Weng et al. EURASIP Journal on Image and Video Processing (2021) 2021:3
https://doi.org/10.1186/s13640-020-00544-0

EURASIP Journal on Image and Video Processing

RESEARCH Open Access

Human activity prediction using saliency-aware motion enhancement and weighted LSTM network

Zhengkui Weng*, Wuzhao Li and Zhipeng Jin

* Correspondence: zkweng19@163.com
Jiaxing Vocational and Technical College, Jiaxing, Zhejiang 314036, China

Abstract

In recent years, great progress has been made in recognizing human activities in complete image sequences. However, predicting human activity earlier in a video is still a challenging task. In this paper, a novel framework named weighted long short-term memory network (WLSTM) with saliency-aware motion enhancement (SME) is proposed for video activity prediction. First, a boundary-prior based motion segmentation method is introduced to use shortest geodesic distance in an undirected weighted graph. Next, a dynamic contrast segmentation strategy is proposed to segment the moving object in a complex environment. Then, the SME is constructed to enhance the moving object by suppressing irrelevant background in each frame. Moreover, an effective long-range attention mechanism is designed to further deal with the long-term dependency of complex non-periodic activities by automatically focusing more on the semantic critical frames instead of processing all sampled frames equally. Thus, the learned weights can highlight the discriminative frames and reduce the temporal redundancy. Finally, we evaluate our framework on the UT-Interaction and sub-JHMDB datasets. The experimental results show that WLSTM with SME statistically outperforms a number of state-of-the-art methods on both datasets.

Keywords: Activity prediction, Weighted long short-term memory network, Dynamic contrast segmentation, Saliency-aware motion enhancement

1 Introduction

At present, most of the researches in the field of activity recognition focus on how to recognize human activity in a complete image sequence [1–3]. However, in practical applications, people is more desirable that the intelligent system can warn of the potential risks in advance so as to stop dangerous acts before they cause serious damage, rather than just recognizing the dangerous activity or detecting the damage caused by it. So, the activity prediction task aims to recognition human activities with less video frames. In a broad sense, the prediction task can be seen as a human activity recognition task with limited observed data. Although great progress has been made in recognizing activities in complete image sequences [4–6], video-based human activity prediction in the early stage is still a challenging task.

Different from the static image classification task, one of the distinctive properties of video-based activity prediction is the dynamic viewpoint and messy background which generally existing in video data. The background in a video may contain irrelevant motion other than the foreground objects that we are actually interested in. Thus, to engage the foreground motion features deeply inside the learning process, Majd and Safabakhsh [7] proposed a motion-aware ConvLSTM network for action recognition. A multiple short-time motion energy image was proposed in [6] to capture human motion information, which helps the CNNs to learn hierarchical local motion features from the input image, but the effectiveness of motion-aware module depends on the video data and the quality of model training. There are also some existing activity representation approaches aim to diminish the interference of various visual occurrences by detecting space-time interest points [8, 9], generating motion trajectories [10], or segmenting the moving object [11] before the motion features are fed into convolutional neural network (CNN). However, these methods suffer from two common drawbacks. First, some subtle motion with key information may be lost if the threshold is oversensitive. Second, it is not easy to extract discriminative features from the selected action-related regions because these foreground regions may not necessarily be spatially coherent [12].

In addition, another unique characteristic of video is the variable-length temporal dimension. Without the guidance of high-level semantic information, the activity representations extracted by these approaches cannot selectively express the most relevant frames in a video [13, 14]. In other words, the methods aforementioned deal with the frame-level features equally, which may unavoidably increase the redundant noise from irrelevant frames. To address the problem, Ryoo [15] proposed a video early activity prediction method based on a Dynamic Bag of Words model which uses encoded features to model the characteristics of early observation video. However, the temporal feature encoding depends on setting parameters manually, so the prediction performance is difficult to be guaranteed. Du et al. [16] proposed a recurrent network framework based on pose attention for human activity prediction. This was the first time that human pose was integrated into a recurrent network and future activity was predicted by the superposition of multiple long short-term memory (LSTM) units. Although the estimation of human posture improves the prediction accuracy, it is hard to ensure the accurate of pose estimation in 2D video data, which is not helpful to the temporal representation of video. Wang et al. [17] extended the generalized time warping (GTW) algorithm by adding temporal constraints over the warping path to encourage the matching in the early portion of an activity, but this matching approach is not sensitive to some periodic human activities like “comb the hair.” Aliakbarian et al. [18] developed a multistage LSTM human activity prediction framework. This framework introduced an action-aware module, and constructed a new loss function to encourage the model to predict the correct activity categories as early as possible, which demonstrate the effectiveness of LSTM for activity prediction. Coincidentally, Lan et al. [19] proposed a hierarchical movemes approach to describe human motion rules at multiple levels by considering the hierarchical characteristics of human activities. Sun et al. [20] investigated the motion map 3D ConvNet to learn a motion map for representing an action video clip, and a discrimination network is also introduced for classifying actions based on the learned motion map.

Recent work in activity prediction has shown benefits of exploiting advanced deep learning based structure. Sun et al [21] developed a graph based relational network to jointly model temporal and spatial interactions among different actors. Wang et al [22] presented a teacher-student learning framework for early action prediction, which achieves knowledge distillation by minimizing the local progressive-wise and global distribution knowledge discrepancy. But how to mine as much action knowledge as possible from the teacher model is one of the major challenges in such framework. Zhao and Wildes [23] proposed a Kalman filter mechanism to ameliorate error accumulation over time, which effectively model the temporal characteristics of activities. From the above analysis, it can be seen that, similar to activity recognition, how to extract the spatial and temporal information from video is also a crucial problem in activity prediction. A sequence-to-sequence framework named RU (rolling-unrolling)-LSTM was proposed in [24], where a multi-modal framework based on LSTM networks to anticipate future actions which is able to summarize past observations while making predictions of the future at different time steps. However, without the guidance of high-level semantic information, the action representations extracted by these models cannot selectively express the most relevant frames in a video.

In response to the aforementioned challenges, we focus on exploring robust spatio-temporal feature and proposing a novel framework, named the weighted long short-term memory (WLSTM) network with saliency-aware motion enhancement (SME) for spatial and temporal modeling. First, a boundary-prior based moving object segmentation method is introduced by combining the shortest geodesic distance with a dynamic contrast segmentation strategy. Then, the SME is constructed to enhance the moving object by suppressing the irrelevant background motion in each frame. In order to further deal with the long-term dependency in complex non-periodic actions, the WLSTM is developed by employing an effective long-range attention mechanism so as to highlight the discriminative frames and suppress the temporal redundancy. As shown in Fig. 1, two modalities which have both been enhanced are fed into WLSTM, i.e., RGB-SME and OF (optical flow)-SME. Both streams of the WLSTM can be trained with stochastic gradient descent (SGD). The proposed WLSTM with SME is extensively validated on two challenging video activity datasets, UT-Interaction [25] and sub-JHMDB [26]. The effectiveness of our framework is proved when compared with state-of-the-art methods. Furthermore, to illustrate the effectiveness of each part of our work, extensive studies are performed to visualize the intermediate processes of WLSTM and SME.

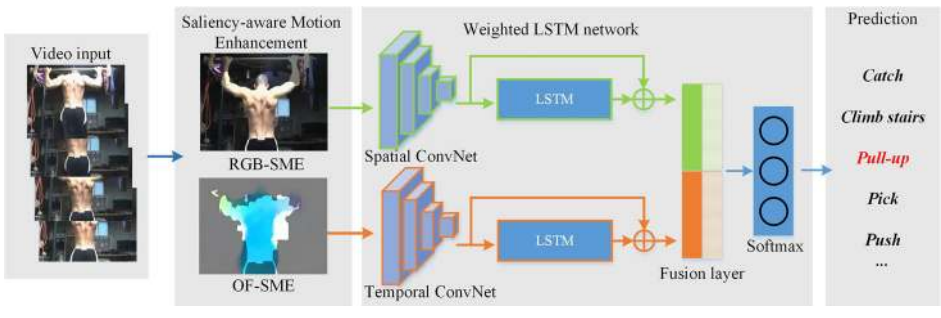


Fig. 1 The overall of the proposed human activity prediction framework

The main contributions of this paper can be summarized as follows:

- A totally automatic saliency-aware motion enhancement (SME) approach is proposed to enhance the motion object from complicated background via a novel boundary prior based motion object segmentation method. It can further guide the learning procedure of our framework, with more focus on the important motion regions where human actions are most likely to occur.
- This paper develops a novel deep learning framework for video activity prediction named weight LSTM (WLSTM). Instead of processing all the frames equally, with a simple but effective long-range attention mechanism, the WLSTM focuses more heavily on the semantic key frames and adaptively eliminating temporal redundancy.

The remainder of this paper is organized as follows. In Section 2, the saliency-aware motion enhancement (SME) and the weighted long short-term memory (WLSTM) are presented. The experimental results are discussed in Section 3. Finally, the conclusion is drawn in Section 4.

2 Method

2.1 Saliency-aware motion enhancement

Figure 2 shows an overview of the proposed saliency-aware motion enhancement. First, each input frame is segmented into superpixels. An undirected weighted graph in which every superpixel is regarded as one node is constructed in each segmented frame, and the relationship between superpixels is presented as the edge between two nodes. In order to extract both appearance and motion information from the action-related region, two types of the edge (static and motion) are extracted from each frame. We observe that the moving objects are always surrounded by the superpixels with large spatiotemporal value of edge. Thus, we define the saliency probability as the shortest geodesic distance from each superpixel to the frame boundary.

Next, an adaptive threshold is designed to gain the coarse saliency map by dividing the superpixels into background superpixel set and foreground superpixel set. Meanwhile, a dynamic contrast segmentation strategy is introduced to obtain the accurate motion regions by computing the geodesic distance of every two superpixels between background set and foreground set. For those foreground superpixels with larger mean

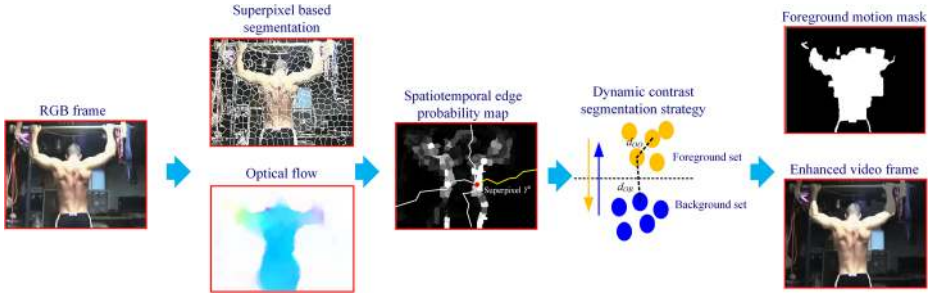


Fig. 2 Overview of saliency-aware motion enhancement

geodesic distance from the background superpixels, we regard them as pseudo foreground superpixels then put them into the background set, and vice versa. Finally, the background in each frame is suppressed by halving the brightness of pixels which are out of the accurate motion region.

2.1.1 Spatio-temporal edge generation

Given a frame sequence $F = \{F^1, F^2, \ldots, F^k\}$, $P_i^k = (x_i^k, y_i^k)$ denotes the ith pixel in frame F^k, where x_i^k and y_i^k represent the abscissa and ordinate of pixel P_i^k, respectively. Considering the experimental performance and computational efficiency, we calculate the spatial edge probability map $\hat{E}_{static}^k(P_i^k) = \hat{E}_{static}^k(x_i^k, y_i^k)$ corresponding to the kth frame F^k and pixel P_i^k using Canny edge operator. Let $\omega_i^k = (u_i^k, v_i^k)$ be the optical flow of the ith pixel in frame F^k, where u_i^k and v_i^k are the horizontal and vertical component, respectively. Then, we calculate the motion gradient $\hat{E}_{motion}^k(\omega_i^k)$ of the optical flow ω_i^k as

$$\hat{E}_{motion}^k(\omega_i^k) = \|\nabla\omega_i^k\| = \|\frac{\partial\omega_i^k}{\partial u_i^k} + \frac{\partial\omega_i^k}{\partial v_i^k}\| \tag{1}$$

Next, we segment the frame into superpixels by SLIC [27]. $Y^k = \{Y_1^k, Y_2^k, \ldots, Y_n^k\}$ denotes the superpixel set in frame F^k. Given the static edge map $\hat{E}_{static}^k$, the edge probability of the superpixel Y_n^k is calculated as the average value of the pixels with the ten largest edge probabilities in Y_n^k. Similarly, the motion gradient magnitude is also obtained using $\hat{E}_{motion}^k$. The above steps generate two superpixel-based edge probability maps, E_{static}^k and E_{motion}^k. Then, a spatiotemporal edge probability map $E^k(Y^k)$ can be obtained as the element-wise product of E_{static}^k and E_{motion}^k:

$$E^k(Y^k) = E_{static}^k \circ E_{motion}^k \tag{2}$$

The reason why we compute $E^k(Y^k)$ is that we consider both static and motion information which can provide useful information for moving object detection.

2.1.2 Geodesic distance based dynamic saliency estimation

For each frame F^k, an undirected weighted graph is constructed as $g^k = \{Y^k, e^k\}$ with superpixels Y^k as nodes and $e^k(m, n)$ as the edges between pairs of adjacent superpixels. Based on the constructed structure, a $|Y^k| \times |Y^k|$ adjacency matrix W^k of the graph is developed, which is defined as the element-wise product of $e^k(m, n)$ and w_{mn}^k, where w_{mn}^k denotes the weight between adjacent superpixels Y_m^k and Y_n^k:

$$W^k = e^k(m, n) \circ w_{mn}^k \tag{3}$$

$$w_{mn}^k = \|E^k(Y_m^k) - E^k(Y_n^k)\| \tag{4}$$

where $E^k(Y_m^k)$ and $E^k(Y_n^k)$ correspond to the spatiotemporal boundary probabilities of superpixels Y_m^k and Y_n^k, respectively. To emphasize the foreground moving objects that have high spatiotemporal edge values or are surrounded by regions with high spatiotemporal edge values, we employ the geodesic distance to compute a coarse object

probability map. The geodesic distance $d_g(Y_m^k, Y_n^k, g^k)$ between any two superpixels $Y_m^k, Y_n^k \in Y^k$ in graph g^k is defined as the cumulative weighted shortest path:

$$d_g\left(Y_m^k, Y_n^k, g^k\right) = \min\left\{\sum_{Y_m^k}^{Y_n^k} \left|W^k \cdot e^k(m,n)\right|\right\} \tag{5}$$

Based on the definition of the geodesic distance d_g, we can see that if a superpixel is outside the motion region, there possibly exists a path to the frame boundaries which does not pass through any superpixels. Thus, the geodesic distance of such a superpixel is relatively small. Conversely, supposing a superpixel is inside the motion region, the superpixel must be surrounded by superpixels with high spatiotemporal edge values, which will increase the geodesic distance to the frame boundaries. As such, the boundary prior saliency value S_n^k for each superpixel Y_n^k is computed by

$$S_n^k = \min_{q \in Q^k}\left\{d_g\left(Y_n^k, q, g^k\right)\right\} \tag{6}$$

where Q^k indicates the superpixels from the four boundaries in each frame F^k. All saliency values in S_n^k are normalized to [0, 1]. Based on the saliency valueS_n^k, a coarse saliency map can be obtained by using a self-adaptive threshold which divides the superpixels into a background set $\hat{B}^k$ and a foreground set $\hat{O}^k$. The self-adaptive threshold θ^k for each frame F^k is computed by

$$\theta^k = \mu\left(S_n^k\right) = \frac{1}{n}\sum_{i \in n} S_i^k \tag{7}$$

where $\mu(\cdot)$ represents the mean value of all superpixels within frame F^k by saliency value S_n^k. Then, the superpixels in each frame F^k can be cataloged into the background and foreground set:

$$\begin{aligned} \hat{O}^k &= \left\{Y_n^k \middle| S_n^k > \theta^k\right\} \\ \hat{B}^k &= Y^k - \hat{O}^k \end{aligned} \tag{8}$$

To obtain the refined foreground region, a dynamic contrast segmentation strategy is introduced by comparing both the inter and intra geodesic distances between the foreground and background sets. Three kinds of distances, namely, d_{OO}, d_{OB}, and d_{BB}, are utilized as the measurement for the superpixels $Y^k = \hat{O}^k \cup \hat{B}^k$ in each frame F^k:

$$\begin{aligned} d_{OO} &= \mu\left(\sum_{Y_n^k, Y_{n+1}^k \in \hat{O}^k}\left[d_g\left(Y_n^k, Y_{n+1}^k, g^k\right)\right]\right) \\ d_{OB} &= \mu\left(\sum_{Y_n^k \in \hat{O}^k, Y_{n+1}^k \in \hat{B}^k}\left[d_g\left(Y_n^k, Y_{n+1}^k, g^k\right)\right]\right) \\ d_{BB} &= \mu\left(\sum_{Y_n^k, Y_{n+1}^k \in \hat{B}^k}\left[d_g\left(Y_n^k, Y_{n+1}^k, g^k\right)\right]\right) \end{aligned} \tag{9}$$

where d_{OO} denotes the mean geodesic distance between two foreground superpixels. The higher the geodesic distance value, the closer the relationship between the

superpixels. Similarly, d_{OB} and d_{BB} are also adopted as the measurement to describe inter and intra relationship among superpixels in one frame. Thus, we define the refined motion object regions O^k as:

$$\begin{aligned} O^k &= \left\{ Y^k \in \hat{O}^k | d_{OO} > d_{OB} \right\} \cup \left\{ Y^k \in \hat{B}^k | d_{BB} \le d_{OB} \right\} \\ B^k &= Y^k - O^k \end{aligned} \tag{10}$$

The main rationale behind Eq. (10) is that both inter and intra differences between the two superpixel sets are taken into consideration in the proposed dynamic contrast segmentation strategy.

2.1.3 Moving object enhancement

Based on the classified superpixels Y^k in each frame F^k, the corresponding saliency map is computed by binarizing the values S_n^k in the coarse saliency map. If $S_n^k(i,j) \in B^k$, then $S_n^k(i,j) = 0$; otherwise, $S_n^k(i,j) = 1$. Next, we suppress the background pixels in F^k to improve the importance of foreground moving objects. We first transform F^k from the RGB color space to the HSI color space, halving the I component of background regions where $S_n^k(i,j) = 0$, then perform the opposite operation to reverse the frame back to the RGB color space. As a result, an SME-based frame can be obtained and is denoted RGB-SME. To construct the OF-SME, we halve the magnitude outside the saliency-aware region of the optical flow in each frame F^k. Figure 3 visualizes the saliency-aware motion enhancement intermediate procedures.

2.2 Weighted LSTM network

2.2.1 Long short-term memory unit

LSTM consists of a series of memory cells, each containing an internal state c_t, which stores information about the input sequence x_t from time 1 to time t. As shown in Fig. 4, these gates are activated by non-linear functions which enable LSTM to model human activity in dynamic environments at different time steps.

LSTM has the following gate control structure:

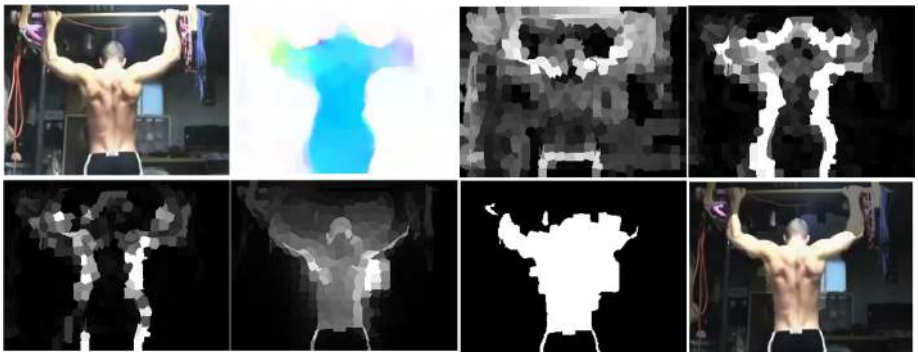


Fig. 3 Visualization of the saliency-aware motion enhancement intermediate procedures. From left to right in the first row: source video frame, optical flow, static edge map, and motion edge map. From left to right in the second row: spatio-temporal edge map, saliency map, foreground motion mask, and enhanced video frame

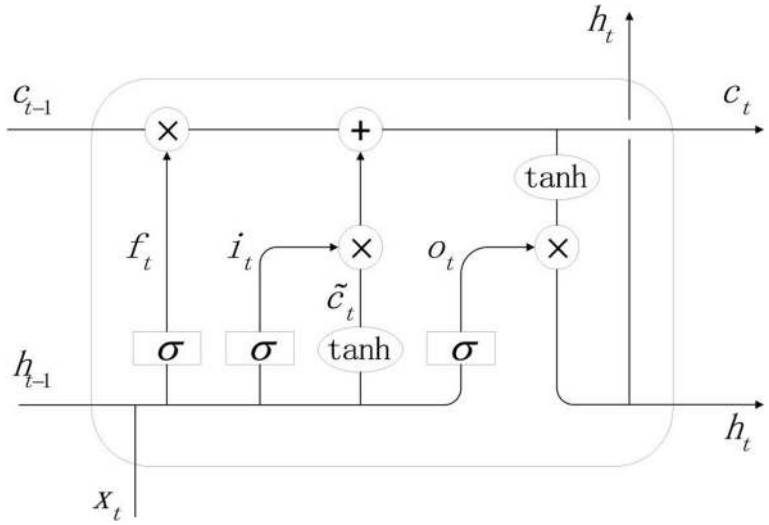


Fig. 4 Structure of an LSTM unit

1) Input gate i_t controls the degree of input information into memory cells to affect the state of the tth memory cell c_t:

$$i_t = \sigma(w_{xi}x_t + w_{hi}h_{t-1} + w_{ci}c_{t-1} + b_i) \tag{11}$$

where σ(·) is the sigmoid activation function; w_{xi}, w_{hi} , and $w_{c\ I}$ are weight matrices; and b_i is the bias.

2) Forget gate f_t regulates the previous state $c_{t\text{-}1}$ of memory cells to control the activation of the current state c_t:

$$f_t = \sigma\left(w_{xf}x_t + w_{hf}h_{t-1} + w_{cf}c_{t-1} + b_f\right) \tag{12}$$

where w_{xf}, w_{hf}, and w_{cf} are the weight matrices, and b_f is the bias. c_t is represented as

$$c_t = f_t c_{t-1} + i_t c_{t-1/2} \tag{13}$$

where $c_{t-1/2}$ is the pre-state of memory cells.

$$c_{t-1/2} = i_t \tanh(w_{xc}x_t + w_{hc}h_{t-1} + b_c) \tag{14}$$

where tanh(·) is the hyperbolic tangent activation function, w_{xc} and w_{hc} are weight matrices, and b_c is the bias.

3) Output gate o_t controls the output of information from memory cells which will affect the future state of LSTM memory cells:

$$o_t = \sigma(w_{xo}x_t + w_{ho}h_{t-1} + w_{co}c_t + b_o) \tag{15}$$

where w_{xo}, w_{ho}, and w_{co} are the weight matrices and b_o is the bias. Finally, the output of an LSTM unit can be represented as:

$$h_t = o_t \tanh(c_t) \tag{16}$$

2.2.2 Weighted LSTM for human activity prediction

In the training process of WLSTM, the complete features $\{(x_1, x_2, \ldots, x_T), y\}$ of the observable image sequence can be obtained from the base convolutional neural network. The goal of our work is to train the WLSTM to predict future activities with partially observable features $\{(x_1, x_2, \ldots, x_t), t < T\}$ from RGB-SME and OF-SME. Therefore, a sequence-to-sequence prediction approach is utilized in the training process of WLSTM. Given a training sample $\{(x_1, x_2, \ldots, x_T)j, Y_j\}_{j=1}^{N}$, a weighted LSTM unit is introduced to learn the complete image sequence features $(X_1, X_2, \ldots, X_T)$ with corresponding label $(y_1, y_2, \ldots, y_T)$. In this way, the WLSTM can predict the future activity label y for incomplete image sequences in the testing process. In addition, in the process of constructing the WLSTM, an effective weighted mechanism is proposed in this paper; it makes LSTM units pay more attention to key frames in image sequences and adaptively removes temporal redundancy so as to better solve the problem of long-term dependence of video frames in activity prediction. The overall framework of the proposed WLSTM is shown in Fig. 5.

No matter what the activity is, there are always some segments or video frames that are irrelevant, redundant, or confusing. The main content of these clips shifts from humans to some irrelevant moving objects which may interfere the model training. In this paper, a weighted LSTM frame is proposed to connect the inception with batch normalization (BN-Inception) [28] network, making video frames with different visual contents play different roles in the result of the activity prediction. The weights of two modalities are calculated by the output of BN-Inception x_t at time t and LSTM unit h^*_{t-1} at time t–1:

$$\alpha^*_t = \exp\left(\tanh\left(w_{x\alpha}x_t + w_{h\alpha}h^*_{t-1} + b_\alpha\right)\right) \tag{17}$$

where exp(·) is the exponential function, $w_{x\alpha}$ and $w_{h\alpha}$ are weight matrices, and b_α is the bias. Then, the input x_t' of WLSTM at time t is weighted by the basic feature x_t and contribution factor α^*_t.

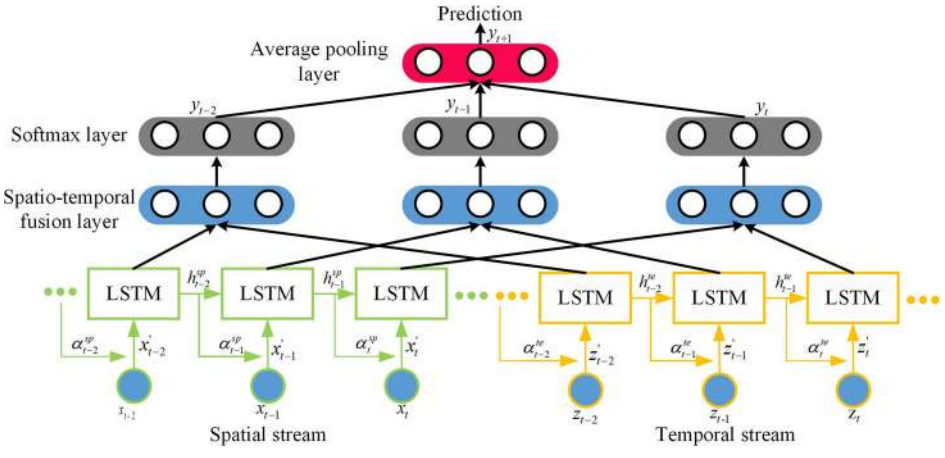


Fig. 5 The overall framework of the proposed weighted LSTM (WLSTM)

$$x_t' = \alpha_t^* x_t \tag{18}$$

In the process of multi-modality fusion, the weighted output of two modalities of the basic network BN-Inception is not concentrated directly $\{(x_1', x_2', ..., x_T'), (z_1', z_2', ..., z_T')\}$. Instead, a continuous temporal feature is constructed by an LSTM unit, and then the time-coded feature $\{(h_1^{sp}, h_2^{sp}, ..., h_T^{sp}), (h_1^{te}, h_2^{te}, ..., h_T^{te})\}$ is fed into a fully connected spatio-temporal feature fusion layer. Finally, the activity at this time is predicted by a softmax layer, which is as follows:

$$(h_t^{sp}, c_t^{sp}) = \mathrm{LSTM}\left(x_t', h_{t-1}^{sp}, c_{t-1}^{sp}\right) \tag{19}$$

$$\left(h_t^{te}, c_t^{te}\right) = \mathrm{LSTM}\left(z_t', h_{t-1}^{te}, c_{t-1}^{te}\right) \tag{20}$$

$$e_t = \tanh\left(W_f\left[h_t^{sp}; h_t^{te}\right] + b_f\right) \tag{21}$$

$$y_t = \mathrm{softmax}\left(W_y e_t + b_y\right) \tag{22}$$

where W_* is the weight matrix, b_* is a bias, c_t is the state vector of the memory cells at time t, and y_t is the activity label at time t. Finally, the video activity of frame t+1 is predicted:

$$y_{t+1} = \frac{1}{T}\sum_{i \le T} y_i. \tag{23}$$

In order to predict earlier the activities occurring in video, we add a time penalty term to the original cross-entropy function; this makes the network loss increase with increasing time in the training process, so the WLSTM can be guided to repair the training error as soon as possible:

$$loss = \sum_{j=1}^{N}\sum_{t=1}^{T} -e^{-(T-t)} \log\left(y_j^k\right) \tag{24}$$

where N is the number of samples and T is the video sampling time.

3 Experimental results and discussion

3.1 Datasets

In order to verify the effectiveness of the proposed video activity prediction framework, two challenging video datasets, UT-Interaction [25] and sub-JHMDB [26], were selected for experiments. Figure 6 shows some sampled frames in these datasets.

UT-Interaction is a human activity dataset, which can be further subdivided into UT-Interaction Set 1 and UT-Interaction Set 2. Two subdatasets both contain 60 videos including six kinds of activities: shake hands, hug, kick, point, punch, and push. In this paper, we used 10-fold leave-one-out cross validation per set to get our experimental results on the UT-Interaction dataset. Each 10-fold cross validation extracts one-tenth of all samples as the test set, and the remaining samples are used as the training set. The test sets are selected 10 times, and the samples are different each time.

The sub-JHMDB dataset collects video data from a wide range of sources and includes 316 videos showing 12 activities: catch, climb stairs, golf, jump, kick ball, pick, pull up, push, run, shoot ball, swing baseball, and walk. There are 19 to 42 video

Fig. 6 Some sampled frames from UT-Interaction and sub-JHMDB

samples for each activity, and the total number of samples is 316. Each video sample contains 15 to 40 frames of 320 × 240 pixels in size. Three kinds of train-test splits are given in the dataset, and the first train-test split was adopted for the experiment.

3.2 Implementation details

In the data preprocessing stage, the TVL1 optical flow algorithm [29] was adopted. Specifically, the optical flow was discretized into the range of [0, 255] to guarantee data consistency with RGB frames. Then, the proposed SME was implemented on both datasets to generate RGB-SME and OF-SME as the input of the base network. In the selection of a convolutional network for basic feature extraction, we chose the publicly available inception with batch normalization network (BN-Inception) as our base network from [28] because of its good performance in terms of efficiency and accuracy. During the training phase, we adopted the mini-batch SGD optimizer by setting the momentum to 0.9 and batch size to 128. For spatial modality (RGB-SME), we initialized the network using pre-trained models from ImageNet; the learning rate was first set to 0.001, and then reduced to 1/10 every 1500 iterations. The training process finally stopped after 3500 iterations. For temporal modality (OF-SME), we used the cross-modality pre-training strategy proposed in [30] by utilizing the learned spatial models to initialize the temporal stream. The learning rate was initialized to 0.005, which was reduced by a factor of 10 after 12,000 and 16,000 iterations. The maximum number of iterations was set to 18,000. For data augmentation, we utilized the random horizontal flipping and scale jittering technique to reduce the risk of overfitting. The WLSTM training parameters mainly included the batch size, learning rate, and optimizer. According to [30], we set the batch training size of the model to 128; the initial learning rate was 0.003, which was halved every 15,000 iterations; and the Adam optimizer was selected. In the WLSTM testing phase, each complete video in the above dataset was divided into 10 subvideos with different observation ratios using equal observation ratio increments. For different subvideos with different observation ratios, we set sample interval $T = 3$ for sampling and activity prediction. Finally, more importance was given to the temporal stream by setting its weight to 1.5 and that of the spatial stream to 1 for the fusion of two streams.

3.3 Experimental results analysis

As shown in Table 1, Table 2, and Table 3, the results of the WLSTM in predicting human activities are presented in this section. Three modalities are used to test the accuracy of the proposed activity prediction framework, namely, RGB-SME, OF-SME, and two-modality fusion with temporal and spatial information. The experimental results show that two-modality fusion as the input achieved superior prediction accuracy on all three datasets. The video prediction accuracy values with two-modality fusion on the UT-Interaction set 1 and UT-Interaction set 2 [25] datasets are 3.2% and 3.0% higher than that with single-input modality, respectively, and 5.7% higher than single input on the sub-JHMDB [26] dataset. It can also be seen from Table 1 and Table 2 that with only half the input video, the WLSTM in this paper achieves 95.0% and 90.2% prediction accuracy on UT-Interaction sets 1 and 2, respectively. This means that when the observation ratio is only 0.5, the prediction accuracy of the WLSTM can reach 95.3% of the complete video activity recognition accuracy, which is very encouraging. Figure 7 shows the activity prediction results for the three data sets when the test video clips were subvideos of different lengths. It is not difficult to see that with the increase of the subvideo observation ratio, the accuracy of activity prediction of the three modes was improved.

In Fig. 7, the accuracy of activity prediction on the three datasets with different subvideo lengths is presented. The video observation ratio interval of the subvideo is set to [0.1, 1]. It is not difficult to see that with the increase of the subvideo observation ratio, the accuracy of activity prediction of all three modalities is improved. The highest prediction performances achieved are 98.3%, 95.1%, and 78.1% on UT-Interaction sets 1 and 2 and sub-JHMDB, respectively. Although UT-Interaction sets 1 and 2 have identical activity categories, the prediction accuracy of UT-Interaction set 2 is slightly lower than that of set 1 because the background of UT-Interaction set 1 is a parking lot with constant light, which simplifies the background while eliminating the interference caused by illumination. Meanwhile, the background in set 1 is mostly static, with no camera jitter. UT-Interaction set 2, however, has a complex background of grassland and tree branches, including background moves (for example, trees move) and contain more camera jitter. So, the accuracy of optical flow is affected in set 2 due to the complex conditions which leads to a degradation in OF-SME performance. That is why OF-SME can outperform RGB-SME with 50% of the clips in set 1 where it need 70% in set 2.

Moreover, the WLSTM (two-modality fusion) proposed in this paper achieves about 50% prediction accuracy on all three datasets when the video observation ratio is only 0.1, which also shows the effectiveness of the WLSTM in video activity prediction tasks. In addition, it can be seen that RGB-SME has better robustness, and its prediction accuracy curve shows almost no change under different video observation ratios. Meanwhile, when the video observation ratio is less than 0.4, the prediction accuracy of the

Table 1 Prediction accuracy comparison of different modalities on UT-Interaction set 1 (%)

Modality	Accuracy with half videos	Accuracy with full videos
RGB-SME	87.6	93.8
OF-SME	93.1	96.4
Two-modality fusion	95.0	98.3

Table 2 Prediction accuracy comparison of different modalities on UT-Interaction set 2 (%)

Modality	Accuracy with half videos	Accuracy with full videos
RGB-SME	84.7	91.5
OF-SME	75.2	92.6
Two-modality fusion	90.2	95.1

OF-SME is poor, and with the increase of the video observation ratio, the prediction accuracy is greatly improved. We conjecture that this is because the optical flow reflects the motion information within adjacent frames. Under a low observational ratio, it is easy to misclassify activity because of the similarities between short-term actions. In summary, WLSTM predicts human activities in the early stage of video, and the fusion of two modes (RGB-SME and OF-SME) can effectively improve the prediction performance of early and complete activities.

From the confusion matrix in Fig. 8, we can see the predicted results of early and complete human activities. In UT-Interaction set 1, hug and punch could be accurately identified in the early stage of activity, and the accuracy of other activities reached more than 90%. When the video observation ratio was 1.0, the average accuracy reached 98.3%, and the push action can be recognized accurately. In UT Interaction set 2, the prediction accuracies of both early and complete videos are lower than those of set 1 because of the increased noise. However, our method still maintains good prediction performance. The prediction accuracies of early and complete activity are 66.8% and 78.1%, respectively. On the sub-JHMDB dataset, the proposed WLSTM is able to accurately predict golf, pull up, push, and swing baseball in the early stage. The corresponding accuracies are 100%, 92%, 90%, and 86%, respectively. The prediction accuracies of catch, climb stairs, jump, kick ball, pick, run, shoot ball, and walk are relatively low: 67%, 67%, 38%, 33%, 75%, 20%, 67%, and 67%, respectively. This is because the activities in the sub-JHMDB dataset are more challenging than those in the UT-Interaction dataset, so the prediction performance is relatively low.

3.4 Evaluation of saliency-aware motion enhancement

In Section 2.1, we introduced the saliency-aware motion enhancement (SME) method where we use a geodesic distance based dynamic saliency estimation for motion region detection. In order to evaluate the impact of motion region detector on the performance of human activity prediction, the motion region is also detected by optical flow (OF), HOG-based human detector [31], and Gaussian mixture model (GMM) [32]. Then, we compare the activity prediction performances with half videos among these three methods on UT-Interaction and sub-JHMDB

Table 3 Prediction accuracy comparison of different modalities on sub-JHMDB (%)

Modality	Accuracy with half videos	Accuracy with full videos
RGB-SME	55.4	71.1
OF-SME	56.7	73.8
Two-modality fusion	66.8	78.1

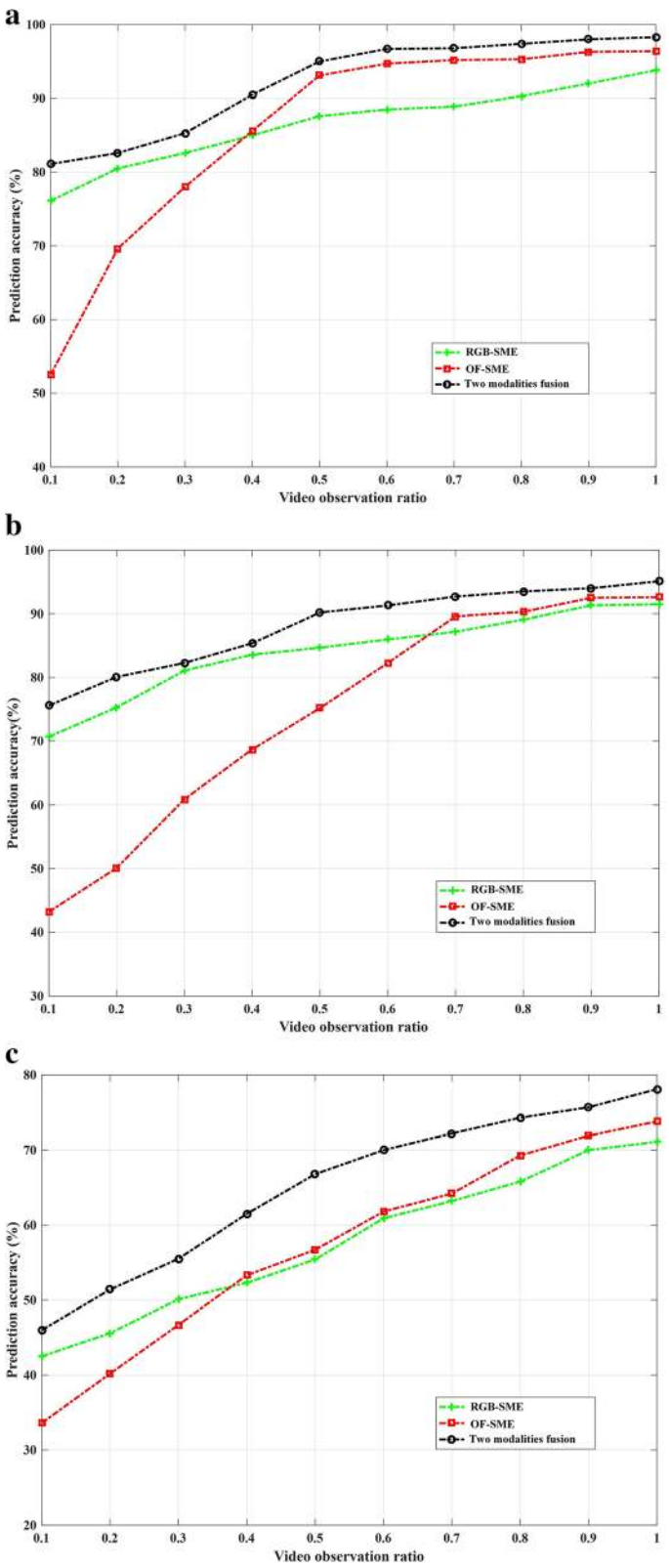


Fig. 7 Accuracy comparison of different modalities for three datasets. **a** UT-Interaction set 1. **b** UT-Interaction set 2. **c** Sub-JHMDB

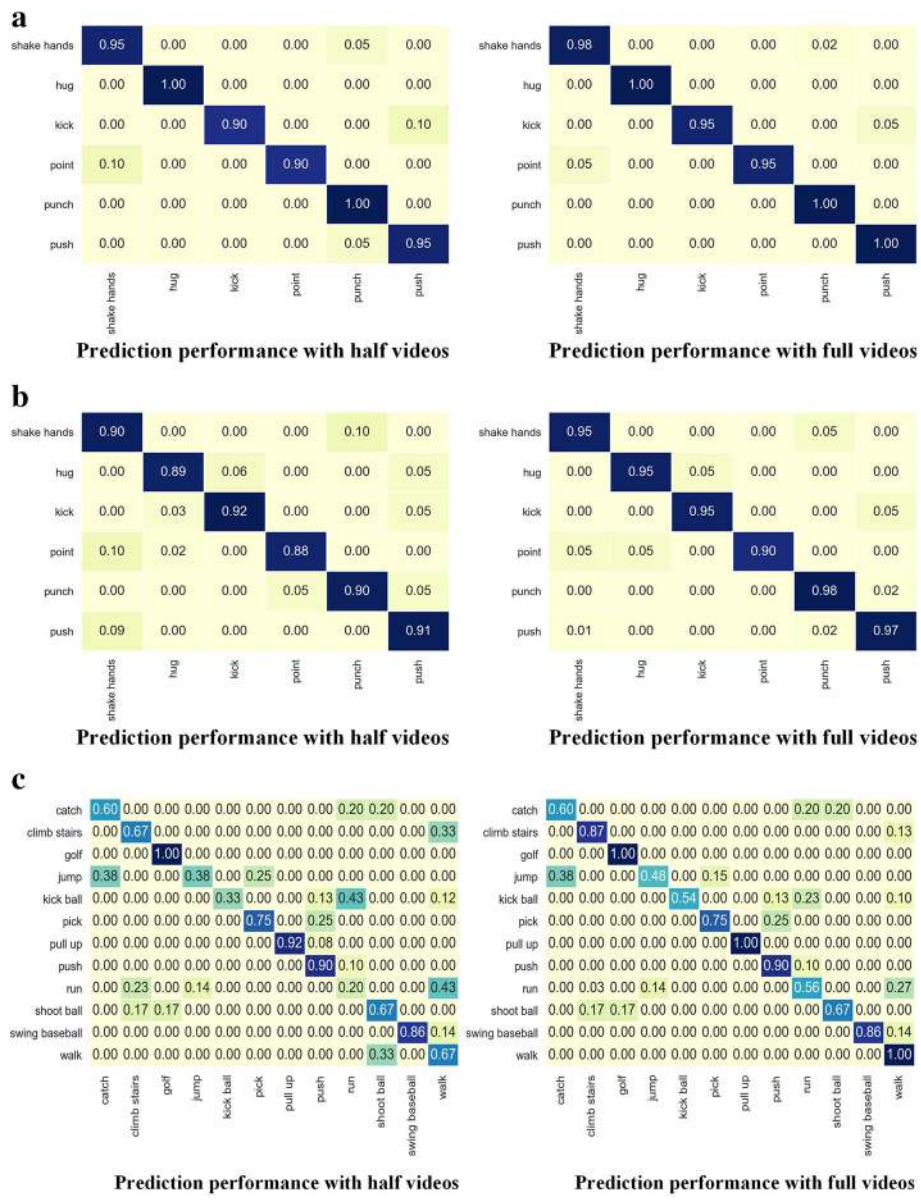


Fig. 8 Comparison of confusion matrices for three datasets (two modalities). **a** UT-Interaction set 1. **b** UT-Interaction set 2. **c** Sub-JHMDB

datasets. Figure 9 illustrates the comparison result and example frames on two datasets from different motion detectors. It is obvious that the proposed SME obtains competitive performance on both two datasets. The reasonable explanation for this phenomenon is that motion detection is the foundation of all subsequent work. So, the quality of WLSTM based feature encoding is highly depending on the integrity of motion information. For most action clips in UT-Interaction, the motion region is relatively clear and easy to be detected by the four motion detectors. Whereas for some action clips in sub-JHMDB, the objects are relatively weak and the background is complex which leads to false detection, thus increases the difficulty of activity prediction.

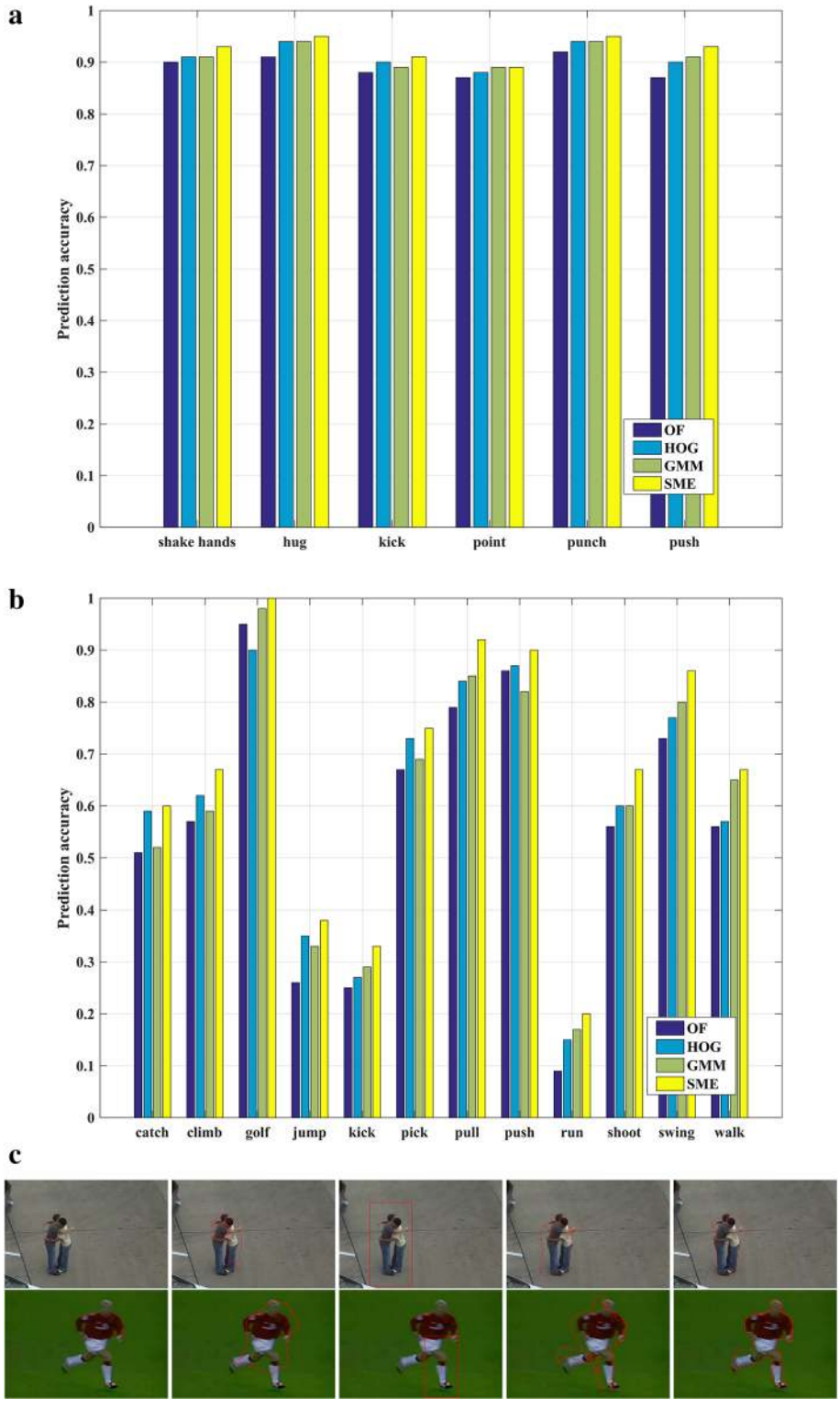


Fig. 9 Comparison of motion detectors. **a** Prediction accuracy on UT-Interaction. **b** Prediction accuracy on sub-JHMDB. **c** Example frames of different motion detectors. From left to right: original, optical flow, HOG human detector, GMM, and proposed SME

Table 4 Ablative study on UT-Interaction dataset (%)

Settings	SME	Weighted module	Accuracy with half videos	Accuracy with full videos
Setting 1	*x*	*x*	81.0	89.6
Setting 2	*x*	✓	88.5	92.1
Setting 3	✓	*x*	81.6	90.2
Setting 4	✓	✓	**92.6**	**96.7**

3.5 Ablative studies

In this section, to verify the effectiveness of the proposed saliency-aware motion enhancement (SME) and weight LSTM (WLSTM) network, various ablative experiments were performed using different settings on two datasets: RGB+OF+LSTM (setting 1), RGB+OF+WLSTM (setting 2), RGB-SME+OF-SME+LSTM (setting 3), and RGB-SME+OF-SME+WLSTM (setting 4). Experimental results on two datasets are shown in Table 4 and Table 5, respectively. The accuracy of the UT-Interaction dataset is the average of set 1 and set 2. The best value in the table is presented in bold. The experimental results show that the setting 4 (RGB-SME+OF-SME+WLSTM) obtains the superior performance among four settings on both two datasets. From Table 4, we can see the weighted module plays a more important role on UT-Interaction dataset. The accuracy of setting 1 and setting 3 is basically the same (81.0/89.6 vs 81.6/90.2). We speculate that this is because the video background of UT-Interaction dataset is relatively simple, so whether to use SME or not has little effect on prediction accuracy, whereas on sub-JHMDB dataset, both SME and weighted module provide strong contribution to the prediction accuracy. It is not hard to find out that without any of SME or weighted module may leads to a decline on prediction accuracy.

3.6 Comparison with state-of-the-art methods

By some extent, the proposed framework can be seen as a kind of spatiotemporal attention. Therefore, we compared our method with these SOTA methods which have spatiotemporal attention model to further evaluate the effectiveness of our method; experiments were performed using the WLSTM on the UT-Interaction dataset and sub-JHMDB dataset. The results of video activity prediction on the UT-Interaction dataset were compared with those of Hierarchical-M [19], TGTW [17], HSOM [13], Dynamic BOW [15], multi-stage LSTM [18], P-TS [21], and RU-LSTM [24]. The results of video activity prediction on the sub-JHMDB dataset were compared with those of PDP [14], confidence-DT [33], WSF-DS [34], RPAN [16], DRN [22], and RGN [23]. The accuracies of activity prediction were tested in the case of half video and full video with observation ratios of 0.5 and 1.0, respectively. The prediction results are shown in Table 6 and Table 7. The accuracy of the UT-Interaction dataset is the average of those

Table 5 Ablative study on sub-JHMDB dataset (%)

Settings	SME	Weighted module	Accuracy with half videos	Accuracy with full videos
Setting 1	*x*	*x*	48.9	59.2
Setting 2	*x*	✓	59.3	63.4
Setting 3	✓	*x*	61.5	69.7
Setting 4	✓	✓	**66.8**	**78.1**

Table 6 Prediction accuracy comparison on UT-Interaction (%)

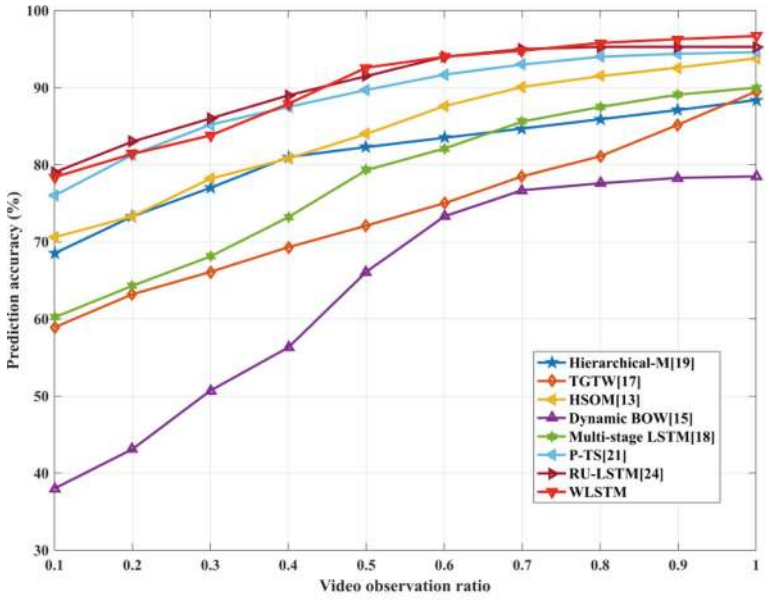
Method	Accuracy with half videos	Accuracy with full videos
Hierarchical-M [19]	82.3	88.4
TGTW [17]	72.1	89.5
HSOM [13]	84.0	93.8
Dynamic BOW [15]	66.1	78.5
Multi-stage LSTM [18]	79.3	90.0
P-TS [21]	89.7	94.6
RU-LSTM [24]	91.4	95.3
WLSTM (two modalities)	**92.6**	**96.7**

of set 1 and set 2 and the best value in the table is presented in bold. It can be seen that the WLSTM (two modalities) proposed in this paper achieves promising results on both datasets, and the prediction accuracy exceeds those of most of the state-of-the-art methods. Among these results, 92.6% prediction accuracy is achieved on the UT-Interaction dataset in the early stage of video behavior, and the average performance is 15.8% better than that of other similar methods. In full video recognition, the WLSTM achieves 96.7% recognition accuracy, which is ahead of the other methods. For sub-JHMDB dataset, our method has a good prediction accuracy of 66.8% in the early stage and 78.1% in the recognition of the complete video. These data show that the human activity prediction method proposed in this paper is effective under various conditions. Owing to the use of graph neural network, DRN [22] gains 71.8% in the prediction task. However, with the increase of video observation ratio, the final recognition performance does not exceed WLSTM. It is noteworthy that RPAN had the highest accuracy for complete video recognition, slightly higher than WLSTM's 78.1%. We conjecture that this is because an attention model of human posture is embedded in RPAN, which improves the recognition accuracy. However, when the video observation ratio is relatively low, the prediction accuracy of RPAN is affected by inaccurate estimation of human posture.

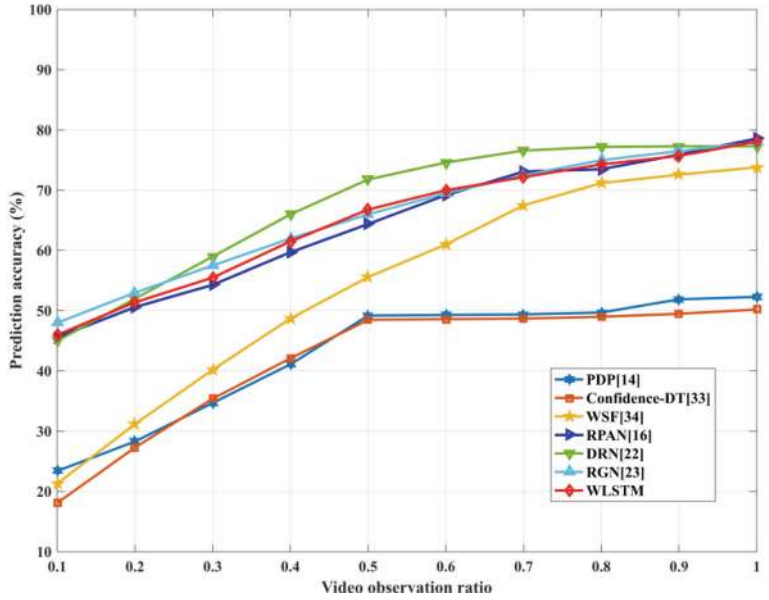
Figure 10 shows a comparison of the activity prediction accuracy under different video observation ratios. The activity prediction method proposed in this paper achieved good results in both datasets, and its prediction performance in the early stage of video surpassed that of all other methods used for comparison. On the UT-Interaction dataset in particular, the WLSTM (two modalities) had a lowest prediction accuracy of 78.4%, a highest prediction accuracy of 96.7%, and a prediction accuracy of 92.6% when the video observation ratio was only 0.5. It is noteworthy that this method

Table 7 Prediction accuracy comparison on sub-JHMDB (%)

Method	Accuracy with half videos	Accuracy with full videos
PDP [14]	49.2	52.3
Confidence-DT [33]	48.5	50.2
WSF [34]	55.6	73.8
RPAN [16]	64.4	**78.6**
DRN [22]	**71.8**	77.3
RGN [23]	66.0	78.0
WLSTM (two modalities)	66.8	78.1



(a) UT-Interaction



(b) sub-JHMDB

Fig. 10 Prediction accuracy comparison under different video observation ratio. **a** UT-Interaction. **b** Sub-JHMDB

maintains high prediction accuracy when the video observation ratios of the two datasets are between 0.5 and 1.0.

3.7 Experimental result visualization

In order to verify the effectiveness of SME in the prediction process, Figure 11 shows the visualization of SME on the two datasets. Each video frame corresponds to an activity. The three images in each group from left to right are the original video frame, the

Fig. 11 Visualization of saliency-aware motion enhancement. **a** Perform well. **b** Perform poorly

saliency image, and the SME-based moving object enhancement. In this paper, eight types of human activities are randomly selected from the UT-Interaction (left column) and sub-JHMDB (right column) datasets for visualization. Most of the videos in the sub-JHMDB datasets are polluted by intense background motion or other noise to some extent. All video frames are pre-processed using the SME proposed in this paper, in which the foreground moving area is made brighter than the background. This also shows that the pre-processed video frames retain more areas with visual information, which can make the convolution operations more focused on these regions. Most of the video frames in the UT-Interaction dataset have a static background, and moving objects can also be well segmented. These visualization results show that the proposed SME performs well in most cases.

Meanwhile, we also observe that the proposed SME is not very effective in the face of small scale motion, which is one of our future research directions. On the other hand, when processing the background pixels, we only reduce the brightness of these pixels by half. Pixels in background set still retain small amount of information, which plays a certain role in the subsequent prediction task. In addition, in the process of feature encoding by WLSTM, the weights of video frames with less information are lower. Therefore, the impact of this problem on the accuracy of prediction is relatively small.

To analyze the importance of learned weights in WLSTM, Fig. 12 shows the learned weights in different video frames. The WLSTM gives priority to the recognition of key video frames (the most representative stages) by assigning higher weights to specific activities (frames with red border). On the contrary, the weights of video frames with less information are lower, which is similar to the natural observation of human eyes. Comparing with the runners on the sports field, WLSTM places higher weights on the frames with foot and kick because the video frames containing the action of kick have higher classification importance for kickball. When the activity is simple, such as pull, the weight distribution of the WLSTM is relatively average.

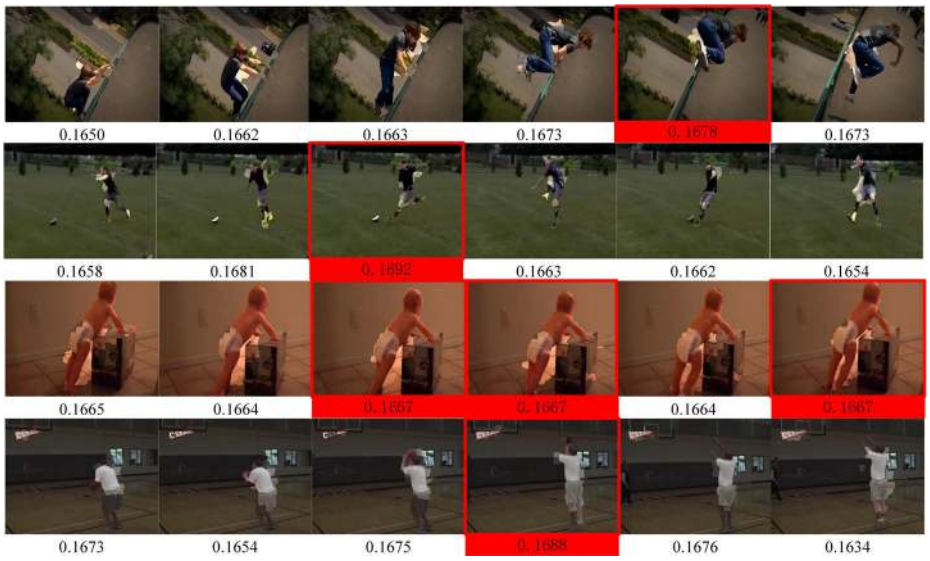


Fig. 12 Visualization of the learned weights by WLSTM

4 Conclusion

This paper proposed a weighted LSTM (WLSTM) network and saliency-aware motion enhancement (SME) to suppress the background noise of video frames, which can effectively reduce the interference of spatially and temporally redundancy. Experiments on two challenging datasets demonstrated the effectiveness of our method. This effectiveness is mainly attributed to the fully automatic moving object enhancement method and the weights of the sampled video frames. The former makes the convolution pay more attention to action-related regions, while the latter provides an effective solution to solve the long-term dependence between frames.

In the future work, we will consider some end-to-end learning frameworks and further explore the impact of spatial-temporal attention models for video activity prediction, especially focus on the models which are real time and able to deal with the multi-scale motion extraction.

Abbreviations

LSTM: Long short-term memory; WLSTM: Weighted long short-term memory; SME: Saliency-aware motion enhancement; ConvLSTM: Convolutional long short-term memory; CNN: Convolutional neural network; SLIC: Simple linear iterative clustering; OF: Optical flow; BOW: Bag of words; SGD: Stochastic gradient descent; BN: Batch normalization; TVL1: Total variation L1 normalization; HOG: Histogram of oriented gradient; GMM: Gaussian mixed model; TGTW: Temporally-weighted generalized time warping; HSOM: Hierarchical self-organizing map; PDP: Pose-based discriminative patch; WSF: Weighted score-level feature; RPAN: Recurrent pose-attention network; P-TS: Progressive teacher-student; DRN: Relational recurrent network; RGN: Residual generator network; RU-LSTM: Rolling unrolling long short-term memory

Acknowledgements

Thanks to all those who have suggested and given guidance for this article.

Authors' contributions

All the authors of this article contributed to this article. ZKW carried out the design and summary of the study; WZL and ZPJ participated in the experimental design and results analysis. All authors read and approved the final manuscript.

Funding

This research was supported by Zhejiang Provincial Natural Science Foundation of China under Grant No.LGF21F02007, LGF20F02003, LY18F030010) and Jiaxing Public Welfare Research Project under Grant No.2020AD30030.

Availability of data and materials
The video activity dataset UT-Interaction and sub-JHMDB is available online at http://cvrc.ece.utexas.edu/SDHA2010/Human_Interaction.html and http://jhmdb.is.tue.mpg.de/challenge/JHMDB respectively.

Competing interests
The authors declare that they have no competing interests.

Received: 25 February 2020 Accepted: 13 December 2020
Published online: 11 January 2021

References

1. L. Wang, Three-dimensional convolutional restricted Boltzmann machine for human behavior recognition from RGB-D video. EURASIP J. Image Video Process. **2018**, 120 (2018)
2. X. Wang, L. Gao, J. Song, et al., Beyond frame-level CNN: saliency-aware 3D CNN with LSTM for video action recognition. IEEE Signal Process. Lett. **24**(4), 510–514 (2017)
3. Z. Weng, Y. Guan, Trajectory-aware three-stream CNN for video action recognition. J. Electron. Imaging **28**(2), 021004 (2018)
4. Z. Weng, Y. Guan, Action recognition using length-variable edge trajectory and spatio-temporal motion skeleton descriptor. EURASIP J. Image Video Process. **2018**, 8 (2018)
5. H. Bilen, B. Fernando, E. Gavves, et al., Action recognition with dynamic image networks. IEEE Trans. Pattern Anal. Mach. Intell. **40**(12), 2799–2813 (2018)
6. A. Abdelbaky, S. Aly, Human action recognition using short-time motion energy template images and PCANet features. Neural Comput. Appl. (2020). https://doi.org/10.1007/s00521-020-04712-1
7. M. Majd, R. Safabakhsh, A motion-aware ConvLSTM network for action recognition. Appl. Intell. **49**(1), 2515–2521 (2019)
8. W. Tian, C. Yang, M. Zhang, et al., Internal transfer learning for improving performance in human action recognition for small datasets. IEEE Access **5**(99), 17627–17633 (2017)
9. I. Laptev, M. Marszalek, C. Schmid, et al., in *Proceedings of IEEE International Conference on Computer Vision and Pattern Recognition*. Learning realistic human actions from movies (2018), pp. 1–8
10. Y. Yun, H. Wang, Motion keypoint trajectory and covariance descriptor for human action recognition. Vis. Comput. **34**(3), 391–403 (2018)
11. Z. Tu, X. Wei, Q. Qin, et al., Multi-stream CNN: learning representations based on human-related regions for action recognition. Pattern Recognit. **79**(2), 32–43 (2018)
12. Z. Tu, Y. Li, J. Cao, et al., MSR-CNN: applying motion salient region based descriptors for action recognition. Proc. IEEE Int. Conf. Pattern Recognit., 3524–3529 (2016)
13. W. Ding, K. Liu, F. Cheng, Learning hierarchical spatio-temporal pattern for human activity prediction. J. Visual Commun. Image Representation **35**(C), 103–111 (2016)
14. S. Cao, K. Chen, R. Nevatia, Activity recognition and prediction with pose based discriminative patch model. Proc. IEEE Winter Conf. Appl. Comput. Vision, 2533–2541 (2016)
15. M. Ryoo, Human activity prediction: early recognition of ongoing activities from streaming videos. Proc. IEEE Int. Conf. Comput. Vision, 3468–3476 (2011)
16. W. Du, Y. Wang, Y. Qiao, RPAN: An end-to-end recurrent pose-attention network for action recognition in videos. Proc. IEEE Int. Conf. Comput. Vision, 3745–3754 (2017)
17. H. Wang, W. Yang, C. Yuan, et al., Human activity prediction using temporally-weighted generalized time warping. Neurocomputing **225**(1), 139–147 (2017)
18. M. Aliakbarian, F. Saleh, M. Salzmann, et al., Encouraging LSTMs to anticipate actions very early. Proc. IEEE Int. Conf. Comput. Vision, 37–46 (2017)
19. T. Lan, T. Chen, T. Savarese, A hierarchical representation for future action prediction. Proc. Eur. Conf. Comput. Vision **2014** (1975-1981)
20. Y. Sun, W. Wu, W. Yu, et al., Action recognition with motion map 3D network. Neurocomputing **297**(4), 33–39 (2018)
21. X. Wang, J. Hu, J. Lai, et al., Progressive teacher-student learning for early action prediction. Proc. IEEE Int. Conf. Comput. Vision Pattern Recognit., 3556–3565 (2019)
22. C. Sun, A. Shrivastava, C. Vondrick, et al., Relational action forecasting. Proc. IEEE Int. Conf. Comput. Vision Pattern Recognit., 273–283 (2019)
23. H. Zhao, R. Wildes, Spatiotemporal feature residual propagation for action prediction. Proc. IEEE Int. Conf. Comput. Vision, 7003–7012 (2019)
24. Guglielmo C, Pasquale C, Antonino F, et al. Knowledge distillation for action anticipation via label smoothing. arXiv preprint, arXiv:2004.07711v1.
25. M. Ryoo, J. Aggarwal, UT-interaction dataset, ICPR contest on semantic description of human activities (SDHA). Proc. IEEE Int. Conf. Pattern Recognit. Workshops, 2–4 (2010)
26. H. Jhuang, J. Gall, S. Zuffi, et al., Towards understanding action recognition. Proc IEEE Int. Conf. Comput. Vision, 3192–3199 (2014)
27. R. Achanta, A. Shaji, K. Smith, et al., SLIC superpixels compared to state-of-the-art superpixel methods. IEEE Trans. Pattern Anal. Mach. Intell. **34**(11), 2274–2282 (2012)
28. S. Ioffe, C. Szegedy, Batch normalization: accelerating deep network training by reducing internal covariate shift. Proc. Int. Conf. Int. Conf. Mach. Learn., 1356–1363 (2015)
29. C. Zach, T. Pock, H. Bischof, A duality based approach for realtime TV-L1 optical flow. Symp. Pattern Recognit., 214–223 (2007)
30. L. Wang, Y. Xiong, Z. Wang, et al., Temporal segment networks: towards good practices for deep action recognition. Proc Eur. Conf. Comput. Vision, 20–36 (2016)
31. Y. Pang, Y. Yuan, X. Li, et al., Efficient HOG human detection. Signal Process. **91**(4), 773–781 (2011)
32. M. Chen, X. Wei, Q. Yang, et al., Spatiotemporal GMM for background subtraction with superpixel hierarchy. IEEE Trans. Pattern Anal. Mach. Intell. **40**(6), 1518–1525 (2018)

33. X. Hu, Y. Jing, Confidence-based human action recognition with different-level features. Proc. Int. Conf. Mach. Learn. Cybern., 63–772 (2018)
34. G. Zhang, S. Jia, X. Li, et al., Weighted score-level feature fusion based on Dempster-Shafer evidence theory for action recognition. J. Electron. Imaging **27**(1), 1–10 (2018)

Publisher's Note