL ENTROPY APPEARANCE MODEL
A random matrix \( \mathbf{R} \in \mathbb{R}^{m \times n} \) with unit length line, maps the data from a high dimension \( x \in \mathbb{R}^m \) to a low dimension \( v \in \mathbb{R}^n \).

\[
\begin{align*}
v \in \mathbb{R}^x
\end{align*}
\]  \hspace{1cm} (1)

Where, \( n \ll m \). In the ideal state, we hope that the distance between the original signals can be conserved. It was pointed out in the J-L lemma that the distance between the points in a vector space can be conserved when they were mapped into a random selected subspace. Baraniuk et al. verified that the random matrix satisfying the J-L lemma met the restricted isometry property (RIP) in the compressed sensing. When the random matrix \( \mathbf{R} \) in equation (1) satisfies the J-L lemma, and \( x \) is compressible, it can be reconstructed from \( v \) with smallest errors, and almost all the information in \( x \) is conserved. The powerful theory supports us to analysis the high dimensional signals using low dimensional random mapping.

Sparse representation has been applied in many visual applications widely. A signal can be represented as a linear combination of a few base vectors using sparse constraint. Local sparse representation has been used to modeling appearance model in object tracking, histogram of sparse coding is used to represent object feature. The method using local block to represent object is useful when there is partial occlusion in object tracking. While the method can't provide enough spatial information needed in tracking.

An analogous local sparse representation is employed in this paper. Sampling image blocks in target region according fixed spatial arrangement using analogous method of grid, so enough spatial structure information is reserved. A sparse matrix with few nonzero elements is showed in figure 1.

![Figure 1. The diagram of sparse matrix](image1)

Where, the black blocks represent nonzero elements in the matrix, and white blocks represent zero elements. The matrix is applied on the target region in the tracking procedure, amount to extract a number of small image blocks with \( n \) lines and \( m \) rows in the target region. A feature is calculated in each image block, and a feature vector is formed using these local features. Then, a high dimensional image region is mapped sparsely into a low dimensional vector.

Image entropy is used to statistic amount of information in the image. One dimensional entropy represents gathering character of distribution of gray levels in image. Let \( p_i \) represents the proportion of pixels which gray value is \( i \) in the image, the form of one dimensional entropy definition is:

\[
H = \sum_{i=0}^{255} p_i \log p_i
\]  \hspace{1cm} (2)

In this paper, the entropy feature and image gridding are combined to represent the target feature. That is, each candidate target image is normalized to the size of \( 32 \times 32 \), then, sampling image blocks in each target region using analogous gridding. This representation using structural image blocks dose well in partial occlusion. Calculating entropy feature in each black according to equation (2), then, a high dimensional image target is mapped to a low dimensional vector. Each image block represents a fixed part of the target region, and gathering them can represent the whole target. The diagram of this kind of structural representation is showed in figure 2.

![Figure 2. The diagram of structural representation](image2)
Let $I(x, y)$ represents the current image frame, $F(x, y) \in I$ represents the candidate target image, matrix $T$ is used to divide the target image $F$ into $m \times n$ local image blocks. 

$$T = \begin{bmatrix} T_{11} & \cdots & T_{1n} \\ \vdots & \ddots & \vdots \\ T_{m1} & \cdots & T_{mn} \end{bmatrix}$$

(3)

Where, each $T_{ij}$ represents a fixed part of the target image, the size is $r \times s$.

$$T_{ij} = \begin{bmatrix} t_{i1} & \cdots & t_{is} \\ \vdots & \ddots & \vdots \\ t_{r1} & \cdots & t_{rs} \end{bmatrix}$$

(4)

Where, $t_{ik}$ represents the coordinates of pixels in candidate target image $F$.

$$t_{ik} = ((i-1)r+l, (j-1)s+k)$$

(5)

Where, $l = 1, 2, \ldots, r, k = 1, 2, \ldots, s$.

Then, the target can be represent as

$$v = TF = [v_1, v_2, \ldots, v_{m \times n}]$$

(6)

The entropy is calculated in each $v_i$, the final structural entropy appearance model is

$$h = [h_1, h_2, \ldots, h_{m \times n}]^T$$

(7)

Where, $h_i$ is the entropy of each image block.

### 3. THE PROPOSED TRACKING METHOD

#### 3.1. Constructing and updating classifier

Naive Bayesian classifier divides the unknown samples into two classes, that is, positive and negative. It is a studying method with supervision, which assume that the given value of one character is independent with others. The posterior probability of positive sample $x$ is

$$p(y=1|x) = \frac{p(x|y=1)p(y=1)}{\sum_{x=1}^n p(x|y=1)p(y=1)} = \sigma \left( \ln \left( \frac{p(x|y=1)p(y=1)}{p(x|y=0)p(y=0)} \right) \right)$$

(8)

Where, $\sigma(z) = 1/(1+e^{-z})$ is a sigama function, $y \in \{0,1\}$ is the binary label of sample $x$.

The definition of the classifier used in the paper is:

$$H_k(x) = \ln \left( \frac{p(x|y=1)p(y=1)}{p(x|y=0)p(y=0)} \right)$$

(9)

The posterior probability of equation(8) can be represented as

$$p(y=1|x) = \sigma \left( H_k(x) \right)$$

(10)

Where, sample $x$ is represented as a feature vector, that is,

$$f(x) = [f_1(x), \ldots, f_K(x)]^T$$

(11)

The elements $f_i(x)$ in $f(x)$ are distributed independently and the prior probability $p(y=0) = p(y=1)$. The classifier $H_k(\cdot)$ in equation (9) can be described with $f(x)$ as

$$H_k(x) = \ln \left( \frac{p(f(x)|y=1)p(y=1)}{p(f(x)|y=0)p(y=0)} \right) = \sum_{k=1}^K h_k(x)$$

(12)

Where,

$$h_k(x) = \ln \left( \frac{p(f_k(x)|y=1)}{p(f_k(x)|y=0)} \right)$$

(13)
This is a discrimination function, the conditional distribution in which meet Gaussian function, that is
\[
p(f_k|y=1) \sim N(\mu_k^1, \sigma_k^1)
\]
\[
p(f_k|y=0) \sim N(\mu_k^0, \sigma_k^0)
\]
(14)

When updating the classifier, the studying parameters are added to update the mean and variance \((\mu_k^1, \sigma_k^1), (\mu_k^0, \sigma_k^0)\) of positive and negative samples.
\[
\mu_k^i \leftarrow \lambda \mu_k^i + (1 - \lambda) \mu^i
\]
\[
\sigma_k^i \leftarrow \sqrt{\lambda (\sigma_k^i)^2 + (1 - \lambda) (\sigma^i)^2 + \lambda (1 - \lambda)(\mu_k^i - \mu^i)^2}
\]
(15)

Where, \(\lambda > 0\) is the studying parameter, \(\sigma^1 = \left[ \frac{1}{n} \sum_{i=1}^{n-1} (f_k(i) - \mu^i)^2 \right], \mu^1 = \frac{1}{n} \sum_{i=1}^{n-1} f_k(i)\).

The variables in the classifier are independent, the \(n\) dimensional problem of multi-variant is reduced to the estimation of \(n\) single variants. Less training samples are needed to gain the accurate position of the target comparing with estimating the covariance matrix. Even there are some drifts in tracking results, the robust estimation still can be gain when the positive samples are extracted surrounding the tracking result in the current frame to update the distribution parameters. The previous accurate information is used to update the distribution parameters as well, which enhance the robustness of the method to those deviant samples.

3.2. The two-phase tracking method

The combination of the structural entropy appearance model and the naïve Bayesian classifier is applied to two-phase tracking method. The new target position is confirmed using search method from coarse search to fine search. In the coarse search phase, some test samples are obtained surrounding the last target location using a large search radius \(\gamma_c\) and a search step length \(\Delta_c\). The coarse target location is obtained by calculating the scores of the classifier. In the fine search phase, some new test samples are obtained surrounding the coarse target location using a smaller search radius \(\gamma_f < \gamma_c\) and search step length \(\Delta_f < \Delta_c\). The final target location is confirmed by computing the score of the classifier (the procedure is shown in figure 3). The size of the target maybe change in the moving procedure, an appropriate scale factor is added to the width and height of the rectangle area, that is, the affine transform parameters.

![Figure 3. Coarse-fine search method for new object location](image)

The red point in the left image in figure 3 represents the center location of the target. The middle image shows the coarse search using a large radius and step length. The right image shows the fine search using a smaller search radius and step length. The blue point represents the center of the final target location.

The proposed structural entropy appearance model combined with naïve Bayesian classifier is used to track target. In the initial phase of the tracking, the target location is labeled manually in the first frame. A number of positive and negative samples are extracted surrounding the previous target center with a radius. Structural entropy feature vectors are extracted according the equation (3-7), and a batch of sample set \(\{x_j, y_j\}_{j=1}^y\) are gained. These initial samples are used to train the naïve Bayesian classifier. In the following frames, selecting enough samples using coarse search method surrounding the previous target location, calculating scores of each sample using classifier, and selecting the samples with the biggest score as the primary estimated location. In the base of the coarse estimation, searching the samples using fine search surrounding the primary location, computing the score of new samples, and selecting the biggest one as the final target location. Finally, selecting enough positive and negative samples surrounding the location to update the classifier using equation (15). The whole method described as

Input: visual sequence \(\{y_t\}_{t=1}^T\)
(1) Labeling the rectangle \( R_t \) in the first frame manually.

(2) Getting the positive and negative samples \( S_t = \{x_i, y_i\}_{i=1}^{W} \).

(3) Training the naïve Bayesian classifier using the samples \( H_t(x) = \sum_{k=1}^{K} h_k(x) \).

(4) For the following frames \( I_t, t > 2 \):
   1. Getting new candidate samples using coarse searching.
   2. Getting new target region \( R_t \) with the highest score using classifier \( h_k(x) \) of the previous frame.
   3. Confirming the final target location by computing score of the new candidate samples gained from fine search.
   4. Getting new positive samples \( P_t = V_{t-1}^+ \cup S_t^{+} \) and negative samples \( N_t = V_{t-1}^- \cup S_t^{-} \).
   5. Training classifier with new training set \( h_k(x) = \ln \left( \frac{\rho(f_k^t(x) | y = 1)}{\rho(f_k^t(x) | y = 0)} \right) \).

Output: location of target in each frame.

4. EXPERIMENTAL RESULTS AND ANALYSIS

Experiments are implemented in MATLAB (R2010b). For each video sequence, the target object is manually labeled in the first frame. Implemented on 4GB RAM, Intel(R) Core(TM) i5-2400 3.10 GHz CPU and no optimization. The radius \( \alpha \) of the positive samples is 4, the number is 45. The inner and outer radius of the negative samples are \( \zeta = 8, \beta = 38 \) respectively, the number is 50. In the coarse search phase, \( \gamma_c = 25, \Delta_c = 4 \), the number of test sample is 121. In the fine search phase, \( \gamma_f = 10, \Delta_f = 1 \), the number of test sample is 305, and the studying parameter of the classifier \( \lambda > 0.85 \). The comparison among the one-phase tracking method (PLS1), two-phase tracking method based on subspace model (PLS) and the proposed method are made. The comparison with PLS is used to demonstrate the robustness and feasibility of the proposed method, and the comparison with one-phase tracking method is used to demonstrate the effectiveness of the proposed method. Each candidate image is normalize to the size of 32 \( \times \) 32, and local patches with size of 16 \( \times \) 16 are extracted overlapped. The quantity and quality analysis about the tracking results are made in the paper. Partial tracking results are showed in Figure 4-7, there are illumination change, pose change, partial occlusion and rotation in the visual sequences. The central position difference is used to evaluate the effectiveness of the tracking method, the form of which is \( \sqrt{(x' - x_0)^2 + (y' - y_0)^2} \) and \((x_0, y_0)\) are the central position of the tracking result and the reference of the target in the current frame, respectively. The central position difference curves of the sequences are showed in Figure 8-11 (abscissa is the number of frames, ordinate is the center position difference (pixels)), and Table 1 shows the average center position difference.

<table>
<thead>
<tr>
<th>Table 1 Average center position difference (pixels)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Coke</td>
</tr>
<tr>
<td>car4</td>
</tr>
<tr>
<td>faceocc2</td>
</tr>
<tr>
<td>David-indoor</td>
</tr>
</tbody>
</table>

Illumination and pose change: In the Coke, car4 and David-indoor sequence shown in figure 4-5 and figure 7, there is significant illumination and pose changes. Although our method lost the target when tracking the David-indoor sequence and Coke sequence, the whole tracking effect is good. The other two methods both lost the target in the Coke sequence specially. In car4 sequence, PLS1 lost the target after 200th frame shown in figure 9, the drift took place in the 291th frame shown in figure 5. While our method tracked the target well. The reason is that the entropy feature which is insensitive to illumination and pose change.

Rotation and partial occlusion: PLS1 is the single phase tracking algorithm. By observing figure 8-11, we can find that the two-phase tracking algorithm is better in tracking effect and more stable than the single phase tracking algorithm. Because two-phase tracking method add revise and improve the accuracy, while the single-phase only learns an appearance model online, thus it is easier to lose the object. Especially when the object rotates and occlude partially (Coke sequence, as shown in figure 4). Ours can track the target well even
though there is partial occlusion (faceocc2 sequence, as shown in figure 6). As can be seen from figure 8-11, the whole tracking effect of our method is well.

The structural entropy appearance model adopted in this paper makes tracking more stable. Even if the object is influenced by significant illumination and pose changes, it can still track the object steadily because of the entropy feature which is insensitive to these factors. Structural representation can deal with partial occlusion in a certain extent. Table 1 shows that the proposed method track well than other two methods.
5. CONCLUSION

Structural entropy appearance model and the naive Bayesian classifier are combined to track the object using two-phase search method. The entropy feature represents texture feature which is insensitive to illumination and pose change. Blocking is used to deal with partial occlusion. The tracking is carried out within the Bayesian inference frame with two-phase searching. The tracking results of different video sequences show that the proposed method is robust to illumination, pose change and partial occlusion.

ACKNOWLEDGEMENT

The paper is funded by the project of “Natural Science Foundation Guidance Plan of Liaoning Province”
REFERENCES


