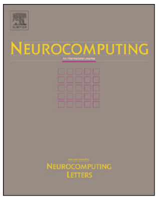
Neurocomputing 136 (2014) 327–336

Contents lists available at ScienceDirect

Neurocomputing

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



Visual tracking with structural appearance model based on extended incremental non-negative matrix factorization

Cheng Qian*, Yanbin Zhuang, Zezhong Xu

Department of Computer Science and Technology, Changzhou Institute of Technology, China

A R T I C L E I N F O

Article history:
Received 30 May 2013
Received in revised form
4 December 2013
Accepted 28 December 2013
Communicated by X. Li
Available online 10 January 2014

Keywords:
Visual tracking
Incremental non-negative matrix factorization
Coding coefficient vector
Naïve Bayes classifier

A B S T R A C T

In this paper, we propose an appearance model based on extended incremental non-negative matrix factorization for visual tracking. With non-negative matrix factorization, each object image patch identified in a frame is regarded as a linear combination of a set of non-negative basis vectors. Then a feature space depicting the appearance of the object is spanned by the set of basis vectors. An encoding matrix encodes the image patches with the coding coefficient vectors. The discrimination of the coding coefficients is investigated. In order to search the image region for the object in a new frame, a group of image patches are sampled around the position of the object in the previous frame. A naïve Bayes classifier evaluates the confidence of each image patch as the candidate for the object image patch. Finally, the image patch with the maximal confidence is taken as the image region of the object. After identification of the object image patch, the mixing matrix and the classifier are updated incrementally. Experimental results show that our algorithm outperforms several state-of-the-art algorithms.

1. Introduction

As a fundamental part of computer vision, visual tracking contributes to behavior understanding, video surveillance and anomaly detection. Currently, a variety of algorithms try to give solutions to object tracking. But, due to camera motion, illumination changes and occlusions, the object is usually subject to variations in appearance, which directly results in difficulty in tracking object under arbitrary circumstance. It has reached a consensus that the performance of tracking relies on the appearance model to a large extent [1,2], and modeling appearance to make the object salient will prompt tracking [3,4].

In an image, an appearance model usually obtains a semantic understanding of scenarios based on low-level features. In order to meet demands of various applications, the appearance model should take different methods for feature extraction and feature refinement into consideration. Numerous appearance models are biased in favor of pixel-wise features as fine-grain descriptions for image information, and the structure of the object is often ignored [5–7]. This is adverse to the organization of the features into a semantic description.

For a sequence of video in short time, image regions for a special object in the frames exhibit time coherence and spatial coherence. While the time coherence gives the continuity of movement, the spatial coherence shows the consistent structural composition of an object in the consecutive frames. This makes it possible to seek compressive and intuitive descriptions for the object. Several methods for finding a low-rank approximation to the high-dimensional image data, like singular value decomposition (SVD) [8], manifold learning [9,10], compress sensing [11,12] and so on, have demonstrated their ability to uncover the implicit structure of the data. Among them, non-negative matrix factorization (NMF) enforces an addition rule to matrix decomposition [13,14] and gives a part-based representation for images. But NMF usually requires much memory to store data. Besides storage, the large scale data also causes much computational effort to obtain the basis vectors spanning a subspace. Recently, incremental non-negative matrix factorization (INMF) is proposed as an online subspace learning method [15–17]. Instead of processing data in batch mode, INMF constructs a subspace in an incremental manner. Especially, it is efficient for dealing with the data stream like video sequence.

In this paper, we propose a novel algorithm for visual tracking. INMF is utilized to create a dictionary updated online. Through solving a non-negative least squares problem, an image patch is encoded with a low-dimensional coding coefficient vector that acts as the feature. Tracking result is derived from the evaluation of likelihoods over the features. The methods presented in [17,18] also concentrate on the usage of NMF in visual tracking. In [17], Wang et al. propose an approach of incremental orthogonal projective non-negative matrix factorization, and then integrate this approach into sparse representation framework so as to get a

[*] Corresponding author.
E-mail address: qcpaper@163.com (C. Qian).

0925-2312/$ - see front matter © 2014 Elsevier B.V. All rights reserved.
http://dx.doi.org/10.1016/j.neucom.2013.12.025

part-based appearance model for tracking. Wu et al. [18] add sparsity constraint and smooth constrain into NMF. They try to find a subspace with locality preservation to account for the variations in the appearance. But our work is different from these two works in original inspiration and the usage of NMF. The former methods place emphasis on the development of the technique of subspace learning. The establishment of the subspace does not take the background into consideration. In nature, their tracking follows the same idea of subspace learning-based tracking [19]. Compared with them, NMF serves as an approach of feature extraction in our method. A classifier is trained to distinguish the object from the background. This classifier absorbs the information from the background, which improves the robustness of the tracking.

The reminder of our paper is organized as follows: Section 2 reviews the related work on visual tracking. Section 3 proposes an appearance model based on a non-negative dictionary, and interprets overall framework for object tracking in detail. Section 4 presents the experimental results and gives analysis to the results. Section 5 concludes this paper.

2. Relate work

Appearance of an object usually varies with different environments. The appearance model designed for a fixed scenario is only confined to some special tracking tasks [7], and they are difficult to be extended to other application contexts. The tendency towards adaptive appearance models gradually dominates the development of tracking algorithms [20].

During tracking, features are picked in each frame, and the adaptive appearance models are renewed with these features. The features, such as histogram [6], Haar feature [21], SIFT [22], HOG [23] and so on, are often good at capturing characteristics of the appearances. Corresponding feature spaces are established in company with these features.

For a generative appearance model, the features representing a specific object are considered as a cluster in the feature space. The center of the cluster is usually taken as the template of the object. The object region is identified through matching candidates with the template [24–26]. Thanks to the emergence of efficient optimizations, the generative appearance models can accomplish a fast matching. Among them, mean-shift tracking [6] uses a weighted color histogram to depict the target. The object image region is found as the peak of a confidence map through mean shift. David et al. [19] devise a linear subspace to depict the image patches from an object. Incremental principal component analysis is utilized to incorporate the variations of the appearance into the low-dimensional subspace. As an extension to the technique of incremental subspace learning, incremental tensor subspace learning [27] is developed to characterize the intrinsic spatio-temporal structures of the object. The object is identified with the distance measurement in the tensor subspace. Sparse representation-based tracking [28–30] creates an over-complete dictionary for the appearance. L1 minimization is used as a tool to measure the similarity between the templates and the candidate. To eliminate the distracters occurring in the tracking, Hong et al. [31] propose a distance metric learning approach to find a subspace, where the distance between the object and the distracters is enlarged under the framework of maximum margin. The object is located through sparse representation in the subspace spanned by the projection vectors. Recently, a least soft-threshold squares' algorithm is exploited to model the reconstruction error with Gaussian noise vector and Laplacian noise vector [32]. Then the tracking problem is transformed into a measurement of the least soft-threshold squares distance. These methods all try to build up templates for matching. But the distance between two feature points in the feature space cannot entirely reflect the similarity between two image regions that they represent. Once the accuracy of the template degrades, the drift of tracking will be amplified.

In contrast to the generative appearance model, the discriminative appearance model not only models the object but also models the background. To make the distinction between object and background, a maximal margin is sought in the feature space [33–36]. Then the tracking is transformed to a binary classification between background and object. After the classifier is trained, the evaluation of the likelihood will consume litter time. This gives the tracker a chance to detect the target in the whole image. As a typical classifier, support vector machine (SVM) gets wide applications in numerous vision tasks. To accomplish visual tracking, Avidan [37] introduces SVM into the classification of image patches from the video clips. After that, Avidan [33] tries to use a boosting method for tracking. Grabern et al. [38] present a feature selection method based on online boosting, and this selection can effectively exploit the features in favor of tracking. In order to overcome the ambiguities in labels of training samples, multiple instance learning is employed to assign the labels to the bags of samples [39]. Following this strategy, the classifier trained becomes more robust to outliers. Sam et al. [35] use a kernelized structured output SVM to provide an adaptive tracking. Xi et al. [40] propose a hashing function to encode the information from multiple cues with a binary code. With the help of LS-SVM learned on the binary codes, a hypergraph propagation method gives the final tracking result.

The tracking approaches with the discriminative appearance models usually employ a static window to scan the image in a large scope. Immersion of pixels from the background into the window of the object region leads to outliers occurring in the training set. This contaminates the cluster of feature points for object samples and blurs the decision boundary. Since the introduction of outliers to the training set is inevitable, it is necessary to bring the decision boundary to prominence.

To impose some structural constraints on feature refinement will enhance robustness of the appearance model remarkably. In contrast to other approaches, our approach focuses on the integration of structure into features extraction. This is helpful for the appearance model to resist against false object, and it improves the robustness of the tracker.

3. Non-negative matrix factorization for tracking

For tracking-by-learning framework, a tracker determines the position of a target from one frame to the next frame, and the online learning improves the adaption of the tracker to changes in appearance. Our method also complies with this framework. We use NMF to generate a non-negative dictionary, which checks "ingredients" of image patches to validate the object image regions. After tracking, the dictionary is updated through INMF.

3.1. Prediction of location

The spatial location of an object in a frame can be represented with a vector of hidden variables $\mathbf{x} = [x, y, w, h, \alpha, \theta]$, where x, y, w, h, α, θ denote coordinates of center, width, height, scale and aspect ratio, respectively. Under the control of the hidden variables, the motion of an object can be viewed as affine transformation of an image region surrounding the object.

Once the target is identified in the current frame, the corresponding hidden variable vector is got. Considering the temporal smoothness, slight perturbations are carried out on the hidden variable vector so as to simulate the motion of an object, and then