Neurocomputing 225 (2017) 139–147

Contents lists available at ScienceDirect

Neurocomputing

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

Human activity prediction using temporally-weighted generalized time warping

Haoran Wang[a,*], Wankou Yang[b], Chunfeng Yuan[c], Haibin Ling[d], Weiming Hu[c]

[a] College of Information Science and Engineering, Northeastern University, Shenyang 110819, China
[b] School of Automation, Southeast University, Nanjing 210096, China
[c] National Laboratory of Pattern Recognition, Institute of Automation, Chinese Academy of Sciences, Beijing 100190, China
[d] Department of Computer and Information Sciences, Temple University, Philadelphia, PA 19122, USA

ARTICLE INFO

Communicated by Dr. Z. Wang

Keywords:
Activity prediction
Time warping
Alignment

ABSTRACT

Different from traditional human activity recognition, human activity prediction aims to recognize an unfinished activity, typically in absence of explicit temporal progress status. In this paper, we propose a new human activity prediction approach by extending the recently proposed generalized time warping (GTW) [20], which allows an efficient and flexible alignment of two or more multi-dimensional time series. More specifically, for each activity video, either complete or incomplete, we first decompose it into a sequence of short video segments. Then, we represent each segment by the local spatial-temporal statistics using the classical bag-of-visual-words model. In this way, the comparison between a query sequence (i.e., containing an incomplete activity) and a reference sequence (i.e., containing a full activity) boils down to the problem of aligning their corresponding segment sequences. While GTW treats different portions of a sequence as equally important, our task is in favor of early portions since an incomplete activity video always aligns from the beginning of a complete one. Thus motivated, we develop a temporally-weighted GTW (TGTW) algorithm for the activity prediction problem by encouraging alignment in the early portion of an activity sequence. Finally, the similarity derived from TGTW is combined with the k-nearest neighbors algorithm for predicting the activity class of an input sequence. The proposed approach is evaluated on several publicly available datasets in comparison with state-of-the-art approaches. The experimental results and analysis clearly demonstrate the effectiveness of the proposed approach.

1. Introduction

Recognizing human activities from videos has attracted an increasing amount of research interest recently. It typically requires to capture enough spatial and temporal information to distinguish different activity classes, while handling the large intra-class variations. Recent surveys can be found in [1–4].

Most of the existing methods usually focus on recognitions of complete activity videos [7,8,15–17]. However, in many real-world scenarios, the system is required to identify intended human activities before they are fully executed. For example, in a surveillance scenario, recognizing the fact that certain objects are missing after they have been stolen may not be meaningful. The system could be more useful if it is able to prevent the theft and catch the thieves by predicting the ongoing stealing activity as early as possible based on live video observations. In the sports video analysis, the capability of predicting the progress or results of a sport game will be highly desirable. In public area, we want to equip a surveillance system that can raise an alarm in advance before any potential dangerous activity happens. In a smart room, people's intention of activity can be predicted by a user-friendly sensor-camera, so that the system will adaptively provide services, even help if necessary.

The intuitive approach is to extend traditional sequential models such as *hidden Markov models* (HMM) to roughly approximate the prediction problem, but these models often meet the problem about how to extract high-dimensional features to provide an effective video representation. The grammar based method [14] shows the effectiveness in prediction of complex activities, but semantic representation requires high resolution videos, and is easily influenced by noises in realistic videos. A popular strategy [6,9,12,13] is to divide a video into a set of consecutive segments, and measure the similarity between corresponding segments in different videos. The correspondence is,

* Corresponding author.
E-mail addresses: wanghaoran@ise.neu.edu.cn (H. Wang), youngwankou@yeah.net (W. Yang), cfyuan@nlpr.ia.ac.cn (C. Yuan), hbling@temple.edu (H. Ling), wmhu@nlpr.ia.ac.cn (W. Hu).

http://dx.doi.org/10.1016/j.neucom.2016.11.004
Received 28 March 2016; Received in revised form 31 August 2016; Accepted 9 November 2016
Available online 15 November 2016
0925-2312/

however, nontrivial to obtain. For an unfinished test video, we usually have no information about the status of its temporal progress, so we can not decide the number of segments which the unfinished test video and the complete training video are separately divided into for the similarity measurement of corresponding segments. So far, it is still an unsolved problem to automatically infer the progress status of an incomplete activity, and this is the essential difference between activity prediction and traditional activity recognition.

In this paper, we propose an activity prediction method based on the alignment of time series. Having noticed that activities from the same class are usually performed in similar action evolution processes, we represent a video sequence by dividing it into a series of short segments. The number of segments thus correlates to the progress status of the activity the sequence carries. To capture the dynamic appearance of each segment, the distribution of the spatial-temporal interest points [19] are summarized in the bag-of-visual-words fashion. In this way, an activity video, either incomplete or complete, is represented by a time series of visual word histograms. In order to compare such time series of different lengths, it is natural to use the time warping algorithms such as the recently proposed *generalized time warping* (GTW) [20] algorithm. The original GTW algorithm, however, treats equally important each video segment, while in action prediction, matching happens more often in the early portion than in the latter portion of the activity. Addressing this issue, we design two different *temporally-weighted GTW* (TGTW) based algorithms to align an unfinished activity video with a full activity video in favor of its early segments. More specifically, one approach intuitively modifies the GTW objective function by a diagonal matrix to weight early portions of activities less than the latter portions. The other approach directly constrains the warping path which essentially determines the alignment position in the complete activity video. Finally, the similarity derived from the TGTW based alignment is combined with the k-nearest neighbors algorithm to predict the activity class. Fig. 1 illustrates the flowchart of our framework.

We evaluate the proposed TGTW-based activity prediction algorithm on three public benchmark datasets including UT-Interaction [36], DARPA-Y1 [38], and UCF Sports datasets [37]. Our method achieves very promising results in all the experiments in comparison with several state-of-the-art solutions.

The remainder of this paper is organized as follows. Section 2 gives a review of related work. Section 3 introduces the proposed approach. Section 4 demonstrates the experimental results. Section 5 concludes this paper.

2. Related work

This section reviews previous work on activity prediction and time series alignment.

2.1. Activity prediction

Human activity prediction is an important and challenging problem, and it is a relatively new topic in computer vision with several notable recent studies. Ryoo [6] represents an activity with an integral or dynamic bag-of-words model to probabilistically formulate the activity prediction problem, and proposes a prediction algorithm to consider the sequential structure formed by video features. Cao et al. [9] recognize human activities from partially observed videos. They take a set of completely observed training video samples of each activity class as the bases, and then use sparse coding to derive the likelihood that a certain type of activity is presented in a partially observed test video. Kong et al. [12,13] extend the support vector machine, and build multiple temporal scale templates to recognize activities with different progress status. These methods all manually divide different videos into the same number of segments, and use the segment index to reflect the temporal status. But for the unfinished activity, it is hard to know the temporal status and the number of segments it should be divided into. In this paper, we also divide the videos into several consecutive segments, but we don't impose the restriction on the same number of segments for different videos. Most importantly, our method automatically searches for the corresponding part of the unfinished activity from the complete activity video.

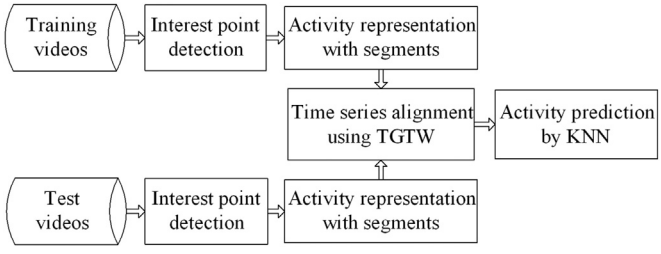


Fig. 1. Flowchart of the proposed framework.

Also, there are some enlightening works which are closely related to the problem of human activity prediction. Xu et al. [10] regularize different activity videos to the same length by upsampling or down sampling, and extract discriminative patches to auto-complete partial videos. Kitani et al. [11] use semantic scene understanding method to predict plausible paths and destinations of pedestrian. Li et al. [14] intend to discover the causal relationships between constituent actions and the predictable characteristics of activities, and decompose an activity into a set of atomic actions in a syntactic way. But the proposed high-level features are only tested on two noise-free datasets. Walker et al. [18] propose an unsupervised method to forecast the possible change of scene with time. Hoai and De la Torre [5] propose max-margin early event detectors to localize the starting and ending frames of an activity.

2.2. Time series alignment in related problems

Aligning sequences of entities is one of the fundamental problems in computer vision, consequently, time series alignment algorithms have been widely explored in computer vision tasks such as curve matching [22], shape matching [23], activity recognition [24], and synthesis of human motion [25]. Despite a large body of literature, it is still a challenging problem to finding correspondences between human activities with distinct styles under large environment change. To account for the variations of the same activity performed by different subjects, Hsu et al. [26] propose to combine *dynamic time warping* (DTW) [21] with a space warping step to transform the style of an action into a new one while preserving original content. Heloir et al. [27] propose a multi-level DTW algorithm to deal with the communicative gestural sequence alignment problem by removing the stylistic part in gesture sequences. Singh et al. [33] establish sub-frame synchronization for video sequences which are acquired via uncalibrated cameras using a bi-directional or symmetrical alignment. Although above methods achieve promising results, there are still limitations in deal with data with multiple modalities and dimensions. Zhou and De la Torre [34] propose the canonical time warping, by extending the algorithms of CCA and DTW, for the spatial-temporal alignment of two multivariate time series, and apply it to align human activity videos of two subjects. Furthermore, they present the GTW algorithm [20] to efficiently align semantically similar multi-modal sequences, using multi-set canonical correlation analysis to find the spatial transformations and considering the temporal warping as a combination of multiple monotonic bases.

View-invariant representation is an important technique in video alignment. Zhou and De la Torre [20] use multi-set canonical correlation analysis to adapt series alignment for handling view-variability in activity videos. Caspi and Irani [28] temporally align videos captured from different cameras with no overlapped fields of view, while the

cameras are attached closely to each other and move jointly in space. Padua et al. [30] use homography-based constraints to align sequences under different viewpoints. They reduce the alignment of N unsynchronized videos into the estimation of a single line, which contains the relations between different sequences without any prior knowledge. Li and Chellappa [31] introduce an efficient method for spatial-temporal alignment by designing a sequential importance sampling algorithm on a Riemannian manifold. Rao et al. [32] propose a rank constraint based similarity measurement to align human activities from different viewpoints. Besides time series alignment, view-invariant representation is also a pervasive problem in other computer vision areas. Junejo et al. [24] present a view-invariant self-similarity descriptor for actions by making use of the affinity matrix between time instances. This descriptor captures the structure of temporal similarities and dissimilarities within an action sequence, and requires neither structure recovery nor multi-view correspondence estimation. Gritai et al. [29] utilize the trajectory matching score and the projective camera model to match actions captured from different viewpoints and performed at different rates. Huang et al. [40] utilize the domain transfer ability of the canonical correlation analysis algorithm to obtain a correlation subspace as a joint representation for different viewpoints.

3. The proposed approach

Human activity prediction aims to infer an unfinished activity given a temporally incomplete video, i.e., before the full execution of an activity. We tackle the task by addressing two main problems: (1) representing the unfinished activity video using temporal order information, and (2) predicting the activity class by measuring the similarity between the unfinished query video and the complete training videos.

3.1. Activity representation for prediction

Our approach takes advantage of the space-time features for human activity representation. For the interest point detection, we use the cuboid detector introduced in [19], which uses separable linear filters for computing the response function of a video sequence. For the local feature description, we adopt the histograms-of-optical-flow (HOF) and histograms-of-oriented-gradients (HOG), which respectively characterize the motion and appearance information of a volume surrounding the interest point. Afterwards, we employ the K-means clustering method to obtain a vocabulary of size K based on the HOF and HOG features of the interest points extracted from the training set. Under the bag-of-visual-words model, each interest point is assigned to the most similar visual word.

A human activity is usually composed of a sequence of simpler actions, each of which contains different spatial-temporal information. Assuming the intra-class activities in different videos are performed in the similar action evolution progress, then we can group a small number of continuous frames as a segment as in [9]. Therefore, a video containing either an unfinished or complete activity is divided into a number of consecutive segments, and the number of segments roughly reflects the temporal status of the activity carried in the video. Each segment is denoted as a vector capturing the distribution of visual words, i.e., bag-of-visual-words, within the segment. Consequently, the final representation of an activity video is a time series of feature vectors, and the length of the series varies for different videos.

In practice, the same activity may be performed in speeds with some minor variation. To deal with such intra-class variation, we use a flexible representation with small perturbation in segment lengths when representing training videos. More specifically, for each training sequence, we generate five representations using segments of lengths $\Delta-4$, $\Delta-2$, Δ, $\Delta+2$, and $\Delta+4$ respectively, where Δ is the average segment length. That is, each activity video corresponds to five time series with different lengths of the segment, and the variety of training time series is largely increased to better approximate the query time series.

In the following we assume that the fully observed training video and the unfinished query video are separately divided into n_t and n_y segments. The training video and query video are thus denoted respectively as $T=\{t_1, t_2, \ldots, t_{n_t}\}$ and $Y=\{y_1, y_2, \ldots, y_{n_y}\}$. Their feature matrices are expressed as two time series X_T and X_Y, where $X_i=[\boldsymbol{x}_1^i, \ldots, \boldsymbol{x}_{n_i}^i] \in \mathbb{R}^{d_i \times n_i}$ and $i \in \{T, Y\}$. The columns of feature matrices are the d-dimensional histogram features extracted from corresponding video segments, i.e., $\boldsymbol{x}_k^T$ from t_k and $\boldsymbol{x}_k^Y$ from y_k.

With these representations, our task boils down to align an unfinished test video (e.g. Y) with the fully observed training videos (e.g., T), and then predict the activity class of the test video according to the similarity derived from the alignment. Because each training video results in five different time series, we separately align each time series with the test video, and adopt the maximum alignment accuracy as the similarity between the training video and the test video. A straightforward solution to our alignment problem is to use existing time warping algorithms. In the next subsections, we review one such choice, generalized time warping (GTW) [20], and we then propose a temporal-weighted generalized time warping (TGTW) algorithm to better meet the requirement of the prediction task.

3.2. Generalized time warping

GTW [20] extends the classical dynamic time warping (DTW) by incorporating a more flexible temporal warping scheme to compensate for temporal changes, and simultaneously allowing feature space manipulation (e.g., dimensionality reduction). It adopts an efficient linear-time optimization by using a Gauss-Newton algorithm.

In our problem setting, given two time series X_T and X_Y, where $X_i=[\boldsymbol{x}_1^i, \ldots, \boldsymbol{x}_{n_i}^i] \in \mathbb{R}^{d_i \times n_i}$ and $i \in \{T, Y\}$, they respectively denote the feature matrices of the training video T and the query video Y computed from the above subsection. For each X_i, GTW adopts a non-linear temporal transformation $W_i=(W_i(k,j)) \in \{0,1\}^{n_i \times l}$ and a low-dimensional spatial embedding $V_i \in \mathbb{R}^{d_i \times d}$, so the resulting sequence $V_i^\top X_i W_i \in \mathbb{R}^{d \times l}$ is aligned with each other in the least-squares algorithm. GTW minimizes the following objective function:

$$J_g(W_T, W_Y, V_T, V_Y) = \| V_T^\top X_T W_T - V_Y^\top X_Y W_Y \|_F^2 + \sum_{i \in \{T,Y\}} (\psi(W_i) + \phi(V_i)), \quad (1)$$

$$\text{s. t.} \quad W_i \in \Psi \quad \text{and} \quad V_i \in \Phi, \quad \forall\, i \in \{T, Y\},$$

where $\psi(\cdot)$ and $\phi(\cdot)$ are regularization terms; Ψ and Φ represent the domains for W_i and V_i. To solve the above non-convex optimization, in [20] an iterative solution is proposed and it contains two key steps: solving for W_i using a Gauss-Newton algorithm and computing V_i using multi-set canonical correlation analysis [35].

Given a warping matrix W, the alignment in GTW is described by a *warping path*, denoted here by $\boldsymbol{p}=(p(1), \ldots, p(l))^\top \in \{1{:}\, n\}^l$, such that $W(k,j)=\delta(p(j)-k)$. In [20], $\boldsymbol{p}$ is further written as a linear combination of a set of non-decreasing functions which compose the columns of basis matrix Q, i.e., $\boldsymbol{p}=Q\boldsymbol{a}$ where $\boldsymbol{a}$ contains the combination coefficients. With the new formulation, W can be parameterized as $W(\boldsymbol{a})$ and a solution is derived to iteratively minimize J_g based on the first order Taylor expansion, and in each iteration V_i and W_i are updated alternately. Details of the solution can be found in [20].

3.3. Activity prediction with TGTW

GTW can be used for aligning a partial sequence to a complete one for activity prediction. However, in activity prediction, an unfinished activity always starts from the beginning of the activity, suggesting that the aligning is in favor of the early portion of an activity. An intuitive example is illustrated in Fig. 2. Under the BOW representation, the

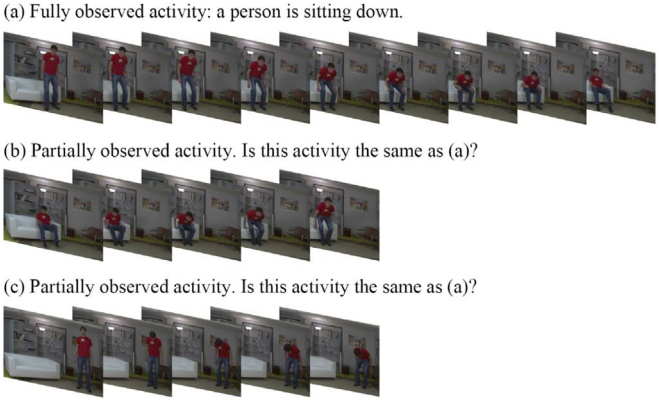


Fig. 2. An intuitive example showing the importance of matching position to activity prediction.

incomplete video (b) is analogous to the ending portion of (a), and (c) is similar to the beginning part of (a). Consequently, it's hard to decide which belongs to the same activity class as (a) using traditional GTW algorithm. But if encourage to match the unfinished action with the early portion of complete one, we can obtain that the alignment accuracy between (a) and (c) is obviously higher than that between (a) and (b). Actually, this is consistent with the fact that both (a) and (c) represent sitting down, and the activity in (b) is standing up.

The above observation motivates us to encode temporal weights during series alignment. We first explore a solution (denoted by TGTW-D) by naturally introducing time sensitive weights. We then derive a more effective solution (denoted by TGTW) that constrains on the warping path in GTW.

TGTW-D–Time Sensitive Cost Function. One way to encode the time sensitive prior is to modify the cost function in GTW by weighting more the alignment in latter portions than in early ones. This can be realized by multiplying a diagonal weighting matrix, denoted by $D = \mathrm{diag}(d_1, d_2, \ldots, d_l) \in \mathbb{R}^{l\times l}$, to the difference in J_g. The weight d_k constrains the corresponding column vector of the aligned feature matrix ($V_T^\top X_T W_T$ or $V_Y^\top X_Y W_Y$), corresponding to an aligned video segment in X or Y. To encourage alignment in the early portion, d_k is defined as increasing with k

$$d_k = 1 + \xi \cdot \frac{k}{l}, \tag{2}$$

where ξ is the variation span of video segments' weights.

Using the weight matrix D, we extend the original GTW to the following minimization problem:

$$\widetilde{J}_{\mathrm{tg}}(W_T, W_Y, V_T, V_Y) = \|V_T^\top X_T W_T D - V_Y^\top X_Y W_Y D\|_F^2 + \sum_{i\in\{T,Y\}} (\psi(W_i) + \phi(V_i)), \tag{3}$$

$$\text{s. t.} \quad W_i \in \Psi \quad \text{and} \quad V_i \in \Phi, \quad \forall\, i \in \{T, Y\}.$$

Due to the similarity between J_g and $\widetilde{J}_{\mathrm{tg}}$, we can solve the above problem similarly as GTW by iteratively first-order Taylor approximating and alternately updating W and V.

The above weighting strategy naturally inhibits the alignment in the ending portion of an activity, and hence improves the original GTW for activity prediction. However, the strategy meanwhile increases the importance of later segments if they are aligned. This accompanied effect probably impedes the performance of activity prediction and inspires us to seek for better alternatives.

TGTW–Time Sensitive Constraints over Warping Paths. Since the alignment is determined by the warping path, an alternative way to constrain the alignment is to penalize paths that have large elements (i.e., positions of the later portion). Given the two sequences T and Y described previously, such constraint can be naturally modeled by the time sensitive importance term $\sum_{t=1}^{l} p_T(t)$, where $\boldsymbol{p}_T = (p_T(1), p_T(2), \ldots, p_T(l))^\top$ is the warping path for T – note that only the constraint on the training (complete) activity sequence is needed. In this way, we have a new extension of GTW, named TGTW, as below:

$$J_{\mathrm{tg}}(W_T, W_Y, V_T, V_Y) = \|V_T^\top X_T W_T - V_Y^\top X_Y W_Y\|_F^2 + \sum_{i\in\{T,Y\}} (\psi(W_i) + \phi(V_i)) + \mu \sum_{t=1}^{l} p_T(t), \tag{4}$$

$$\text{s. t.} \quad W_i \in \Psi \quad \text{and} \quad V_i \in \Phi, \quad \forall\, i \in \{T, Y\}.$$

where μ is the regularization weight.

The formulation of TGTW in (4) is similar to the original GTW, with an additional regularization term. We use the same technique as in [20] to solve TGTW, i.e., to iteratively and alternately solving for V and W. Notice that, when W is fixed, the term $\mu \sum_{t=1}^{l} p_T(t)$ becomes a negligible constant, and hence the step for updating V (when fixing W) is identical as the original GTW [20].

When updating W for fixed V, we modify the solution in [20] to include the regularization term. Briefly speaking, for $i \in \{T, Y\}$, the warping matrices are first parameterized by the warping paths $\boldsymbol{p}_i$, which are further represented as linear combinations of non-decreasing basis Q_i in the form of $\boldsymbol{p}_i = Q_i \boldsymbol{a}_i$. Subsequently, updating W is equivalent to minimizing w.r.t. a_T,a_Y the following objective function $J_{\mathcal{W}}$:

$$J_W(\boldsymbol{a}_T, \boldsymbol{a}_Y) = \|V_T^\top X_T W (Q_T \boldsymbol{a}_T) - V_Y^\top X_Y W (Q_Y \boldsymbol{a}_Y)\|_F^2 + \sum_{i\in\{T,Y\}} \psi(\boldsymbol{a}_i) + \mu \mathbf{1}_l^\top Q_T \boldsymbol{a}_T, \tag{5}$$

$$\text{s. t.} \quad W(Q_i \boldsymbol{a}_i) \in \Psi, \quad \forall\, i \in \{T, Y\},$$

where $\mathbf{1}_l$ is a vector of ones of size l.

Imitating the optimization process in [20], we linearize the parameterized objective function $J_{\mathcal{W}}$ by performing the first order Taylor approximation on it. Therefore, the minimization of Eq. (5) is transformed to a quadratic programming problem which can be solved by classical methods.

By optimizing J_{tg} through alternately updating W_i and V_i, the query activity video is aligned with all the training videos. For each training video, the alignment error is computed as:

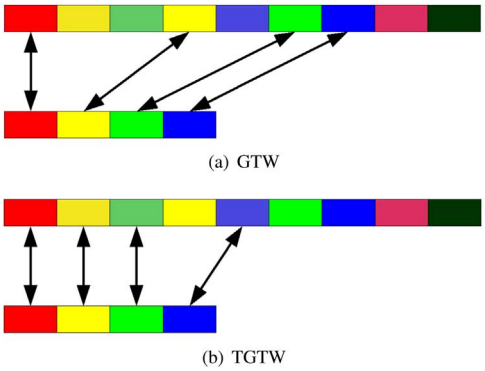
$$R = \|V_T^\top X_T W_T - V_Y^\top X_Y W_Y\|_F^2. \tag{6}$$

For an ongoing activity, our ultimate goal is to predict the class it belongs to according to the incomplete information. Based on our TGTW method, the query unfinished activity is aligned with each of the training videos. Obtaining all the alignment errors, the k-nearest neighbors algorithm is used to assign the query activity to the class which generates the minimum alignment error. Note that, the proposed TGTW strategy can be combined with other learning tools for activity prediction. In this paper, we focus on the alignment using TGTW, and therefore use k-NN for better understanding and fair evaluation.

Fig. 3 is a toy illustration for the difference between TGTW and GTW. In (a), though achieving the minimum matching cost, GTW matches the partial sequence to all over the full sequence. By contrast, our solution in (b) nicely constrains the matching to the starting portion, which is more appropriate for activity prediction. GTW only targets minimizing the alignment error, but TGTW considers a trade-off between the alignment error and the warping path.

4. Experimental results

To evaluate the capability of the proposed activity prediction method, three challenging datasets, i.e., the UT-Interaction dataset [36], the DARPA-Y1 dataset [38], and the UCF Sports dataset [37] are used in the experiments. Examples of these datasets are shown in



(a) GTW

(b) TGTW

Fig. 3. A toy example showing the difference between GTW and TGTW. (a) GTW targets minimizing the alignment error. (b) TGTW inclines to align the short series with the beginning part of the long series which is consistent with the requirement of the prediction task.

Fig. 4. The methods in [6,9] also use the interest point features to represent the activities. So we follow the same parameter setting with them, and generate a vocabulary with 800 visual words by K-means clustering for all three datasets in order to compare the performances with previous methods in the same condition. For each dataset, five percent of the average full activity length is considered as a video segment, and then an unfinished or complete activity is represented as a time series. For the TGTW-D model, the coefficient ξ is empirically set to 0.5 for three datasets. To align different time series in the TGTW model, the weight coefficient μ is important to the alignment accuracy. Therefore, we test different values for it on each dataset in the following subsections, and empirically set the coefficient $\mu = 0.004$ for the UT-Interaction and UCF Sports datasets, and set $\mu = 0.002$ for the DARPA-Y1 dataset. Because the videos on the DARPA-Y1 dataset are longer than those on the other two datasets, the value of the time sensitive importance term in TGTW corresponding to DARPA-Y1 is obviously larger. So a smaller trade-off weight is set on DARPA-Y1. The other parameter settings in time warping algorithm are identical as original GTW [20]. We simulate the technique in GTW, and optimize the temporal alignment problem using a Gauss-Newton algorithm. It has the linear complexity in terms of the length of the sequence. The average cost time to align two sequences is about 2.8 s using Matlab on a laptop with a 2.9 GHz Intel CPU and a 4 GB memory.

We compare our methods with several state-of-the-art methods, including integral bag-of-words (Ry_int) [6], dynamic bag-of-words (Ry_dyn) [6], sparse coding (SC) [9], and sparse coding on a mixture of segments (MSSC) [9]. Ry_int models the evolution of feature distributions as observations increase to predict partially observed activities. Based on Ry_int, Ry_dyn considers a video as a sequence of ordered intervals, and combines more activity structure information. MSSC uses features of video segments as bases, and apply sparse coding to construct the test video. MSSC adopts more bases corresponding to different temporal lengths than traditional sparse coding.

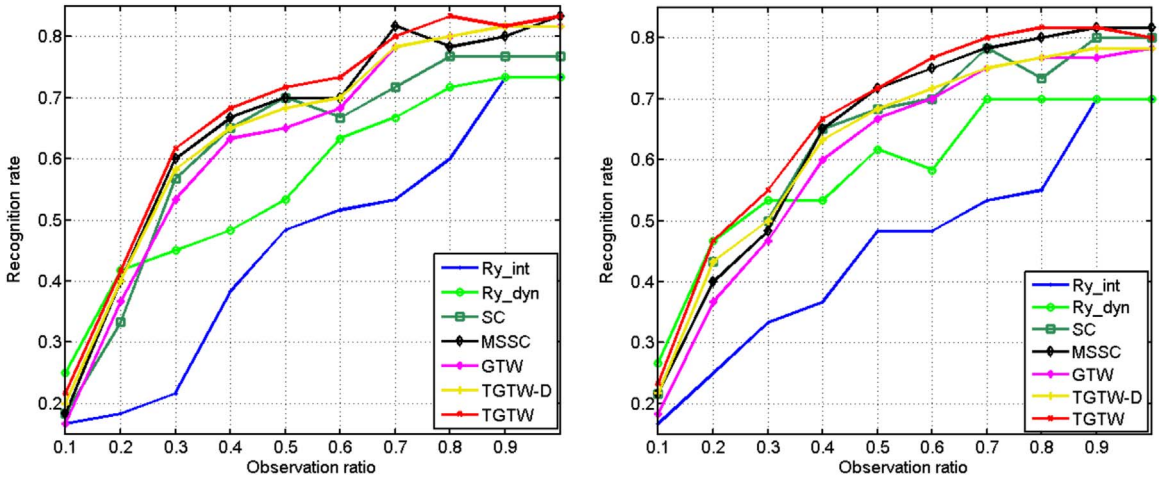
4.1. Experiments on the UT-interaction dataset

The UT-Interaction dataset has been used in the first Contest on Semantic Description of Human Activities [36]. This dataset contains action sequences of six interactions: *hug*, *kick*, *point*, *punch*, *push*, and *hand-shake*. For classification, 120 video segments cropped based on the ground-truth bounding boxes and time intervals are provided by the dataset organizers. These segments are further divided into two sets, and each set has 60 segments with 10 segments per class. Set 1 is captured at a parking lot and Set 2 at a lawn. Importantly, this dataset contains complex activities having sufficient temporal durations. Therefore, it is suitable for the experiments on activity prediction.

We follow the evaluation method in [6,9], and report the recognition rates corresponding to different observation ratios. For the performance evaluation, we use the leave-one-out cross validation, cycling each sample as the test video one at a time. The average accuracy over all tests is used as a quantitative metric of the performance.

Fig. 5 illustrates the prediction results on the UT-Interaction dataset. Different from traditional fully observed activity recognition, prediction is to classify a partially observed video. So we report the accuracies corresponding to different observation ratios. In most observation ratios, the proposed TGTW outperforms other methods on both sets by using the time sensitive importance term to constrain

Fig. 4. Representative frames from videos on three datasets. From top to bottom: the UT-Interaction dataset, the DARPA-Y1 dataset, the UCF Sports dataset.



(a) Set 1 (b) Set 2

Fig. 5. Prediction results on the UT-Interaction dataset.

the warping path. TGTW-D can also regulate the alignment position, but it simultaneously changes the weights of different video segments in alignment measurements. It obtains good performance but a little lower than TGTW. The accuracies of baseline GTW are more close to the proposed TGTW-D method in high observation ratios. By utilizing multi-scale segments to handle intra-class variations, MSSC performs better than SC, and outperforms TGTW-D and GTW in some observation ratios. The accuracy of Ry_dyn is higher than that of Ry_int, which ignores temporal relations among video segments.

In most existing methods for activity prediction, manual video segmentation is a preliminary operation, which is also a crucial step in time series alignment. Notice that, existing methods need some human intervention to achieve the segment correspondence between different activity sequences. However, the proposed TGTW-D and TGTW methods both automatically match the video segments in partially observed activity with corresponding parts in the complete sequence without any manual annotation. Obviously, our method is more appropriate to deal with automatic activity prediction problem.

In this paper, we compute the average length of complete activities in a dataset, and uniformly set five percent of it (i.e., consecutive six frames) as a segment. Moreover, the performances corresponding to different segment lengths, such as four, eight, and ten frames per segment, are tested using the whole UT-Interaction dataset (including set1 and set2). Fig. 6 illustrates the comparison results. When we respectively set six, eight, and ten consecutive frames as a segment, the prediction performances are approximative in most observation ratios. The result corresponding to four frames per segment is a little lower than others. The reason is that short video segment contains fewer action information, which decreases the sequence alignment accuracy. In general, the proposed prediction method is robust to the segment length variance.

In TGTW algorithm, the time sensitive importance term constrains the alignment position, which is the essential difference from the original GTW. In order to test the influence of time sensitive importance term on prediction rates, we compare the performances corresponding to different weight coefficients μ as shown in Fig. 7. If $\mu = 0$, the formulation of TGTW is transformed into original GTW. When the observation ratio is high, the accuracies of GTW and TGTW are comparative. However, TGTW corresponding to different nonzero weight coefficients all performs better than GTW when the observation ratio is below 0.6. Through constraining the alignment position, the proposed time sensitive term improves the original GTW algorithm and makes it more effective to recognize partially observed activity.

4.2. Experiments on the DARPA-Y1 dataset

To validate the effectiveness of our TGTW-based methods, we further conduct experiments on the DARPA-Y1 dataset. It is challenging in several aspects: the actor size in the same activity class varies

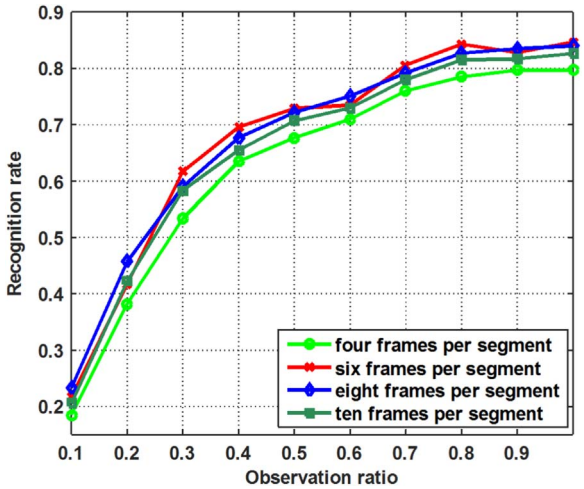


Fig. 6. Prediction results corresponding to different segment lengths.

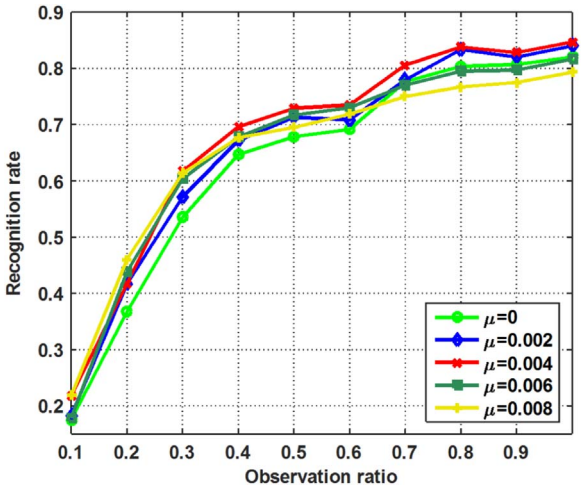


Fig. 7. Prediction results corresponding to different weight coefficients on the UT-Interaction dataset.

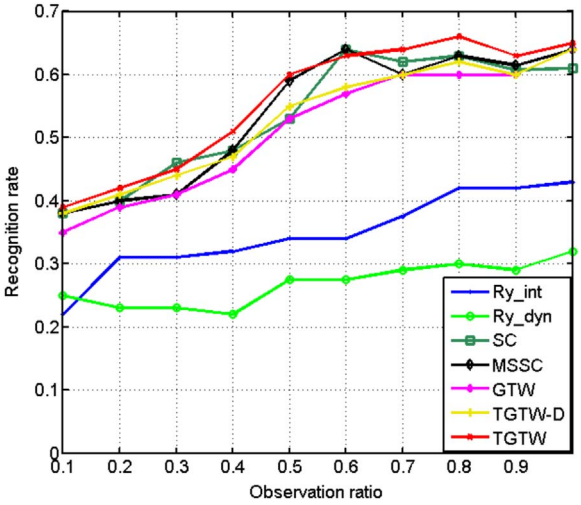


Fig. 8. Prediction results on the DARPA-Y1 dataset.

significantly in different videos; the overhead time for an activity class varies from one video to another; activities are captured from different camera viewpoints; and backgrounds are more complex due to shadows and non-uniform illuminations. Being a subset of videos from the Year-1 corpus of the DARPA Mind's Eye program [38], DARPA-Y1 contains videos from seven human activities: *fall*, *haul*, *hit*, *jump*, *kick*, *push*, and *turn*.

Fig. 8 illustrates the prediction results on the DARPA-Y1 dataset. Under the leave-one-out cross validation, the proposed TGTW outperforms other methods in most observation ratios. TGTW-D, GTW, MSSC, and SC all obtain comparable performances. The accuracies of the proposed TGTW methods are both increased from the baseline GTW. Different from the results on the UT-Interaction dataset, the integral bag-of-words outperforms the dynamic bag-of-words method. But the recognition rates of them are both not as good as above methods. They require two sequences have the same number of segments and the similarity is measured between corresponding segments. The rigid correspondences restrict the performance of them in large intra-class variation datasets. However, the proposed two GTW-based methods are more flexible in series matching. These results further validate the effectiveness of our methods. Furthermore, we test the influence of the time sensitive term on this dataset, and record the prediction accuracies corresponding to different weight coefficients as shown in Fig. 9. Obviously, in low observation ratios, the proposed TGTW ($\mu \neq 0$) outperforms traditional GTW (μ=0).

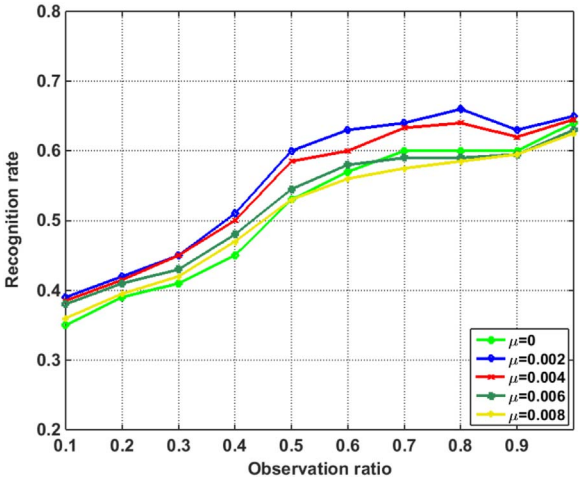


Fig. 9. Prediction results corresponding to different weight coefficients on the DARPA-Y1 dataset.

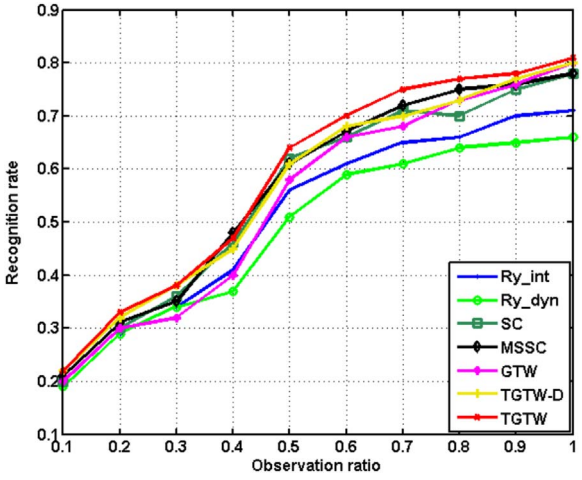


Fig. 10. Prediction results on the UCF Sports dataset.

It validates the effectiveness of the proposed time sensitive term for the prediction of activity videos with low observation ratios. However, through the comparisons of different nonzero weight coefficients, larger weights don't result in better performances. The reason is that the time sensitive term only constrains the alignment position, but it doesn" increase the alignment accuracy.

4.3. Experiments on the UCF sports dataset

The UCF Sports is a challenging dataset for the activity prediction task. It is a collection of 150 broadcast sports videos of ten different classes of activities, including *dive*, *golf swing*, *kick*, *lift*, *horseback ride*, *run*, *skate*, *swing-bench*, *swing-side*, and *walk*. The videos are captured in dynamic and cluttered environments from a wide range of camera views and realistic scenes. Most previously reported results on this dataset use leave-one-out manner. But Lan et al. [39] propose to split the dataset into disjoint training and test sets to avoid the background regularity for evaluation. We follow Lan's strategy to evaluate recent prediction methods.

Fig. 10 shows the prediction results on the UCF Sports dataset. The proposed TGTW-based methods again achieve promising performances, and TGTW outperforms TGTW-D in all observation ratios. When the length of unfinished video is less than thirty percents of the whole activity video, the performance of baseline GTW is lower than others, but it obtains competitive results when the observation ratio is high. MSSC and SC have similar recognition rates in most observation ratios. The results of Ry_int and Ry_dyn are not as well as others which is similar to the DARPA-Y1 dataset. The performances on the large intra-class variation dataset further validate that the proposed TGTW-based methods are effective for the prediction task. To further validate the importance of the proposed warping path constraint, we also adopt different weight coefficients to test the prediction results on the UCF Sports dataset. The performances illustrated in Fig. 11 validate the robustness of the proposed TGTW model.

The activity prediction task is more challenging than traditional fully observed human behavior analysis. Besides the common difficulties (e.g., losing spatial and temporal information in activity representation) in traditional activity recognition, the prediction task has to deal with some new problems (e.g., searching for the corresponding part between different activity videos). Because of these challenges, current methods still have high prediction errors on benchmark datasets, and need many improvements in the future.

5. Conclusion

In this paper, we have proposed integrating temporal prior in time

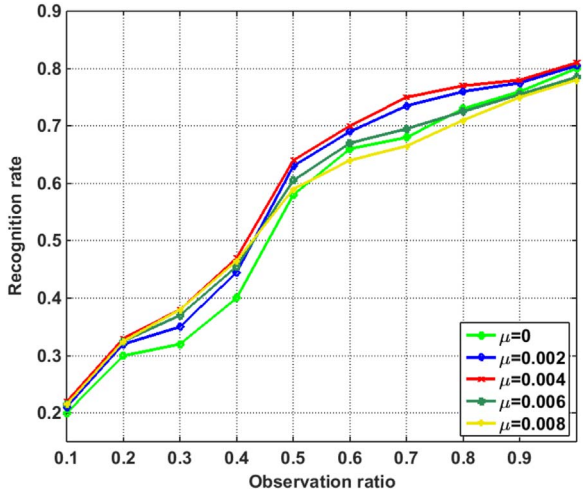


Fig. 11. Prediction results corresponding to different weight coefficients on the UCF Sports dataset.

series alignment for predicting unfinished human activities. In particular, we have extended the recently proposed generalized time warping (GTW) algorithm by adding temporal constraints over the warping path to encourage the matching in the early portion of an activity, which is desired by the nature of activity prediction. The proposed algorithm, named temporally weighted GTW (TGTW), has been validated for activity prediction on three publicly available benchmark datasets. In all experiments, TGTW shows excellent results and outperforms several state-of-the-art activity prediction algorithms.

Acknowledgements

We thank F. Zhou and F. De la Torre for sharing the GTW code online. This work is supported in part by National Natural Science Foundation of China (61603080, 61473086, 61375001), the Fundamental Research Funds for the Central Universities of China (N150403006), the NSF of Jiangsu Province (BK20140566, BK20150470), and China Postdoctoral science Foundation (2014M561586).

References

[1] L. Liu, L. Shao, X. Li, K. Lu, Learning Spatio-Temporal representations for action recognition: a genetic programming approach, IEEE Trans. Cybern. 46 (1) (2016) 158–170.
[2] P.V.K. Borges, N. Conci, A. Cavallaro, Video-based human behavior understanding: a survey, IEEE Trans. Circuits Syst. Video Technol. 23 (11) (2013) 1993–2008.
[3] P. Turaga, R. Chellappa, V.S. Subrahmanian, O. Udrea, Machine recognition of human activities: a survey, IEEE Trans. Circuits Syst. Video Technol. 18 (11) (2008) 1473–1488.
[4] J.K. Aggarwal, M.S. Ryoo, Human activity analysis: a review, ACM Comput. Surv. (CSUR) 43 (3) (2011).
[5] M. Hoai, F. De la Torre, Max-Margin early event detectors, in: Proceedings of IEEE International'l Conference Computer Vision and Pattern Recognition, 2012, pp. 2863–2870.
[6] M.S. Ryoo, Human activity prediction: early recognition of ongoing activities from streaming videos, in: Proceedings of IEEE International'l Conference Computer Vision, 2011, pp. 1036–1043.
[7] N. Raman, S. Maybank, Activity recognition using a supervised non-parametric hierarchical HMM, Neurocomputing 199 (2016) 163–177.
[8] N. Harbi, Y. Gotoh, A unified spatio-temporal human body region tracking approach to action recognition, Neurocomputing 161 (2015) 56–64.
[9] Y. Cao, D. Barrett, A. Barbu, S. Narayanaswamy, H. Yu, A. Michaux, Y. Lin, S. Dickinson, J. M. Siskind, S. Wang, Recognizing Human Activities from Partially Observed Videos, in: Proceedings of IEEE International'l Conference Computer Vision and Pattern Recognition, 2013, pp. 2658–2665.
[10] Z. Xu, L. Qing, J. Miao, Activity auto-completion: predicting human activities from partial videos, in:Proceedings of IEEE International'l Conference Computer Vision, 2015, pp. 3191–3199.
[11] K. Kitani, B. Ziebart, J. Bagnell, M. Hebert, Activity Forecasting, Proceedings European Conference Computer Vision, 2012, pp. 201–214.
[12] Y. Kong, D. Kit, Y. Fu, A discriminative model with multiple temporal scales for action prediction, in:Proceedings of European Conference Computer Vision, 2014, pp. 596–611.
[13] Y. Kong, Y. Fu, Max-margin action prediction machine, IEEE Trans. Pattern Anal. Mach. Intell. 38 (9) (2015).
[14] K. Li, J. Hu, Y. Fu, Modeling complex temporal composition of actionlets for activity prediction, in: Proceedings of European Conference Computer Vision, 2012. pp. 286–299.
[15] F. Moayedi, Z. Azimifar, R. Boostani, Structured sparse representation for human action recognition, Neurocomputing 161 (2015) 38–46.
[16] H. Wang, C. Yuan, W. Hu, H. Ling, W. Yang, C. Sun, Action recognition using nonnegative action component representation and sparse basis selection, IEEE Trans. Image Process. 23 (2) (2014) 570–581.
[17] H. Wang, C. Yuan, W. Hu, C. Sun, Supervised class-specific dictionary learning for sparse modeling in action recognition, Pattern Recognit. 45 (11) (2012) 3902–3911.
[18] J. Walker, A. Gupta, M. Hebert, Patch to the Future: unsupervised visual prediction, in: Proceedings of IEEE International'l Conference Computer Vision and Pattern Recognition, 2014, pp. 3302–3309.
[19] P. Dollar, V. Rabaud, G. Cottrell, S. Belongie, Behavior recognition via sparse spatiotemporal features, in: IEEE Int'l Workshop on Visual Surveillance and Performance Evaluation of Tracking and Surveillance, 2005, pp. 65–72.
[20] F. Zhou, F. De la Torre, Generalized time warping for multi-modal alignment of human motion, in: Proceedings of IEEE International'l Conference Computer Vision and Pattern Recognition, 2012, pp. 1282–1289.
[21] L. Rabiner, B. Juang, Fundamentals of Speech Recognition, Prentice Hall, Inc., Upper Saddle River, NJ, USA, 1993.
[22] T. Sebastian, P. Klein, B. Kimia, On aligning curves, IEEE Trans. Pattern Anal. Mach. Intell. 25 (1) (2003) 116–125.
[23] H. Ling, D. Jacobs, Shape classification using the inner-distance, IEEE Trans. Pattern Anal. Mach. Intell. 29 (2) (2007) 286–299.
[24] I. Junejo, E. Dexter, I. Laptev, P. Perez, View-independent action recognition from temporal self-similarities, IEEE Trans. Pattern Anal. Mach. Intell. 33 (1) (2011) 172–185.
[25] W. Pan, L. Torresani, Unsupervised Hierarchical Modeling of locomotion Styles, in: Proceedings of ACM International'l Conference Machine Learning, 2009, pp. 785–792.
[26] E. Hsu, K. Pulli, J. Popovic, Style translation for human motion, ACM Trans. Graph. 24 (3) (2005) 1082–1089.
[27] A. Heloir, N. Courty, S. Gibet, F. Multon, Temporal alignment of communicative gesture sequences, Comput. Animat. Virtual Worlds 17 (3) (2006) 347–357.
[28] Y. Caspi, M. Irani, Aligning non-overlapping sequences, Int. J. Comput. Vis. 48 (1) (2002) 39–51.
[29] A. Gritai, Y. Sheikh, C. Rao, M. Shah, Matching trajectories of anatomical landmarks under view-point, anthropometric and temporal transforms, Int. J. Comput. Vis. 84 (3) (2009) 325–343.
[30] F. Padua, R. Carceroni, G. Santos, K. Kutulakos, Linear sequence-to-sequence alignment, IEEE Trans. Pattern Anal. Mach. Intell. 32 (2) (2010) 304–320.
[31] R. Li, R. Chellappa, Aligning Spatio-Temporal Signals on A Special Manifold, Proceedings European Conference Computer Vision, 2010, pp. 547–560.
[32] C. Rao, A. Gritaiand, M. Shah, T. Syeda-Mahmood, View-Invariant Alignment and Matching of Video Sequences, in: Proceedings of IEEE International'l Conference Computer Vision, 2003, pp. 939–945.
[33] M. Singh, I. Cheng, M. Mandal, A. Basu, Optimization of symmetric transfer error for sub-frame video synchronization, in: Proceedings of European Conference Computer Vision, 2008, pp. 554–567.
[34] F. Zhou, F. De la Torre, Canonical time warping for alignment of human behavior, Adv. Neural Inf. Process. Syst. (2009) 2286–2294.
[35] M.A. Hasan, On Multi-set Canonical Correlation Analysis, International Joint Conference on Neural Networks, Jun. 2009.
[36] M.S. Ryoo, J.K. Aggarwal, An Overview of Contest on Semantic Description of Human Activities (SDHA), Data Set http://cvrc.ece.utexas.edu/SDHA2010/, 2010.
[37] M.D. Rodriguez, J. Ahmed, M. Shah, Action Mach A Spatio-Temporal Maximum Average Correlation Height Filter for Action Recognition, in: Proceedings of IEEE International'l Conference Computer Vision and Pattern Recognition, 2008, pp. 1–8.
[38] DARPA. Video Dataset from DARPA Mind's Eye Program. http://www.visint.org, 2011.
[39] T. Lan, Y. Wang, G. Mory, Discriminative Figure-Centric Models for Joint Action Localization and Recognition, in: Proceedings of IEEE International'l Conference Computer Vision, 2011. pp. 2003–2010.
[40] C. Huang, Y. Yeh, Y. Wang, Recognizing actions across cameras by exploring the correlated subspace, in: Proceedings of European Conference Computer Vision, 2012, pp. 342–351.

Haoran Wang received the B.S. degree from the Department of information science and technology, Northeast University, China, in 2008, and the PhD degree from School of Automation, Southeast University, China, in 2015. Since spring 2015, he has been an Assistant Professor at Northeastern University, China. His research interests include computer vision, pattern recognition, and machine learning.

Wankou Yang received his B.S., M.S. and Ph.D. degrees at the School of Computer Science and Technology, Nanjing University of Science and Technology (NUST), PR.China, 2002, 2004 and 2009 respectively. Now he is an Associate Professor with School of Automation, Southeast University. His research interests include pattern recognition, computer vision, digital machine learning.

Chunfeng Yuan received the B.S. degree and the MS degree in computer science from Qingdao University of Science and Technology, China, in 2004 and 2007, respectively, and the PhD degree in computer science from the Institute of Automation (CASIA), Chinese Academy of Sciences, Beijing, China, in 2010. Since then she has been an assistant professor at the CASIA. Her research interests and publications range from statistics to computer vision, including sparse representation, motion analysis, action recognition, and event detection.

Haibin Ling received the B.S. degree in mathematics and the MS degree in computer science from Peking University, China, in 1997 and 2000, respectively, and the PhD degree from the University of Maryland, College Park, in Computer Science in 2006. From 2000–2001, he was an assistant researcher at Microsoft Research Asia. From 2006–2007, he worked as a postdoctoral scientist at the University of California Los Angeles. After that, he joined Siemens Corporate Research as a research scientist. In 2008, he joined Temple University where he is now an Associate Professor. Dr. Ling's research interests include computer vision, medical image analysis, human computer interaction, and machine learning. He received the Best Student Paper Award of ACM UIST in 2003 and the NSF CAREER Award in 2014. He serves on the editorial board of IEEE Trans. on Pattern Analysis and Machine Intelligence and Pattern Recognition, and served as Area Chairs for CVPR 2014 and CVPR 2016.

Weiming Hu received the PhD degree from the Department of Computer Science and Engineering, Zhejiang University, in 1998. From April 1998 to March 2000, he was a postdoctoral research fellow with the Institute of Computer Science and Technology, Peking University. Currently, he is a professor in the Institute of Automation, Chinese Academy of Sciences. His research interests include visual surveillance, and filtering of Internet objectionable information.