Real-Time RGB-D Activity Prediction by Soft Regression

Jian-Fang Hu[2,3,4], Wei-Shi Zheng[2,3(✉)], Lianyang Ma[4], Gang Wang[4], and Jianhuang Lai[1,3]

[1] Guangdong Key Laboratory of Information Security Technology, Guangzhou, China
[2] Key Laboratory of Machine Intelligence and Advanced Computing, MOE, Guangzhou, China
wszheng@ieee.org
[3] Sun Yat-sen University, Guangzhou, China
hujianf@mail2.sysu.edu.cn, stsljh@mail.sysu.edu.cn
[4] Nanyang Technological University, Singapore, Singapore
wanggang@ntu.edu.sg, lianyangma2012@gmail.com

Abstract. In this paper, we propose a novel approach for predicting ongoing activities captured by a low-cost depth camera. Our approach avoids a usual assumption in existing activity prediction systems that the progress level of ongoing sequence is given. We overcome this limitation by learning a soft label for each subsequence and develop a soft regression framework for activity prediction to learn both predictor and soft labels jointly. In order to make activity prediction work in a real-time manner, we introduce a new RGB-D feature called "local accumulative frame feature (LAFF)", which can be computed efficiently by constructing an integral feature map. Our experiments on two RGB-D benchmark datasets demonstrate that the proposed regression-based activity prediction model outperforms existing models significantly and also show that the activity prediction on RGB-D sequence is more accurate than that on RGB channel.

Keywords: Activity prediction · RGB-D · Soft regression

1 Introduction

Recognizing activities before they are fully executed is very important for some real-world applications like visual surveillance, robot designing [1,2], and clinical monitoring [3]. Activity prediction is to predict ongoing activities using the observed subsequences that only contain partial activity execution.

Existing action/activity prediction model [4] requires manual labeling of a large amount of video segments, which however record mostly low level actions shared among activities, and thus it is too expensive and sometimes is difficult to label. Although an alternative way is to simply label a subsequence[1] as the

[1] In this work, the subsequence of an activity means the accumulation of consecutive segments from the start of the activity.

B. Leibe et al. (Eds.): ECCV 2016, Part I, LNCS 9905, pp. 280–296, 2016.
DOI: 10.1007/978-3-319-46448-0_17

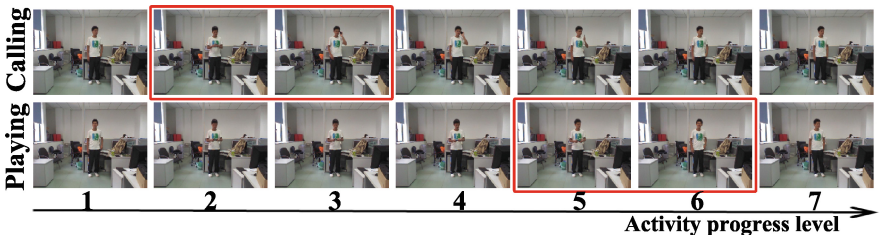


Fig. 1. Snapshots from activities *calling with a cell-phone* and *playing with a cell-phone*. As presented in the first row, it is hard to recognize the activity when its progress level is less than 3. However, if the segment with temporal interval [3, 4] (marked with red box) are provided, it becomes clear that the subject is performing the activity *calling with a cell-phone*. We also observe that the subsequences at progress level 2 (temporal interval [1, 2]) in the two activities contain the same action "taking out a cell-phone". (Color figure online)

label of the full sequence [5], sometimes, this naive labeling would make ambiguity for predicting ongoing activity. It is because an activity sequence consists of several segments, and segments from different activities could be similar so that each segment could be ambiguous when used to predict the label of the full activity. Taking Fig. 1 as an example, the action "*taking out a cell-phone*" appears in both the activity sequences "*calling with a cell-phone*" and "*playing with a cell-phone*", so it will mislead the predictor learning if we directly treat the subsequence of taking out phone in the first row as "*calling with phone*" while define the similar subsequence in the second row as "*playing with a cell-phone*".

To address the above problem, we learn a soft label for the subsequences of each progress level. The soft label tells how likely the subsequence is performing the activity depicted in the corresponding full sequence. Introducing the soft labels for subsequences can alleviate the confusion caused by the fact that subsequences from different activities could contain similar activity contents. A regression-based activity prediction model is therefore developed to regress these soft labels. A characteristic of our regression-based prediction model is to learn the activity predictor and the soft labels jointly without any prior knowledge set for soft labels. In this way, our modeling avoids overfitting caused by the ambiguity of subsequence on prediction and meanwhile makes the prediction model work in a discriminant way on partially observed activity sequences. And more importantly, by learning soft labels, the usual assumption on given the progress level of subsequence [6,7] is not necessary for our model.

In addition, most of activity prediction works in literatures focus on predicting human activities from RGB sequences by matching visual appearance and human motion among the activities executed at different progress levels [3,4,6,8,9]. However, RGB sequences are intrinsically limited in capturing highly articulated motions due to the inherent visual ambiguity caused by clothing similarity among people, appearance changes from view point difference, illumination variation, cluttered background and occlusions [10,11]. The recently introduced

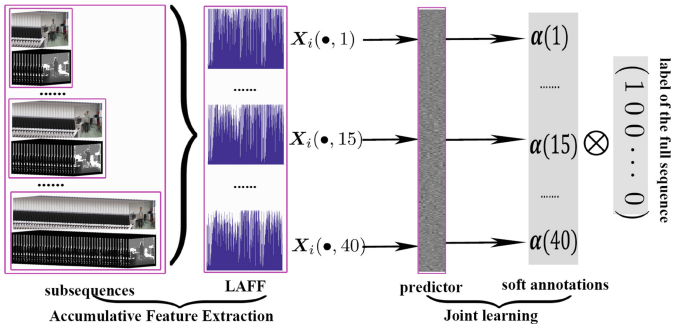


Fig. 2. A graphic illustration of our soft regression model. $\otimes$ is Kronecker product.

low-cost depth cameras can alleviate the ambiguity due to the availability of more modal data describing activity such as depth of scene and 3D joint positions of human skeleton. Hence, in this work, we explore real-time activity prediction with the assistance of depth sensors.

Towards making our prediction model work on RGB-D sequence in real-time, we design "local accumulative frame feature (LAFF)" to characterize the activity context of RGB-D sequence with arbitrary activity progress levels. The RGB-D context will include the appearance, shapes and skeletons of human body parts, manipulated objects and even scene (background). By employing the popularly used integral map computing technique, we demonstrate that the formulated LAFF can be efficiently computed in a recursive manner and thus be suitable for real-time prediction. The flowchart of our method is illustrated in Fig. 2.

In summary, the main contributions are: (1) A soft regression-based activity prediction model is formulated for overcoming the usual assumption that the progress level of ongoing activity is given; and (2) A local accumulative frame feature (LAFF) is developed for real-time activity prediction on RGB-D sequences. We claim that the prediction on RGB-D sequences works much better than that on RGB videos only. To verify our claim, we have tested our method on two RGB-D activity sets. The proposed method can obtain more reliable performances for predicting activities at varied progress levels. It can process more than 40 frames per second on a normal PC using MATLAB without elaborate optimization of programming, which can be used for real-time activity prediction.

2 Related Work

Activity prediction. In many real-world scenarios like surveillance, it would be more important to correctly predict an activity before it is fully executed. Many efforts are on developing early activity detectors or future activity prediction systems [7–9,12–14]. For example, [12,15] explored the application of max-margin learning in early event recognition and detection. [8] developed an early activity prediction system according to the change of feature distribution as more

and more video streams were observed. Lan et al. proposed to represent human movements in a hierarchical manner and employ a max-margin learning framework to select the most discriminative features for prediction [3]. [4] proposed to mine some sequential patterns that frequently appear in the training samples for prediction. Recently, [7] extended the max-margin event detector for activity prediction and obtained the state-of-the-art results on several benchmark sets. However, it is assumed in [7] that the progress level of ongoing activity is provided along with the observed sequence even in the test phase, which renders their method unrealistic in the real-world applications as it is hard to obtain the progress level of ongoing activity until it has been fully executed and observed.

Recent researches on human activity prediction is mainly focusing on predicting activities from ongoing RGB sequences, while less work has been reported on RGB-D sequences captured by low-cost depth cameras. In this paper, we consider the prediction of RGB-D activity sequence and develop a real-time system for predicting human activities without any additional prior information about the progress level of ongoing activity. The most closest to our approach is the online RGB-D activity prediction system in [5]. However, the system in [5] is based on frame-level prediction and the long-term motions are discarded in their model. Moreover, the subsequences with partial activity executions are not exploited for prediction, which renders their method less accurate for activity prediction.

Activity recognition with monocular video sensor. Human activity recognition is a long-term research topic and it has attracted a lot of attentions in the past decades. A large number of considerable progresses have been made for developing robust spatiotemporal features (Cuboids [16], interest point clouds [17], HOG3D [18], dense trajectory [19], and two-stream CNN [20] etc.) and feature learning techniques (sparse coding [21], max-margin learning [22–24], Fisher vector [19] etc.). Activity recognition aims at developing algorithms and systems for after-of-the-fact prediction of human activity, where activity sequence is entirely observed. Consequently, the activity recognition methods cannot be directly used for activity prediction task, which needs to work on ongoing sequences.

Activity recognition with depth cameras. The emergence of Kinect device has lit up the research of human activity recognition with depth cameras in these years. In the literatures, how to acquire a robust feature representation for the depth sequences is one of the most fundamental research topics. A lot of RGB video descriptors have been extended in order to characterize 3D geometries depicted in depth sequences [25–29]. For example, [25,26] developed their depth descriptors by extending the idea of constructing histogram of oriented gradient [27]. Considering the close relationship between human pose(skeleton [30]) and activity, some researchers seeked to represent human activities using positional dynamics of each skeleton joint [31–33] or joint pairs [34–39]. Human activity may contain complex interactions between the actor and objects, using depth and skeleton channels is not sufficient for describing the interactions. RGB channel is also utilized for feature representation [10,11,40]. In this paper, we construct a

RGB-D sequence feature by combining the local descriptors extracted from color patterns, depth patterns and skeletons. Different from the previous work that extracts features for off-line computation [10,11,40], we formulate our feature modeling in a recursive manner so that it can be computed in real-time.

It is worth noting that the conception of soft label has also been recently explored in [41] for improving RGB-D activity recognition, where the authors allow the human annotators to assign a soft label for the video segment with ambiguity. However, it requires manual setting on labeling.

3 Our Approach

3.1 Problem Statement

We concern a real-time prediction system for identifying ongoing activity sequence. In activity prediction, the activity depicted in the observed sequence is always uncompleted before it has been fully executed. In this work, we consider a more realistic setting for this problem. Unlike the activity prediction considered in [6,7], we do not assume that the progress level of ongoing activity is known in the test phase, as it is hard (if not impossible) to have a surveillance system obtain the progress level of ongoing activity until it has been fully executed. In this work, we propose a predictor that can be generally used for predicting an (ongoing) activity sequence at any progress level.

Notation. Throughout this paper, we use bold uppercase characters to denote matrices and bold lowercase characters (or Greek letters) to denote vectors. For any matrix $\boldsymbol{A}$, we use $\boldsymbol{A}(i,\cdot)$, $\boldsymbol{A}(\cdot,j)$ and $\boldsymbol{A}(i,j)$ to denote the $i-th$ row, the $j-th$ column and the (i,j)-element of $\boldsymbol{A}$, respectively. $\boldsymbol{A}^T$ denotes the transpose matrix of $\boldsymbol{A}$. In this work, we consider the Frobenius norm $||\boldsymbol{A}||_F$ and $L_{1,2}$-norm $||\boldsymbol{A}||_{1,2}$ of a matrix and the l_2 norm $||\boldsymbol{a}||_2$ of a vector. The $L_{1,2}$-norm for matrix $\boldsymbol{A} \in \mathbb{R}^{m\times n}$ is defined as:

$$||\boldsymbol{A}||_{1,2} = \sum_{j=1}^{n} \sqrt{\sum_{i=1}^{m} \boldsymbol{A}(i,j)^2} = \sum_{j=1}^{n} \boldsymbol{r}(j). \tag{1}$$

Here, $\boldsymbol{r}(j)$ represents the l_2 norm of the $j-th$ column $A(\cdot,j)$. Then we can obtain the generalized gradient[2] $\frac{\partial||\boldsymbol{A}||_{1,2}}{\partial \boldsymbol{A}(i,j)} = \frac{\partial \boldsymbol{r}(j)}{\partial \boldsymbol{A}(i,j)} = \frac{\boldsymbol{A}(i,j)}{\boldsymbol{r}(j)}$. This equation indicates that the gradient of $||\boldsymbol{A}||_{1,2}$ with respect to $\boldsymbol{A}$ can be easily obtained by performing a column-wise normalization on the matrix $\boldsymbol{A}$. Here we denote this column-wise normalization operator as $\mathcal{L}$ for convenience.

3.2 Local Accumulative Frame Feature (LAFF)

Since existing RGB-D sequence features [10,11,40] are not for online feature extraction, they are less applicable for online activity prediction. To overcome

[2] We would add a small positive constant ϵ to $\boldsymbol{r}(j)$ when it is zero.

this problem, we propose an effective RGB-D sequence feature representation by employing the popularly used integral map computing technique. In details, we extract local HOG descriptors from RGB and depth patches around each body part and extract relative skeleton features for each frame in order to capture activity contexts including human motions, appearance, shapes of human body parts, the manipulated objects and even the scene. In order to further capture temporal structure, a $3-level$ temporal pyramid is constructed by repeatedly partitioning the observed sequence into increasingly finer sub-segments along temporal dimension. The features of the frames found in each sub-segment are accumulated together using a mean pooling method. The concatenation of all accumulative features forms our local accumulative frame feature (LAFF).

In the following, we show that LAFF can be calculated efficiently by constructing an integral feature map $\boldsymbol{I}$:

$$\boldsymbol{I}(\cdot, T) = \sum_{t=1}^{T} \boldsymbol{F}(\cdot, t). \tag{2}$$

where $\boldsymbol{F} \in \mathbb{R}^{d \times T}$ is the local features extracted from frames in the sequence, d denotes the feature dimension and T is the number of total frames. We can compute the accumulative feature $\boldsymbol{x}$ between frames t_1 and t_2 ($t_2 > t_1$) as follows:

$$\boldsymbol{x} = \frac{\boldsymbol{I}(\cdot, t_2) - \boldsymbol{I}(\cdot, t_1 - 1)}{t_2 - t_1 + 1}. \tag{3}$$

Therefore, the LAFF features with 7 temporal intervals ($1 + 2 + 4$ sub-segments in the 3-level pyramid) can be efficiently computed from the formulated integral feature map using Eq. (3). This enables online and real-time computation.

3.3 Model Formulation

We assume that training activity sequences contain complete activity executions. To train an activity predictor, similar to existing works [7,8], we uniformly divide the fully observed training sequences into N segments. Let $V(\cdot, \cdot, \cdot)$ be the full sequence, and we use a vector $\boldsymbol{\pi}$[3]$\in \mathbb{R}^{N+1}$ to indicate the temporal locations of the segments. For example, $V(\cdot, \cdot, \boldsymbol{\pi}(1) : \boldsymbol{\pi}(2))$ represents the sequence of the first segment. Always, we call $V(:, \boldsymbol{\pi}(1), \boldsymbol{\pi}(n+1))$ an activity's subsequence of progress level n. And correspondingly, its *observation ratio* can be defined as $\frac{n}{N}$.

Let $\{(\boldsymbol{X}_1, \boldsymbol{y}_1), (\boldsymbol{X}_2, \boldsymbol{y}_2), ..., (\boldsymbol{X}_I, \boldsymbol{y}_I)\}$ be the training data that consist of I examples from L classes, where $\boldsymbol{y}_i \in \mathbb{R}^N$ is a label vector of $\boldsymbol{X}_i$, each $\boldsymbol{X}_i \in \mathbb{R}^{d \times N}$ has N instances, and each instance $\boldsymbol{X}_i(\cdot, n)$ is represented by the LAFF feature of the subsequence of progress level n. The label vector $\boldsymbol{y}_i$ is a binary vector, having its $j-th$ entry set to 1 if it is from the $j-th$ class and 0 otherwise.

Indeed, the unfinished subsequences are quite different from the full sequences because the contents contained in the unobserved duration may include some important conception for the complete activity definition.

[3] Intuitively, we need the vector to satisfy the boundary constraint $\boldsymbol{\pi}(1) = 1$, $\boldsymbol{\pi}(N+1) = T$ and the monotonicity constraint $\boldsymbol{\pi}(t_1) \leq \boldsymbol{\pi}(t_2)$ for any $t_1 \leq t_2$.

Since a subsequence is ambiguous, labeling it as the label of its full sequence could make confusing. To overcome this problem, we learn a soft label for each subsequence and define the label of the subsequence with a progress level n as $\boldsymbol{\alpha}(n)\boldsymbol{y}_i$ where $0 \leq \boldsymbol{\alpha}(n) \leq 1$. $\boldsymbol{\alpha}(n)\boldsymbol{y}_i$ can be conceived as how likely the subsequence is from the activity class $\boldsymbol{y}_i$. Using and learning soft labels can alleviate the confusion caused by the fact that subsequences from different activities could contain similar activity content (See Fig. 1 for example). In addition, it also enables the prediction of ongoing activity at any progress level in our modeling.

To learn the soft labels rather than setting them empirically, we form a soft regression model for learning them and activity predictor jointly as follows:

$$\min_{\boldsymbol{W},\boldsymbol{\alpha}} \sum_{i=1}^{I}\sum_{n=1}^{N} \overbrace{\boldsymbol{s}(n)||\boldsymbol{W}^T\boldsymbol{X}_i(\cdot,n) - \boldsymbol{y}_i\boldsymbol{\alpha}(n)||_{1,2}}^{Prediction\ loss\ term} + \frac{\xi_2}{2} \overbrace{||\boldsymbol{W}||_F^2}^{Regularization\ term}$$
$$s.t. \quad \boldsymbol{\alpha}^T\boldsymbol{e}_N = 1, 0 \leqslant \boldsymbol{\alpha} \leqslant 1, \xi_2 \geq 0. \tag{4}$$

where $\boldsymbol{W} \in \mathbb{R}^{d\times C}$ is the transformation matrix of the multi-class discriminative linear predictor and it is constrained by a conventional ridge regularization. Since the prediction loss of subsequences at different progress levels should contribute differently to the prediction, we introduce a $\boldsymbol{s}(n)$ to weight each regression loss. By denoting S as the diagonal matrix generated by $\boldsymbol{s}$, the prediction loss can be expressed in a matrix form as $||(\boldsymbol{W}^T\boldsymbol{X}_i - \boldsymbol{y}_i\boldsymbol{\alpha}^T)\boldsymbol{S}||_{1,2}$. The $L_{1,2}$ norm is used to measure the regression loss because it is robust to noise and outliers [42].

In the above formulation, we constrain $\boldsymbol{\alpha}(N) = \boldsymbol{\alpha}^T\boldsymbol{e}_N = 1$ to ensure that a strong label can be derived if the entire sequence is observed, where $\boldsymbol{e}_N$ is a binary vector with only the $N-th$ entry being 1. In addition, we also restrict all entries in $\boldsymbol{\alpha}$ within $[0, 1]$.

In order to make sure the variation of soft label is smooth, we further impose a consistency constraint on $\boldsymbol{\alpha}$ as follows:

$$\min_{\boldsymbol{W},\boldsymbol{\alpha}} \sum_{i=1}^{I}\sum_{n=1}^{N} \overbrace{||(\boldsymbol{W}^T\boldsymbol{X}_i - \boldsymbol{y}_i\boldsymbol{\alpha}^T)\boldsymbol{S}||_{1,2}}^{Prediction\ loss\ term} + \frac{\xi_1}{2} \overbrace{||\nabla\boldsymbol{\alpha}||_2^2}^{Consistency\ term} + \frac{\xi_2}{2} \overbrace{||\boldsymbol{W}||_F^2}^{Regularization\ term}$$
$$s.t. \quad \boldsymbol{\alpha}^T\boldsymbol{e}_N = 1, 0 \leqslant \boldsymbol{\alpha} \leqslant 1. \xi_1, \xi_2 \geq 0. \tag{5}$$

Consistency term $||\nabla\boldsymbol{\alpha}||_2^2$. This constraint is to enforce the label consistency between subsequences. We compute the gradient of soft labels and measure its norm in order to control the variations of soft labels between subsequences. The effect of the consistency term is controlled by ξ_1. As the gradient operator $\nabla\boldsymbol{\alpha}$ is a linear operator, the consistency term can be rewritten in a matrix form $\boldsymbol{G}\boldsymbol{\alpha}$ equivalently, where we set $\boldsymbol{G}$ as $\begin{pmatrix} 1 & -1 & 0 & 0 \\ 0 & \ddots & \ddots & 0 \\ 0 & 0 & 1 & -1 \end{pmatrix} \in \mathbb{R}^{N-1\times N}$. In this way, we can rewrite our soft regression model as follows:

$$\min_{\boldsymbol{W},\boldsymbol{\alpha}} \sum_{i=1}^{I} ||(\boldsymbol{W}^T\boldsymbol{X}_i - \boldsymbol{y}_i\boldsymbol{\alpha}^T)\boldsymbol{S}||_{1,2} + \frac{\xi_1}{2}||\boldsymbol{G}\boldsymbol{\alpha}||_2^2 + \frac{\xi_2}{2}||\boldsymbol{W}||_F^2$$
$$s.t. \quad \boldsymbol{\alpha}^T\boldsymbol{e}_N = 1, 0 \leqslant \boldsymbol{\alpha} \leqslant 1, \xi_1, \xi_2 \geq 0. \tag{6}$$

3.4 Model Optimization

We solve our soft regression model (6) using a coordinate descent algorithm which would optimize over one parameter at each step while holding the others fixed. The optimization is achieved by iterating over the following two steps. At step 1, we optimize the predictor $\boldsymbol{W}$ with $\boldsymbol{\alpha}$ fixed. At step 2, we optimize it over $\boldsymbol{\alpha}$ with $\boldsymbol{W}$ fixed. The objective function (6) can be monotonically decreased with a guaranteed convergence. We provide some details in the following.

STEP 1. For fixed soft labels $\boldsymbol{\alpha}$, optimize the predictor $\boldsymbol{W}$:

$$\min_{\boldsymbol{W}} \sum_{i=1}^{I} ||(\boldsymbol{W}^T\boldsymbol{X}_i - \boldsymbol{y}_i\boldsymbol{\alpha}^T)\boldsymbol{S}||_{1,2} + \frac{\xi_2}{2}||\boldsymbol{W}||_F^2. \tag{7}$$

This is an unconstrained optimization problem and we can solve it with a standard gradient descent method. Let us denote the matrix $(\boldsymbol{W}^T\boldsymbol{X}_i - \boldsymbol{y}_i\boldsymbol{\alpha}^T)\boldsymbol{S}$ as $\boldsymbol{M}_i$. Then the gradient of the objective function with respect to $\boldsymbol{W}$ can be given by $\boldsymbol{G} = \sum_{i=1}^{i=I} \boldsymbol{X}_i\boldsymbol{S}\pounds(\boldsymbol{M}_i) + \xi_2\boldsymbol{W}$, where $\pounds$ is the column-wise normalization operator defined previously.

STEP 2. For fixed predictor $\boldsymbol{W}$, optimize the soft labels $\boldsymbol{\alpha}$:

$$\min_{\boldsymbol{\alpha}} \sum_{i=1}^{I} ||(\boldsymbol{W}^T\boldsymbol{X}_i - \boldsymbol{y}_i\boldsymbol{\alpha}^T)\boldsymbol{S}||_{1,2} + \frac{\xi_1}{2}||\boldsymbol{G}\boldsymbol{\alpha}||_2^2 \tag{8}$$
$$s.t. \quad \boldsymbol{\alpha}^T\boldsymbol{e}_N = 1, 0 \leqslant \boldsymbol{\alpha} \leqslant 1. \tag{9}$$

It is hard to directly solve the above problem because the existence of the sparse constraint in the prediction term and the bounded constraints in Eq. (9). Here, we introduce a method to find an approximate solution based on the popularly used projected gradient descent method. Firstly, we optimize the following problem without any constraint

$$\min_{\boldsymbol{\alpha}} \sum_{i=1}^{I} ||(\boldsymbol{W}^T\boldsymbol{X}_i - \boldsymbol{y}_i\boldsymbol{\alpha}^T)\boldsymbol{S}||_{1,2} + \frac{\xi_1}{2}||\boldsymbol{G}\boldsymbol{\alpha}||_2^2. \tag{10}$$

The above unconstrained problem can be optimized using a gradient descent based method. Specially, given the t-th step estimator $\boldsymbol{\alpha}_t$, the new updated point can be obtained by projecting $\boldsymbol{\alpha}_t - \tau\boldsymbol{g}_t$ into the feasible solution space $\{\boldsymbol{\alpha} \in \mathbb{R}^N | 0 \leq \boldsymbol{\alpha} \leq 1, \boldsymbol{\alpha}^T\boldsymbol{e}_N = 1\}$. The gradient $\boldsymbol{g}_t$ is given by $\sum_{i=1}^{i=I} \boldsymbol{y}_i^T\pounds(\boldsymbol{W}^T\boldsymbol{X}_i\boldsymbol{S} - \boldsymbol{y}_i\boldsymbol{\alpha}_t\boldsymbol{S})\boldsymbol{S} + \xi_1\boldsymbol{G}\boldsymbol{\alpha}_t$. Here τ is the iteration step size and an optimal step size is determined by a line search method within each iteration in order to monotonically decrease the objective function (10).

3.5 Prediction

Given a probe ongoing activity sequence (the progress level is unknown), we first extracted the corresponding LAFF feature $\boldsymbol{x}$ using the online constructed integral map $\boldsymbol{I}$. Then the prediction was made by finding the label that has the maximum score in $\boldsymbol{W}^T\boldsymbol{x}$. Our method can predict ongoing activity without given the progress level.

4 Experiments

We evaluated our methods on two benchmark 3D activity datasets: *Online RGB-D Action* dataset [5] and *SYSU 3DHOI* dataset [11].

4.1 Compared Methods and Implementation

We have implemented the following approaches using the same proposed features (i.e. LAFF) for comparison:

SVM on the Finished Activities (SVM-FA). As the simplest baseline, it trains a generic activity classifier on the completely executed sequences and the partial activity subsequences were not used for the training. During the test phase, all the ongoing subsequences were predicted using the learnt activity classifier. Comparison to this baseline is to demonstrate that the subsequences containing partial activity executions are important for the predictor learning.

Brute-force Prediction using SVM (BPSVM). It learns an activity predictor from all the available subsequences. In this baseline, we assigned the label of each subsequence with the label of its full sequence. That means these labels were not soft labels as described in our methods. This baseline is introduced in order to show the benefits of using soft labels. We denote it as "BPSVM".

Multiple Stages SVM (MSSVM). We trained a SVM predictor on the sequences obtained at each progress level separately. While in the test phase, we followed the same assumption in [7] that the progress level of ongoing activity is known and thus we can directly make the prediction using the predictor specifically trained for that progress level. Although practical system is hard to have a chance to obtain the progress level of ongoing activity sequence until the activity has been completely executed, it still serves as a good reference for evaluating activity prediction models. We denote this baseline as "MSSVM".

Other related activity prediction methods. In addition to the above comparison, we also compared the state-of-the-art activity prediction systems developed for RGB-D sequences [5] and RGB video clips [6,8].

For implementation of our proposed model, the regularization parameters ξ_1 and ξ_2 were set as 5000 and 1 throughout all the experiments, respectively. The total subsequence number N was set as 40. The weight $s(n)$ can be understood as prior weighting on the regression loss of the $n-th$ subsequence in Eq. (5). In general the loss of the subsequence at the end of the sequence is more important as the type of action becomes more clear. Hence, we increased $s(n)$ in Eq. (5) from 0.25 to 1 uniformly. Its influence will be studied in Sect. 4.4.

4.2 Results on Online RGB-D Action Dataset

The Online RGB-D Action Dataset (ORGBD) was collected for online activity recognition and activity prediction [5]. Some activity examples are shown in Fig. 3. For evaluation, we used the same-environment evaluation protocol detailed in [5], where half of the subjects were used for training a predictor and the rest were used for testing. In this setting, there are totally 224 RGB-D sequences of sixteen subjects, including seven human-object interaction activities (*drinking*, *eating*, *using laptop*, *reading cell phone*, *making phone call*, *reading book* and *using remote*). The mean accuracies were computed via a 2-fold validation.

Fig. 3. Some examples from the ORGBD and SYSU 3DHOI sets. The first and last two rows present samples from ORGBD and SYSU 3DHOI set, respectively.

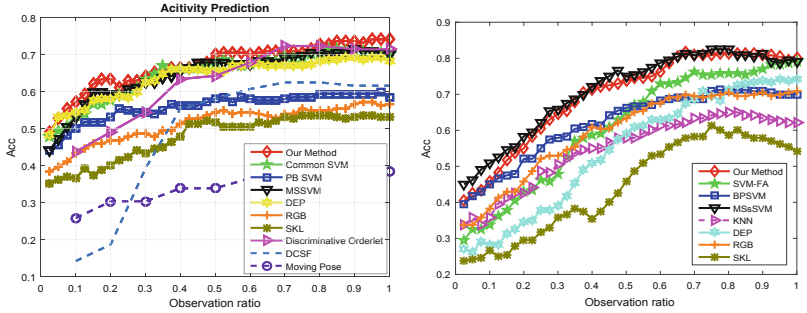


Fig. 4. Comparison results on ORGBD (left) and SYSU 3DHOI (right) sets.

We compared with the baselines and other related prediction methods as described in Sect. 4.1. The results are presented in Fig. 4 and Table 1. As shown, our method can produce better prediction results at most of the observation ratios than the competitors. We also find that the performance gap became larger if fewer activity frames were observed. This is as expected because our soft regression model explicitly makes use of the subsequences that contains partial

Table 1. Prediction (%) on ORGBD set. The last row provides the mean accuracies over all the observation ratios. "Ob. ratio", "MP", "DCSF", and "DO" denote observation ratio, "Moving Pose [5,43]", "DSTIP+DCSF [5,44]", and "Discriminative Order-let [5]".

Ob. ratio	RGB	DEP	SKL	MSSVM	BPSVM	SVM-FA	MP	DCSF	DO	Our Method
10 %	43.3	**54.5**	36.6	53.1	54.5	53.6	25.9	14.3	43.8	**57.1**
60 %	54.5	**67.4**	50.5	**67.4**	64.7	**67.4**	33.9	55.4	63.4	**70.1**
100 %	56.7	68.3	53.1	70.1	66.1	70.1	38.4	61.6	**71.4**	**74.1**
Mean	51.3	63.9	47.8	64.7	61.8	**64.9**	34.3	49.5	63.0	**67.2**

activity executions for obtaining a reliable predictor. By carefully examining the comparisons of our method and the baselines BPSVM and SVM-FA, we find that the introduced soft label learning mechanism can significantly improve prediction performance, and it also outperformed MSSVM which predicts ongoing activities with known progress level using multiple pre-trained predictors.

In addition, we also compared the state-of-the-art prediction algorithms reported on this set [5,44]. The proposed method outperformed the state-of-the-art method (discriminative order-let model) [5] by a large margin (more than 10 percent) when only 10 % of the sequence were used. If full sequence was provided, our predictor still performed better and obtained an accuracy of 74.1 %, which is 2.7 % higher than the discriminative order-let model. This suggests that the long-duration motion information ignored by the frame-level prediction model [5] is very important for identifying human activities, especially at early activity stages where the observed still activity evidence (such as human pose and object appearance etc.) is not sufficient for accurate activity recognition.

4.3 Results on SYSU 3D Human-Object Interaction Set

The SYSU 3D Human-Object Interaction Set (SYSU 3DHOI Set)[4] consists of 12 different activities from 40 participants, including *playing with a cell-phone*, *calling with a cell-phone*, *mopping* and *sweeping* etc. Some snapshots of the activities are presented in Fig. 3. In this dataset, each activity involves a kind of human-object interaction, and the motions and manipulated objects by participant performing are similar among some activities [11]. For evaluation, we employed the cross-subject setting popularly used in RGB-D activity recognition. In particular, we trained our predictor and all compared methods using the samples performed by the first 20 subjects and then tested on the rest subjects.

The prediction comparison results are presented in the second column of Fig. 4 and Table 2. As shown, our predictor can obtain a good performance at most of the progress levels and it significantly outperformed BPSVM and SVM-FA. Since activities at the same progress level often contain more similar activity

[4] It can be downloaded from http://isee.sysu.edu.cn/~hujianfang/.

Table 2. Prediction (%) on SYSU 3DHOI set. "Ob. ratio" denotes the observation ratio. The last row provides the mean accuracies over all the observation ratios.

Ob. ratio	RGB	DEP	SKL	kNN	MSSVM	BPSVM	SVM-FA	Our Method
10 %	38.3	28.3	26.7	35.8	**50.8**	45.0	33.8	**45.8**
60 %	67.5	62.9	53.3	61.3	**78.8**	68.3	72.9	**76.3**
100 %	70.8	74.2	54.2	62.1	**79.2**	70.0	**79.2**	**80.0**
Mean	59.5	54.4	44.5	54.7	**71.1**	61.9	61.0	**69.6**

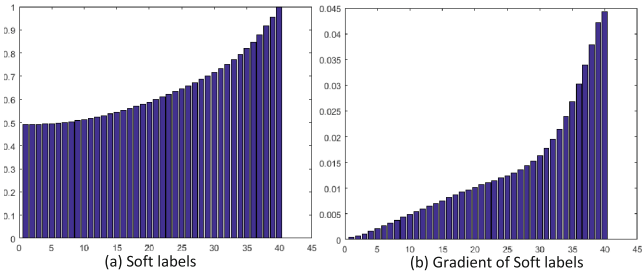


Fig. 5. Example soft labels learned on SYSU 3DHOI set. The vertical axis indicates the values for the soft labels and the horizontal axis is the index of subsequence.

context in this set, additionally using the progress level of an activity to train predictors is beneficial. So, MSSVM performs slightly better than our method at the early stages, but both perform comparably after that, and our method performs better when complete activity sequences were observed (i.e., the conventional activity recognition task). It again demonstrates that the generated large amount of unfinished subsequences can be used to benefit the task of activity recognition.

The learnt soft labels are presented in Fig. 5. We can observe that the soft labels starts around 0.5. This is as expected because some activities can be easily recognized at early stages by activity context (e.g., shapes and textures of objects). In general, the soft labels increase as more about the action is observed.

4.4 More Evaluations

RGB vs. RGB-D prediction. Intuitively, the RGB-D activity prediction can be casted as a RGB activity prediction problem by discarding the depth and skeleton modalities and RGB activity prediction methods can be easily implemented. Here, we tabulated the results on the SYSU 3DHOI set obtained by methods (DBOW [8], SC, and MSSC [6]) developed in [6][5] as well as our method in Table 3. As shown, our soft regression model have a significant advantage over these methods even using the same input data (RGB data). From the results, we can conclude that a RGB-D based prediction system has its unique benefit.

[5] The original codes are downloaded from http://www.visioncao.com/index.html.

Table 3. Comparisons of our method with RGB activity prediction methods.

Observation ratio	10 %	20 %	30 %	40 %	50 %	60 %	70 %	80 %	90 %	100 %	Mean
DBOW [8]	31.7	40.0	43.8	46.7	52.1	54.2	58.8	59.6	62.1	62.5	51.1
SC [6]	30.4	41.3	50.8	53.3	57.1	57.9	57.9	58.8	60.4	61.3	52.9
MSSC [6]	30.4	40.8	47.1	55.0	56.7	59.6	57.5	60.8	62.1	62.9	53.3
Our Method (RGB)	38.3	45.8	52.9	60.8	63.3	67.5	69.6	70.4	70.4	70.8	61.0
Our Method(RGB-D)	45.8	55.0	64.6	71.3	73.8	76.3	80.8	81.3	80.8	80.0	70.9

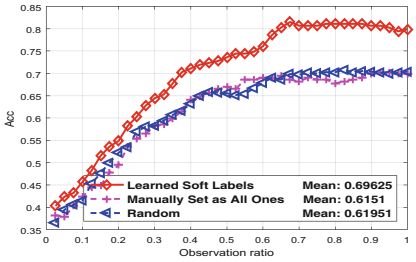
Evaluation on the elements used in the RGB-D. Results in Tables 1 and 2 and Fig. 4 show that the predictor learned from the combination of RGB, depth and skeleton channels is better than only using one of them. This is reasonable because RGB, depth and skeleton sequences indeed characterize activities from different aspects, and any single channel is intrinsically limited in overcoming the inherent visual ambiguity caused by human (object) appearance changes, cluttered background, view variation, occlusions and etc.

Benefits of learning soft labels $\boldsymbol{\alpha}$. For comparison, we implemented two baselines, where different strategies were employed to manually determine the soft labels: (1) we set all the elements of $\boldsymbol{\alpha}$ as 1; and (2) we randomly generated a set of soft labels for our regression model with 20 replicates. As shown in Fig. 6(a), the prediction accuracy decreased a lot if the soft labels were simply defined as the label of the whole sequence or randomly generated.

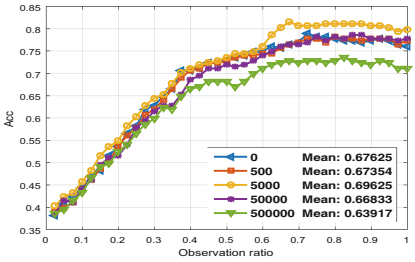
The influence of consistency constraint. We studied the influence of the parameter ξ_1, which is employed to control the effect of the consistency term in our model (5). Figure 6(b) shows the performances of setting ξ_1 as 0, 500, 5000, 50000 and 500000, respectively. As shown, our method can obtain a promising result with $\xi_1 = 5000$. In general, a small or large ξ_1 would lead to a lower prediction accuracy. Especially, when ξ_1 is larger than a certain number (e.g. 500000), the soft labels learnt by our method became useless as all of its entries will be the same and thus an unreliable prediction result was obtained.

Impact of the $\boldsymbol{s}$. In our regression model (5), we use a vector $\boldsymbol{s}$ to control the contribution of the regression losses caused by the subsequences of different progress levels. Here, we tested its influence. In the evaluation, we considered five different settings for $\boldsymbol{s}$, which were presented in the Fig. 6(c). The prediction performance obtained by each setting was presented in Fig. 6(d) with the same color. As shown, an $\boldsymbol{s}$ with incremental items performed better than the constant or even diminishing ones. This is desirable, because the subsequences in the latter progress levels should contain more activity execution and thus can provide more strong information for our predictor learning.

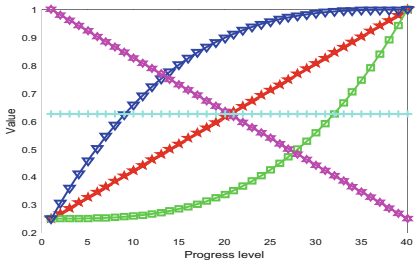
The convergence of the model. Our method converged to a minimum after a limited number of iterations. We empirically observed that 400 iterations were sufficient for obtaining a reliable solution in our experiments.



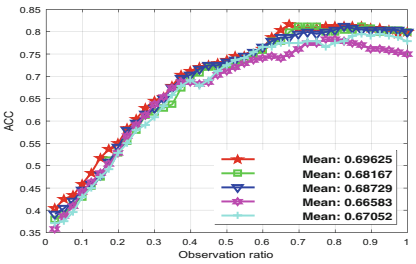
(a) Evaluation on learning soft labels.



(b) Evaluation on the consistency term.



(c) Visualization of used in the evaluation.



(d) Prediction performance for the presented in (c)

Fig. 6. More evaluations on the system performance.

The speed of prediction. We report the average speed (in fps) of the developed human activity prediction system. Our system can identify an activity of ongoing RGB-D sequences in real-time. Especially, it processed more than 40 frames per second using MATLAB on a normal desktop PC (CPU i5-4570), which is about 15 fps faster than the prediction system developed in [5].

5 Conclusions

We have developed a real-time RGB-D activity prediction system to identify ongoing activities under a regression framework. In such a regression framework, we learn soft labels for regression on subsequences containing partial activity executions so that it is not necessary to assume that the progress level of each subsequence is given. We learn both the soft labels and predictor jointly. In addition, a new RGB-D sequence feature called "local accumulative frame feature (LAFF)", which can be computed efficiently by constructing an integral feature map, is designed to characterize activity contexts. We demonstrate the effectiveness of our approach on RGB-D activity prediction and show that depth information is important for achieving much more robust prediction performance.

Acknowledgment. This work was supported partially by the NSFC (No.61573387, 61472456, 61522115, 61661130157), the GuangDong Program (No.2015B010105005), the Guangdong Science and Technology Planning Project (No.2016A010102012), and Guangdong Program for Support of Top-notch Young Professionals (No.2014TQ01X779).

References

1. Koppula, H.S., Saxena, A.: Anticipating human activities using object affordances for reactive robotic response. IEEE Trans. Pattern Anal. Mach. Intell. **38**(1), 14–29 (2016)
2. Koppula, H.S., Gupta, R., Saxena, A.: Learning human activities and object affordances from RGB-D videos. Int. J. Robot. Res. **32**(8), 951–970 (2013)
3. Lan, T., Chen, T.-C., Savarese, S.: A hierarchical representation for future action prediction. In: Fleet, D., Pajdla, T., Schiele, B., Tuytelaars, T. (eds.) ECCV 2014. LNCS, vol. 8691, pp. 689–704. Springer, Heidelberg (2014). doi:10.1007/978-3-319-10578-9_45
4. Li, K., Fu, Y.: Prediction of human activity by discovering temporal sequence patterns. IEEE Trans. Pattern Anal. Mach. Intell. **36**(8), 1644–1657 (2014)
5. Yu, G., Liu, Z., Yuan, J.: Discriminative orderlet mining for real-time recognition of human-object interaction. In: Cremers, D., Reid, I., Saito, H., Yang, M.-H. (eds.) ACCV 2014. LNCS, vol. 9007, pp. 50–65. Springer, Heidelberg (2015). doi:10.1007/978-3-319-16814-2_4
6. Cao, Y., Barrett, D., Barbu, A., Narayanaswamy, S., Yu, H., Michaux, A., Lin, Y., Dickinson, S., Siskind, J., Wang, S.: Recognize human activities from partially observed videos. In: IEEE Conference on Computer Vision and Pattern Recognition, pp. 2658–2665 (2013)
7. Kong, Y., Kit, D., Fu, Y.: A discriminative model with multiple temporal scales for action prediction. In: Fleet, D., Pajdla, T., Schiele, B., Tuytelaars, T. (eds.) ECCV 2014. LNCS, vol. 8693, pp. 596–611. Springer, Heidelberg (2014). doi:10.1007/978-3-319-10602-1_39
8. Ryoo, M.: Human activity prediction: Early recognition of ongoing activities from streaming videos. In: International Conference on Computer Vision, pp. 1036–1043 (2011)
9. Xu, Z., Qing, L., Miao, J.: Activity auto-completion: predicting human activities from partial videos. In: IEEE International Conference on Computer Vision, pp. 3191–3199 (2015)
10. Wang, J., Liu, Z., Wu, Y., Yuan, J.: Learning actionlet ensemble for 3D human action recognition. IEEE Trans. Pattern Anal. Mach. Intell. **20**(11), 914 (2013)
11. Hu, J.F., Zheng, W.S., Lai, J., Zhang, J.: Jointly learning heterogeneous features for RGB-D activity recognition. In: IEEE International Conference on Computer Vision and Pattern Recognition, pp. 5344–5352 (2015)
12. Hoai, M., De la Torre, F.: Max-margin early event detectors. Int. J. Comput. Vis. **107**(2), 191–202 (2014)
13. Kitani, K.M., Ziebart, B.D., Bagnell, J.A., Hebert, M.: Activity forecasting. In: Fitzgibbon, A., Lazebnik, S., Perona, P., Sato, Y., Schmid, C. (eds.) ECCV 2012. LNCS, vol. 7575, pp. 201–214. Springer, Heidelberg (2012). doi:10.1007/978-3-642-33765-9_15
14. Vondrick, C., Pirsiavash, H., Torralba, A.: Anticipating the future by watching unlabeled video (2015). arXiv preprint arXiv:1504.08023
15. Huang, D., Yao, S., Wang, Y., Torre, F.: Sequential max-margin event detectors. In: Fleet, D., Pajdla, T., Schiele, B., Tuytelaars, T. (eds.) ECCV 2014. LNCS, vol. 8691, pp. 410–424. Springer, Heidelberg (2014). doi:10.1007/978-3-319-10578-9_27
16. Dollár, P., Rabaud, V., Cottrell, G., Belongie, S.: Behavior recognition via sparse spatio-temporal features. In: IEEE International Workshop on Visual Surveillance and Performance Evaluation of Tracking and Surveillance, pp. 65–72 (2005)

17. Bregonzio, M., Gong, S., Xiang, T.: Recognising action as clouds of space-time interest points. In: IEEE Conference on Computer Vision and Pattern Recognition 2009, pp. 1948–1955 (2009)
18. Klaser, A., Marszałek, M., Schmid, C.: A spatio-temporal descriptor based on 3D-gradients. In: British Machine Vision Conference, pp. 275–281. British Machine Vision Association (2008)
19. Wang, H., Schmid, C.: Action recognition with improved trajectories. In: IEEE International Conference on Computer Vision, pp. 3551–3558 (2013)
20. Simonyan, K., Zisserman, A.: Two-stream convolutional networks for action recognition in videos. In: Proceedings of Advances in Neural Information Processing Systems, pp. 568–576 (2014)
21. Yang, X., Tian, Y.L.: Action recognition using super sparse coding vector with spatio-temporal awareness. In: Fleet, D., Pajdla, T., Schiele, B., Tuytelaars, T. (eds.) ECCV 2014. LNCS, vol. 8690, pp. 727–741. Springer, Heidelberg (2014). doi:10.1007/978-3-319-10605-2_47
22. Zhu, J., Wang, B., Yang, X., Zhang, W., Tu, Z.: Action recognition with actons. In: Proceedings of the IEEE International Conference on Computer Vision, pp. 3559–3566 (2013)
23. Hu, J.F., Zheng, W.S., Lai, J., Gong, S., Xiang, T.: Exemplar-based recognition of human-object interactions. IEEE Trans. Circ. Syst. Video Technol. **26**(4), 647–660 (2016)
24. Hu, J.F., Zheng, W.S., Lai, J., Gong, S., Xiang, T.: Recognising human-object interaction via exemplar based modelling. In: International Conference on Computer Vision (2013)
25. Oreifej, O., Liu, Z.: Hon4d: histogram of oriented 4D normals for activity recognition from depth sequences. In: IEEE International Conference on Computer Vision and Pattern Recognition, pp. 716–723 (2013)
26. Yang, X., Tian, Y.: Super normal vector for activity recognition using depth sequences. In: IEEE International Conference on Computer Vision and Pattern Recognition, pp. 804–811 (2014)
27. Dalal, N., Triggs, B.: Histograms of oriented gradients for human detection. In: IEEE International Conference on Computer Vision and Pattern Recognition, vol. 1, pp. 886–893 (2005)
28. Lu, C., Jia, J., Tang, C.K.: Range-sample depth feature for action recognition. In: IEEE International Conference on Computer Vision and Pattern Recognition, pp. 772–779 (2014)
29. Wang, J., Liu, Z., Chorowski, J., Chen, Z., Wu, Y.: Robust 3D action recognition with random occupancy patterns. In: Fitzgibbon, A., Lazebnik, S., Perona, P., Sato, Y., Schmid, C. (eds.) ECCV 2012. LNCS, vol. 7573, pp. 872–885. Springer, Heidelberg (2012)
30. Shotton, J., Sharp, T., Kipman, A., Fitzgibbon, A., Finocchio, M., Blake, A., Cook, M., Moore, R.: Real-time human pose recognition in parts from single depth images. Commun. ACM **56**(1), 116–124 (2013)
31. Hussein, M.E., Torki, M., Gowayyed, M.A., El-Saban, M.: Human action recognition using a temporal hierarchy of covariance descriptors on 3D joint locations. IJCAI **13**, 2466–2472 (2013)
32. Xia, L., Chen, C.C., Aggarwal, J.: View invariant human action recognition using histograms of 3D joints. In: IEEE International Conference on Computer Vision and Pattern RecognitionW, pp. 20–27 (2012)

33. Du, Y., Wang, W., Wang, L.: Hierarchical recurrent neural network for skeleton based action recognition. In: IEEE International Conference on Computer Vision and Pattern Recognition, pp. 1110–1118 (2015)
34. Yang, X., Tian, Y.: Eigenjoints-based action recognition using naive-bayes-nearest-neighbor. In: IEEE International Conference on Computer Vision and Pattern RecognitionW, pp. 14–19 (2012)
35. Ofli, F., Chaudhry, R., Kurillo, G., Vidal, R., Bajcsy, R.: Sequence of the most informative joints (SMIJ): a new representation for human skeletal action recognition. JVCIR **25**(1), 24–38 (2014)
36. Lillo, I., Soto, A., Niebles, J.C.: Discriminative hierarchical modeling of spatio-temporally composable human activities. In: IEEE International Conference on Computer Vision and Pattern Recognition, pp. 812–819 (2014)
37. Zanfir, M., Leordeanu, M., Sminchisescu, C.: The moving pose: An efficient 3D kinematics descriptor for low-latency action recognition and detection. In: IEEE International Conference on Computer Vision, pp. 2752–2759 (2013)
38. Shahroudy, A., Ng, T.T., Yang, Q., Wang, G.: Multimodal multipart learning for action recognition in depth videos. IEEE Trans. Pattern Anal. Mach. Intell
39. Liu, J., Shahroudy, A., Xu, D., Wang, G.: Spatio-temporal lstm with trust gates for 3D human action recognition. In: European Conference on Computer Vision (2016)
40. Wei, P., Zhao, Y., Zheng, N., Zhu, S.C.: Modeling 4D human-object interactions for event and object recognition. In: International Conference on Computer Vision, pp. 3272–3279 (2013)
41. Hu, N., Lou, Z., Englebienne, G., Kröse, B.: Learning to recognize human activities from soft labeled data. In: Proceedings of Robotics Science and Systems, Berkeley, USA (2014)
42. Li, Z., Liu, J., Tang, J., Lu, H.: Robust structured subspace learning for data representation. IEEE Trans. Pattern Anal. Mach. Intell. **37**(10), 2085–2098 (2015)
43. Gupta, A., Davis, L.S.: Objects in action: an approach for combining action understanding and object perception. In: IEEE Conference on Computer Vision and Pattern Recognition, pp. 1–8 (2007)
44. Xia, L., Aggarwal, J.: Spatio-temporal depth cuboid similarity feature for activity recognition using depth camera. In: IEEE International Conference on Computer Vision and Pattern Recognition, pp. 2834–2841 (2013)