Digital Signal Processing 79 (2018) 102–115

Contents lists available at ScienceDirect

Digital Signal Processing

www.elsevier.com/locate/dsp

# Visual object tracking based on sequential learning of SVM parameter

Vijay K. Sharma [a,*], K.K. Mahapatra [b]

[a] Department of Electronics and Telecommunication Engineering, National Institute of Technology, Raipur, C.G., 492010, India
[b] Department of Electronics and Communication Engineering, National Institute of Technology, Rourkela, Odisha, 769008, India

ARTICLE INFO

*Article history:*
Available online xxxx

*Keywords:*
Computer vision
Visual tracking
Sequential learning of SVM parameter
Subspace
Sparse DCT

ABSTRACT

In this paper, a training algorithm is proposed to online (sequentially) learn the SVM based parameter. The proposed online learning method is used to construct discriminative classifier for visual object tracking application, where new examples are available in each successive frame of a video. The iterative method is based on maximizing the magnitude of sum of projection values of a positive example and a negative example which are closest to the hyperplane formed by parameter to be updated. In the proposed training framework, even if there is non-accurate labeling of the training examples (which are received online), it is possible to learn the parameter that ensures maximum value from the classification function. The learned parameter along with some examples which are nearest to hyperplane are used for the construction of object likelihood model. Using likelihood model, tracking instance most similar to the target is selected, where the target candidates are generated using particle filter framework. An object representation is also learned based on sparse DCT coefficients. This representation contains the basic structure of the target appearance. The proposed object tracking method performs better than state-of-the-art trackers in a number of challenging video sequences. When high dimension feature vectors are used, instead of simple raw pixels based feature vectors, to represent the training examples, the performance of the object tracking is better in a number of video sequences even without integrating sparse DCT coefficient based object representation as an additional model.



## 1. Introduction

Tracking a generic object in a video, called visual object tracking, is a very important and challenging field in computer vision. The objective of visual object tracking is to correctly estimate the position and size of a target object in the successive video frames. There are several applications that require tracking information. These applications include visual surveillance, human–computer interaction (HCI), traffic monitoring, video compression [1–4] and automatic analysis of sports video [5,6]. The object tracking is also used in medical applications and biological research [7–9].

The support vector machine (SVM) is a supervised learning method that, given a set of training data, learns a optimal hyperplane to separate one class of data from the other [10]. The SVM has been used for pattern classification in several different applications [11–16]. The optimal hyperplane is the one which separates the data of two different classes with maximum distance of separation or margin, where the margin is 2-times the distance between hyperplane and a data point (on either side of the hyperplane) closest to the hyperplane. Fig. 1 shows a hyperplane in a 2-dimensional space with margin and support patterns [10].

Since the SVM training is performed off line (all examples should be available prior to training), it is not possible to use the SVM classifier in the applications that require on line learning. In case of object tracking, the on line learning is a must as new examples are received in each frame of a video sequence. The tracked samples are positive examples while the negative examples should be obtained from each video frame for the construction of efficient classifier. The classifier parameter learned using on line learning method contains the temporal variations in the example.

The discrete cosine transform (DCT) has excellent decorrelation property [17,18]. The DCT coefficients are uncorrelated as opposed to raw pixels of an image, wherein there is presence of low to significantly high pixel correlations. The coefficients represent the well defined frequency content in the image. Low frequency represents the overall structure of an image. An image can be reconstructed with the help of only some low frequency coefficients, although there is quality degradation due to loss of high frequency

* Corresponding author.
*E-mail addresses:* vijay4247@gmail.com (V.K. Sharma), kkm@nitrkl.ac.in (K.K. Mahapatra).

https://doi.org/10.1016/j.dsp.2018.04.008
1051-2004/

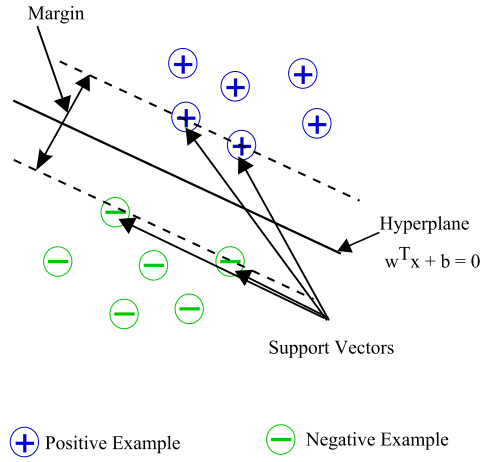


**Fig. 1.** Hyperplane in 2-dimensional space, showing support vectors.

information. By using a proper method that discards some of the frequency contents, an appearance model can be learned that preserves the structure of an on line appearance of the target.

In this paper, SVM parameter vector is sequentially (online) learned to construct a discriminative classifier for object tracking. The main idea is to approximate the learned parameter in an iterative way so as to maximize the magnitude of projection values of sum of two vectors, one from positive example set and another from negative example set, which are closest to the hyperplane to be learned. In contrast to conventional method, the proposed framework of discriminative parameter learning takes into account the inaccuracy of the tracked object in some of the video frames. The learned SVM parameter along with some examples which are nearest to hyperplane are used for object likelihood model construction. Eigenbasis vector based subspace appearance model is also utilized for likelihood construction. Three intermediate tracking instances are obtained, two using proposed discriminative models and one using subspace reconstruction. Then for choosing the best target among them (three intermediate tracked instances), their correlations are compared with raw pixels based appearance models. For deciding the final tracking instance, the instance obtained using learned parameter model is given higher weight than the instances obtained using other two models. In other words, the other two models work as helping models in case the tracked target using learned parameter vector is not highly accurate. Three types of appearances are used based on raw pixels. One of them learns the basic target structure which is constructed using sparse DCT coefficients. The tracking performance of the proposed method is better as compared to the available state-of-the-art trackers in a number of challenging video sequences.

The remaining section of the paper is as follows. Related works are discussed in section 2. Section 3 briefs the construction of SVM classifier. Section 4 presents our proposed SVM parameter learning method. The proposed tracking method along with sparse DCT based appearance model is described in section 5. Experimental results are given in section 6. We conclude the paper in section 7.

## 2. Related works

### 2.1. Appearance model in visual object tracking

The object tracking has been an active research area due to several challenges involved. One of the most encountered challenges is the object appearance variation which occurs from one frame to another in the form of deformation, motion blur, pose change, and illumination variation. Scale change, partial occlusion and clutter are the other challenges involved in visual object tracking. The discriminative and generative modeling are the two statistical learning techniques to online learn a mathematical model of the target. The generative model does not contain the background information [19–22]. In the discriminative appearance technique, a classifier is learned to separate the object from the background [23–25].

In generative method for visual object tracking, a structure most similar to the target is searched in the neighborhood of the current object in an image [19,20,26–28]. Amongst generative models, sparse representation based appearance for visual tracking is a recent approach [22,29–37]. A target candidate is sparsely represented by a target templates set and a trivial templates set (representing occlusion). Regularized L-1 norm minimization formulation is used to obtain sparse coefficients. Wang et al. in [37] proposed a fast solution to L-1 regularization, wherein the target templates set is eigenbasis vectors obtained from incremental principal component analysis (PCA) [20]. The eigenbasis vectors (a set of linearly independent vectors) represent a subspace, where the dimension of the subspace is equal to the number of basis vectors used. In [38], decision-theoretic learning based adaptive NormalHedge algorithm is used. This algorithm uses an adaptive mechanism to determine the amount of historic information to be used for target state estimation.

Raw pixel intensity (or color) is a simple visual feature. Spatially weighted color histogram based appearance is constructed by weighting each pixel in the target region. Pixels at the center are given higher weight as compared to pixels at the boundary of the target [19,39]. Intensity gradient and color histogram based visual features are used in [40]. In [41], spatial-color mixture of Gaussian (SMOG) based appearance model is used.

To increase the robustness, highly complex feature descriptors such as, scale invariant feature transform (SIFT) [42], speeded-up robust features (SURF) [43], histogram of oriented gradient (HOG) [44], and local binary pattern (LBP) [45] can be used. SIFT and HOG represent gradient based features while LBP represents texture based features. SIFT based feature for object tracking is used in [46–48]. SURF based feature has been used in [49,50], while in [51], LBP has been used for visual tracking. Gabor wavelet transform (GWT) is another way to extract robust feature to model an object. In [52], amplitudes of GWT coefficients are used for local feature as Gabor wavelets are localized in both time and frequency domains. Multiple objects can be tracked from the video using data association scheme [53,54]. Calibrated binocular camera is used for tracking in [55], wherein it is argued that binocular geometry constraints and information fusing from two channels can deal with target undergoing occlusion, scale variation and out-of-view.

The DCT is a class of orthogonal transform which decorrelates the data using a set of basis vectors. Since its introduction in [17], the DCT has been used in many signal and image processing applications due to its excellent decorrelation property. The role of DCT in image processing, especially in image compression [18,56–60] is substantial. This is due to the fact that all natural images contain a great deal of redundant pixels. In the transform domain, we have flexibility to discard (for compression) or scale (for enhancement) some of the coefficients due to spectral separation [61]. There is no significant change in the overall appearance of the reconstructed image after removing some of the coefficients representing high frequency. In [62], feature vectors are constructed by selecting some of the DCT coefficients for face recognition system.

Discriminative tracker treats the object tracking as a binary classification problem. Zeng et al. in [63] proposed a tracking algorithm consisting of global and local parts. The global part is a discriminative model based on holistic features. Kernel correlations based tracker which exploits the fast Fourier transform (FFT) and inverse FFT to compute the computationally complex convolutions