J. Cent. South Univ. (2020) 27: 101−113
DOI: https://doi.org/10.1007/s11771-020-4281-6

Springer

Interaction behavior recognition from multiple views

XIA Li-min(夏利民), GUO Wei-ting(郭炜婷), WANG Hao(王浩)

School of Automation, Central South University, Changsha 410083, China

Abstract: This paper proposed a novel multi-view interactive behavior recognition method based on local self-similarity descriptors and graph shared multi-task learning. First, we proposed the composite interactive feature representation which encodes both the spatial distribution of local motion of interest points and their contexts. Furthermore, local self-similarity descriptor represented by temporal-pyramid bag of words (BOW) was applied to decreasing the influence of observation angle change on recognition and retaining the temporal information. For the purpose of exploring latent correlation between different interactive behaviors from different views and retaining specific information of each behaviors, graph shared multi-task learning was used to learn the corresponding interactive behavior recognition model. Experiment results showed the effectiveness of the proposed method in comparison with other state-of-the-art methods on the public databases CASIA, i3Dpose dataset and self-built database for interactive behavior recognition.

Key words: local self-similarity descriptors; graph shared multi-task learning; composite interactive feature; temporal-pyramid bag of words

Cite this article as: XIA Li-min, GUO Wei-ting, WANG Hao. Interaction behavior recognition from multiple views [J]. Journal of Central South University, 2020, 27(1): 101−113. DOI: https://doi.org/10.1007/s11771-020-4281-6.

Open Science Identity (OSID)

1 Introduction

In recent years, video-based human behavior analysis [1−8] has attracted widespread attention from computer vision researchers, since it has wide application prospects in the visual monitoring system, human-computer interaction, sports analysis and other aspects.

The changes of light and observation angle make the recognition difficulty increase in single-view environment. Moreover, it may not be able to capture the ideal behavior characteristics in the current observation angle. Therefore, many researchers have tried to use multi-view methods to solve such problems. SHEN et al [9] used the three-joint point set to represent the action pose and look for the invariant of the rigid motion consisting of a three-node set between two frames. LI et al [10] proposed a generative Bayesian model not only jointly taking the features and views into account, but also learning a discriminant representation across distinctive categories. LI et al [11] learned a low dimensional manifold and reconstructed the 3D model by modeling the dynamic process. These multi-view algorithms usually need to know the angle between different perspectives in advance, which severely limits their applications. Therefore, researchers pay more attention to view-invariant feature learning. For example, ZHENG et al [12] proposed source domains and target domains dictionaries that are simultaneously learned to constitute a convertible dictionary pair, so that the same action has the same sparse representation at two different perspectives. LIU et al [13] used a bidirectional graph to model visual word bags,

Foundation item: Project(51678075) supported by the National Natural Science Foundation of China; Project(2017GK2271) supported by Hunan Provincial Science and Technology Department, China
Received date: 2018-11-05; **Accepted date:** 2019-06-16
Corresponding author: XIA Li-min, PhD, Professor; Tel: +86-13974961656; E-mail: xlm@mail.csu.edu.cn; ORCID: 0000-0002-2249-449X

which transformed a bag of visual-words (BOVW) action model into a bag of bilingual-words (BOBW) model with significant stability from different perspectives. JUNEJO et al [14] used a self-similarity matrix and a Support Vector Machine (SVM) classifier to assign a separated SVM classifier to each view, and apply a fusion method to achieve the final result. However, the correlation between different views is lacked. GAO et al [15] proposed a multi-view discriminative structured dictionary learning with group sparsity and graph model (GM-GS-DSD L). HSU et al [16] used a temporal-pyramid BOW to represent local spatio-temporal descriptors. HAO et al [17] utilized the sparse coding algorithm to transfer the low-level features of various views into the discriminative and high-level semantics space, and employed the multi-task learning approach for joint action modeling, while the divergence of low-level features in different perspectives will affect the action modeling.

At the same time, there are a lot of research results in the field of single-person behavior recognition, but few studies on interaction behavior recognition. In addition to the difficulties of motion recognition, such as the dithering of image acquisition equipment, the change of illumination intensity in the scene, and the occlusion of secondary object, the main problem in the recognition of two-person interaction behavior is to describe the body posture and complex spatio-temporal relationship. At present, there are two kinds of recognition methods for two-person interaction behavior. One is based on overall interaction recognition, while the other is based on individual segmentation. The former method mainly regards the two parts of the interaction behavior as a whole to describe the characteristics represented by overall spatio-temporal relationship, and recognizes interaction behavior by matching the test sample with the training sample. YU et al [18] used pyramid spatio-temporal relationship matching to check interactive actions. YUAN et al [19] proposed to construct spatio-temporal context kernel functions for interactive video matching and recognition. BURGHOUTS et al [20] improved the accuracy of interactive behavior recognition by introducing spatio-temporal layout to improve the inter-class discrimination ability of spatio-temporal features. LI et al [21] proposed a random forest method based on GA training and an effective spatio-temporal matching method to realize the recognition and understanding of interaction behavior. Overall methods treat the interactive behavior as a single-person action, without the need to separate the individual actions of the feature, and the processing idea is simple. However, the kind of method can not accurately represent the intrinsic attributes of interaction, which may cause limited accuracy. It often needs complex feature representation and matching methods to ensure accuracy. The latter is to understand the interaction behavior as a spatio-temporal combination between individual sub-actions. In the process of recognition, the meaning of a single individual action in the interaction behavior is recognized first, then the final recognition result is obtained by combining the spatio-temporal relationship between two individuals. KONG et al [22] proposed a training SVM based recognition model to identify interactive actions. SLIMANI et al [23] proposed a method based on symbiotic visual dictionary. The method is simple and easy to implement, but the recognition accuracy is limited. In a word, individual segmentation method either needs to track and detect the limbs of the human body, or needs to recognize the atomic movement. In complex interaction scenarios, it is difficult to get part of the human body and identify the atomic movement accurately due to occlusion and other factors.

In view of the advantages on multi-view recognition and the difficulties on interactive behavior recognition above, we propose a multi-view interactive behavior recognition method. In the process of bottom interactive feature extraction, the individual motion context and the global motion context of each video frame are spliced into composite interactive feature, which possess well discriminative description to represent complex interactive behavior. For solving the influence of observation angle changes on human interactive behavior recognition, self-similarity matrix (SSM) [14] based on composite interactive feature is constructed, since it has affine invariance and projection invariance. However, the absolute value of the self-similarity matrix element is related to the feature of each video frame, and further away from the diagonal elements, the less reliable the value. Therefore, we extract the local features on diagonal of SSM and represent them by temporal pyramid word-bag model to obtain view-invariant interactive features, which reduce the effect of high-speed motion and retain the temporal information. It only is robust to the changing view, but ignores the relevant information between different interactive behaviors and different views. To address this problem, we propose a graph shared multi-task learning function (GSMTL) and classify human interactive behavior by reconstruction of the minimum label information error. Our method is more discriminative for interactive behavior recognition and has well robustness to view changes.

2 Interactive behavior features

In order to describe interactive behavior in video frames, first we need to extract and track interest points from input image sequences. Then, the individual motion context and the global motion context are constructed since they have been proven to be efficient for capture motion relationships at the individual and global levels [24]. Finally, the two motion contexts are connected to composite interactive feature with a highly discriminative description.

2.1 Interest points detection

In the actual operation, the filter is obtained by transforming the scale factor of Gaussian kernel function [25], and then convoluted with the video sequence to get image sequences in different scales.

$$\boldsymbol{L}(x,y,\sigma)=\boldsymbol{G}(x,y,\sigma)\otimes \boldsymbol{I}(x,y) \tag{1}$$

where $\boldsymbol{L}(x, y, \sigma)$ denotes the scale space; $\boldsymbol{I}(x, y)$ denotes the input image; $\boldsymbol{G}(x, y, \sigma)$ is a Gaussian kernel function with scale factor σ.

$$\boldsymbol{G}(x,y,\sigma)=\frac{1}{2\pi\sigma^2}\mathrm{e}^{-\frac{x^2+y^2}{2\sigma^2}} \tag{2}$$

where σ is the scale factor and the Harris-Laplace multi-scale detection auto-correlation matrix [25] is:

$$\boldsymbol{M}=\mu(x,y,\sigma_{\mathrm{I}},\sigma_{\mathrm{D}})=\sigma_{\mathrm{D}}^2 g(\sigma_{\mathrm{I}})\otimes \begin{pmatrix} L_x^2(x,\sigma_{\mathrm{D}}) & L_xL_y(x,\sigma_{\mathrm{D}}) \\ L_xL_y(x,\sigma_{\mathrm{D}}) & L_y^2(y,\sigma_{\mathrm{D}}) \end{pmatrix} \tag{3}$$

where x and y are the pixel coordinates of the image; σ_{I} is the integral scale; σ_{D} is the differential scale. Generally, $\sigma_{\mathrm{I}}=s\sigma_{\mathrm{D}}$ and the empirical value s=0.6. Multi-scale Harris detects the response of points on each scale space image.

$$R=\det(\mu(x,y,\sigma_{\mathrm{I}},\sigma_{\mathrm{D}}))-\alpha\,\mathrm{trace}^2(\mu(x,y,\sigma_{\mathrm{I}},\sigma_{\mathrm{D}}))>T \tag{4}$$

where α=0.04−0.06; T is the threshold value used to control the number of extraction corner points. The larger the R is, the more likely it is the corner point.

2.2 Composite interactive feature extraction

After obtaining interest points and boundaries, we will use them to represent a video frame. Let $Q_t=(B_t, P_t)$ be the video frame of time duration t. Here, B_t are the set of boundaries and P_t are the interest points. Thus, $Q=[Q_1, Q_2, \cdots, Q_T]$ denotes a video containing T frames. Now, we begin to construct individual movement context and global movement context.

First, the geometric centers of individual motion and global motion are calculated, as shown below:

$$C_t^h=(x_i,y_i) \tag{5}$$

$$C_t^{\mathrm{In}}=\frac{1}{H}\sum_{h=1}^{H}C_t^h \tag{6}$$

where C_t^h denotes geometric center position for the h-th human boundary and the t-th frame; C_t^{In} denotes average value of geometric center of the existing human objects for the t-th frame; and H is the number of people in video.

Next, the gradients between interest points and the individual motion center, the global motion center are calculated respectively in each frame (see Eqs. (7) and (8)).

$$g_t^h(j)=\{(C_t^h-p_t^{h,j})\mid j=1:NP_t^h\} \tag{7}$$

$$g_t^{\mathrm{In}}(j)=\{(C_t^{\mathrm{In}}-p_t^j)\mid j=1:NP_t\} \tag{8}$$

where $p_t^{h,j}$ means the j-th interest point for the h-th boundary and the t-th frame; $g_t^h(j)$ denotes the gradient between the centroid for the h-th boundary and the j-th interest point belonging to the h-th boundary; p_t^j is the j-th interest point for the t-th frame; $g_t^{\mathrm{In}}(j)$ denotes the gradient between the global motion center and the j-th interest point. And $NP_t=\sum_{h=1}^{H}NP_t^h$, where NP_t^h means interest points for the h-th boundary and the t-th frame; NP_t means all interest points for the t-th frame. Each motion context is a histogram of size N_{B}, where N_{B} indicates the number of sub-regions in the boundary with the geometric center as the reference point. For each frame, the two motion contexts are calculated as follows:

$$\lambda_t^h = [F^h(1), F^h(2), \cdots F^h(N_{\mathrm{B}})] \tag{9}$$

$$\lambda_t^{\mathrm{In}} = [F^{\mathrm{In}}(1), F^{\mathrm{In}}(2), \cdots F^{\mathrm{In}}(N_{\mathrm{B}})] \tag{10}$$

where λ_t^h and λ_t^{In} are the individual motion context and global motion context, respectively. Each bin of the histogram is the sum of the magnitude of gradients.

$$F^h(l) = \sum_{j \in NP_t^h} \mathrm{mag}(g_t^h(j)) \cdot \delta[\mathrm{ang}(g_t^h(j)), \mathrm{rang}(l)] \tag{11}$$

$$F^{\mathrm{In}}(l) = \sum_{j \in NP_t} \mathrm{mag}(g_t^{\mathrm{In}}(j)) \cdot \delta[\mathrm{ang}(g_t^{\mathrm{In}}(j)), \mathrm{rang}(l)] \tag{12}$$

where mag(•) and ang(•) denote magnitude and angle of interest point gradient, respectively. $\delta[\mathrm{ang}(g_t^h(j)), \mathrm{rang}(l)]$ is an indicator function, it equals 1 when $\mathrm{ang}(g_t^h(j)) \in \mathrm{rang}(l)$ and 1, otherwise, it equals 0. rang(l) is the angle range of the l-th bin in histogram.

The final composite interactive feature vector for the t-th frame is represented by the concatenation of these two motion contexts, with a size of $N_{\mathrm{B}}\times(H+1)$.

$$p_t = [\lambda_t^h, \cdots, \lambda_t^H, \lambda_t^{\mathrm{In}}] \tag{13}$$

3 Self-similarity matrix representation of interaction behavior

The same human action will make different visual effect and lead to the differences of the feature extracted from interaction behavior, due to different taken angles. Self-similarity matrix reflects the relationship between image sequences, which has affine invariance and projection invariance. Feature self-similarity matrix discards the feature of the frame and only preserves the feature difference between frames. The feature difference is represented by the distance between the two feature descriptors and it has little to do with the view position. In the case of similar interactive behaviors at two different moments, it will be a short distance between the two feature descriptors. However, it will be a opposite situation for two different interactive behaviors. We use self-similarity matrix to describe interaction behavior, since it has a good performance on interaction behavior represented between different views.

3.1 SSM construction of composite interaction features

Given a video image sequence $R=\{r_1, r_2, \cdots, r_N\}$, SSM is defined as follows:

$$\boldsymbol{D} = [r_{i,j}]_{i,j=1,2,\cdots N} = \begin{bmatrix} 0 & r_{12} & r_{13} & \cdots & r_{1N} \\ r_{21} & 0 & r_{23} & \cdots & r_{2N} \\ r_{31} & r_{32} & 0 & \cdots & r_{3N} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ r_{N1} & r_{N2} & r_{N3} & \cdots & 0 \end{bmatrix} \tag{14}$$

$$r_{ij} = \overline{r_i r_j} = \left\| r_i - r_j \right\| \tag{15}$$

where ‖·‖ indicates the distance between low-level feature vectors. The elements on the diagonal should be zero, since they represent the distance between the feature vector and themselves. Meanwhile, the distance between r_i and r_j is equal which the distance between r_j and r_i. It is clear that $\boldsymbol{D}$ is a symmetric matrix. The mode of self-similarity matrix depends on the features and distance metrics.

In this paper, we define r_{ij} as the Euclidean distance between the composite interactive features. Therefore, the self-similarity matrix based on composite interactive feature is defined as follows:

$$\boldsymbol{CIM} = [r_{i,j}]_{i,j=1,2\cdots N} = \begin{bmatrix} 0 & p_{12} & p_{13} & \cdots & p_{1N} \\ p_{21} & 0 & p_{23} & \cdots & p_{2N} \\ p_{31} & p_{32} & 0 & \cdots & p_{3N} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ p_{N1} & p_{N2} & p_{N3} & \cdots & 0 \end{bmatrix} \tag{16}$$

$$p_{ij} = \overline{p_i p_j} = \left\| p_i - p_j \right\| \tag{17}$$

3.2 Local feature descriptor

The absolute values of the SSM's elements are related to the features of each frame, and the further away from the diagonal, the less reliable the pixel. To solve this problem, JUNEJO et al [26] used the local feature descriptor to represent the self-similarity matrix. Each local descriptor is represented by a directional gradient histogram of a semicircle region image. The center of the semicircle region is on the diagonal of the self-similarity matrix (see Figure 1), and the diameter of the semicircle represents the time span. To obtain the vector $\boldsymbol{h}_i^a = [h_{i,b}^a]_b'$, b=1, 2, ⋯, 8, we can make a calculation for the 8-direction gradient histogram for the i-th time at the region A. The local descriptor at the i-th time is derived from the sequentially connection of 11 region-histograms. If the region overflows the boundary of the self-similarity matrix, the vector of the region is set to a zero vector. In Ref. [26], to reduce the influence of motion speed, the temporal information is discarded and all local descriptors are used to compose bag-of-word to represent the features of the whole human movement. In the BOW method, a large number of local descriptors are randomly selected and a vocabulary containing K visual words is obtained by K-means clustering. In subsequent training and classification, each local descriptor is represented by a word nearest to it. All local descriptors are represented by the frequency of the occurrence of the word, forming a histogram of the frequency of the occurrence of the word. At this point, each video sequence can be represented by a histogram.

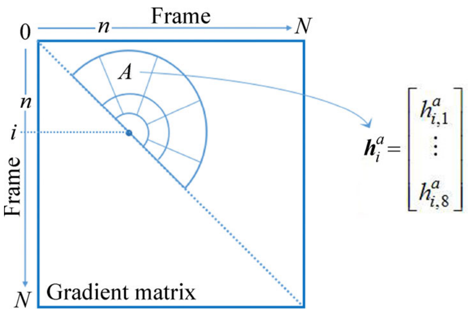

Figure 1 Local descriptor for self-similarity matrix

3.3 Temporal-pyramid BOW

Although the tradition BOW[27] method reduces the influence of motion speed, it ignores the temporal information, which leads to a low ability to recognize interactive behavior in reverse time sequence. In this work, the temporal-pyramid BOW (see Figure 2) is applied to describing the local SSM features, which is a coarse-to-fine representation consisting of the different levels connection of all histograms derived from diverse time slices. Supposing that there are L levels in the pyramid structure, we divide entire interactive behavior features into 2^l partial feature slices for level l, where $0\le l\le L-1$. A standardized histogram is calculated from each slice, and the histograms of all segments are joined together to construct temporal-pyramid BOW model. Temporal-pyramid BOW differs from the traditional BOW method in which it retains temporal information of interactive behavior with two or more higher levels because of more time segments, while at level 0, it's only a traditional BOW providing tolerance of variation of temporal pattern. We can maintain the superiority of different levels by combining with these levels. The final view-invariant interactive behavior feature x is represented by the different levels connection of all histograms derived from diverse time slices with dimension of $K\times(2^L-1)$. K is size of BOW model codebook.

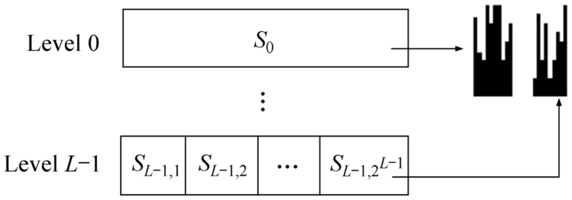

Figure 2 Temporal-pyramid BOW model

4 Interactive behavior recognition based on multi-view and graph shared multi-task learning

The view-invariant interactive feature was obtained in the previous section only by considering the robustness of the same interactive behavior under different views, while ignoring the related information between different interactive behaviors and different views. In this section, we propose a graph shared multi-task learning algorithm, which explores potential relationships from different views by grouping video samples of different views into the graph set [28], and mines the correlation between different interactive behaviors and retain the unique information of each behavior by learning the corresponding interactive behavior recognition model.

4.1 Objective function

In this work, assume that there are J interactive behaviors in V views. For each type of interactive behavior, V video sample can be obtained at the same time, and video interactive feature in each single view is defined as $x^v \in R^{(K\times(2^L-1))\times 1}$. Based on the above ideas, we collectively formulate the multi-task learning model as $F(x^v,W)=\{f(x^v,w_j)\}_{j=1}^J$, where $w_j \in R^{(K\times(2^L-1))\times 1}$ is the model for the j-th task; $W=[w_1,w_2,\cdots w_J]\in R^{(K\times(2^L-1))\times J}$ is the model for joint MTL problem; $\boldsymbol{f}(x^v,w_j)=(\boldsymbol{x}^v)^{\mathrm{T}}w_j$ denotes a single-task learning model based on paired features, which can be defined as a linear prediction model; $X_j^v=\{x_{js}^v\}_{s=1}^S\in R^{(K\times(2^L-1))\times S}$ denotes the feature set for the j-th task and the v-th view, where S is the number of samples. $X_j=[X_j^1, X_j^2,\cdots,X_j^v]$ contains the specific partwise features of all training samples in multiple views for the j-th model learning. $X^v=\{X_j^v\}_{j=1}^J$ denotes the feature set for entire interactive behavior training set for the v-th view.

$Y_j=\{y_{js}\}_{s=1}^S\in R^S$ is the corresponding labels of the j-th task and $Y=\{Y_j\}_{j=1}^J$ corresponds the labels for all tasks. In this work, the shared information of all tasks and the specific information of each task are studied at the same time, that is $w^j=u+v^J$, where the model parameter w^j of each task is composed of the shared information u of all tasks and the specific information v^j of each task. In addition, in order to explore the potential relationship between different views, the video samples from different views were built into a graph set, and different coefficients were distributed in each sample. Therefore, we formulate the objective function as:

$$\langle u^*,\boldsymbol{V}^*,b^*\rangle=\arg\min\nolimits_{(u,\boldsymbol{V},b)}\{L(b,X,Y,u,\boldsymbol{V})+\Omega(u,\boldsymbol{V})+\Gamma(b)\} \tag{18}$$

where $L(b, X, Y, u, \boldsymbol{V})$ means the empirical loss function of joint multi-task learning and can be computed by $L(b,X,Y,u,\boldsymbol{V})=\sum_{j=1}^J\|\beta^1(X_j^1)^{\mathrm{T}}(u+v^j)+\cdots+\beta^V(X_j^V)^{\mathrm{T}}(u+v^j)-Y_j\|_2^2$ where $b=[\beta^1, \beta^2, \cdots, \beta^V]$ denotes the weight of each view which is automatically learned. Thus, this term enables to be rewritten as $L(b,X,Y,u,\boldsymbol{V})=\sum_{j=1}^J\|b(X_j)^{\mathrm{T}}(u+v^j)-Y_j\|_2^2$.

Besides, objective function has two penalty terms:

1) $\Omega(u,\boldsymbol{V})=\lambda_1\|Y-Xu\|+\lambda_2\|u\|_2^2+\lambda_3\|\boldsymbol{V}\|_{2,1}$. The first term is used to limit the error of the shared part model parameters; the second term is used to control the model complexity of the shared part; and the third term is used to limit the columns of the matrix $\boldsymbol{V}$. Each column of $\boldsymbol{V}$ can be viewed as a specific feature of each task. If the J tasks are similar, the number and value of non-zero columns in $\boldsymbol{V}$ should be small. When all tasks are equal, the matrix $\boldsymbol{V}$ should approach a zero matrix.

2) Convex hull term (CHT) $\Gamma(b)$. In this section, we constructed a convex hull term for evaluating the interaction behavior correlation from different views. The sum of sub-term is estimated as 1. And in Eq. (18), the model will reconstruct all samples from different views for the j-th task. Meanwhile, all vectors will be adjusted appropriately. Thus, not only the correlation between different views will be mined, but also different samples could be combined together. Therefore, CHT can be defined as $\Gamma(b)=\rho_1\|b\|_1+\rho_2\|be-1\|_2^2$, meaning sum of every regular constants and control weighted coefficient. The first term in $\Gamma(b)$ is used to control sparsity. The corresponding coefficient will be small if the correlation between samples is not strong. Thus, this sample could be considered having low impact. The sum of second terms is used to make sum of all convex coefficients equal to 1.

In summary, the objective function Eq. (18) can be further defined as Eq. (19):

$$\min g(u,\boldsymbol{V},b)=\sum_{j=1}^J\left\|b(X_j)^{\mathrm{T}}(u+v^i)-Y_j\right\|_2^2+\lambda_1\|Y-Xu\|+\lambda_2\|u\|_2^2+\lambda_3\|\boldsymbol{V}\|_{2,1}+\rho_1\|b\|_1+\rho_2\|be-1\|_2^2 \tag{19}$$

4.2 Solution

In order to solve the energy minimization problem of the objective function, the iterative minimization method [29] is used to solve the three parameters, b, u and $\boldsymbol{V}$ in Eq. (18). Specifically, the following two steps are repeated until convergence: 1) fixing b, minimizing $\langle u^*,\boldsymbol{V}^*,b^*\rangle$ over u and $\boldsymbol{V}$; 2) fixing u and $\boldsymbol{V}$, minimizing $\langle u^*,\boldsymbol{V}^*,b^*\rangle$ over b.

When b is fixed, Eq. (19) can be explicitly abbreviated to Eq. (20):

$$\min g(u,\boldsymbol{V},b)=\sum_{j=1}^J\left\|b(X_j)^{\mathrm{T}}(u+v^i)-Y_j\right\|_2^2+\lambda_1\|Y-Xu\|+\lambda_2\|u\|_2^2+\lambda_3\|\boldsymbol{V}\|_{2,1} \tag{20}$$

The optimization problem in Eq. (20) can be solved by the accelerated proximal method (APM). Because $\lambda_3\|\boldsymbol{V}\|_{2,1}$ is a non-smooth convex function, Eq. (20) is divided into two parts:

$$f(u,\boldsymbol{V})=\sum_{j=1}^J\left\|b(X_j)^{\mathrm{T}}(u+v^i)-Y_j\right\|_2^2+\lambda_1\|Y-Xu\|+\lambda_2\|u\|_2^2 \tag{21}$$

$$r\left(\boldsymbol{V}\right)=\lambda_3\left\|\boldsymbol{V}\right\|_{2,1} \tag{22}$$

Linearizing Eq. (21) obtains:

$$F_{u_t,\boldsymbol{V}_t,l_t}\left(u,\boldsymbol{V}\right)=f\left(u_t,V_t\right)+\left\langle u-u_t,\nabla_u f\left(u_t,\boldsymbol{V}_t\right)\right\rangle+$$
$$\frac{l_t}{2}\left\|u-u_t\right\|_2^2+\left\langle \boldsymbol{V}-\boldsymbol{V}_t,\nabla_{\boldsymbol{V}} f\left(u_t,\boldsymbol{V}_t\right)\right\rangle+\frac{l_t}{2}\left\|\boldsymbol{V}-\boldsymbol{V}_t\right\|_F^2 \tag{23}$$

where u_t and $\boldsymbol{V}_t$ represent the values of u and $\boldsymbol{V}$ at the t-th iteration; l_t stands for step size. Thus, it can be obtained:

$$\left(u_{t+1},\boldsymbol{V}_{t+1}\right)=\underset{u,\boldsymbol{V}}{\arg\min}\frac{l_t}{2}\left\|u-\left(u_t-\frac{1}{l_t}\nabla_u f\left(u_t,\boldsymbol{V}_t\right)\right)\right\|_2^2+$$
$$\frac{l_t}{2}\left\|\boldsymbol{V}-\left(\boldsymbol{V}_t-\frac{1}{l_t}\nabla_{\boldsymbol{V}} f\left(u_t,\boldsymbol{V}_t\right)\right)\right\|+r\left(\boldsymbol{V}\right) \tag{24}$$

Finally, we can get:

$$u_{t+1}=u_t-\frac{1}{l_t}\nabla_u f\left(u_t,\boldsymbol{V}_t\right) \tag{25}$$

$$v_{t+1}^{(j)}=\max\left(0,1-\frac{\lambda_3}{l_t\left\|s_{t+1}^{(j)}\right\|_2}\right)s_{t+1}^{(j)},$$

$$s_{t+1}=\boldsymbol{V}_t-\frac{1}{l_t}\nabla_{\boldsymbol{V}} f\left(u_t,\boldsymbol{V}_t\right) \tag{26}$$

Furthermore, when u and $\boldsymbol{V}$ are fixed, we transform the objective function Eq. (18) into the following form:

$$\left\langle b\right\rangle=\arg\min{}_{(\mathrm{b})}\left(\sum\nolimits_{j=1}^{J}\left\|b(X_j)^{\mathrm{T}}(u+v^j)-Y_j\right\|_2^2+\right.$$
$$\left.\rho_1\left\|b\right\|_1+\rho_2\left\|be-1\right\|_2^2\right) \tag{27}$$

$$\left\langle b\right\rangle=\arg\min{}_{(\mathrm{b})}\left(\sum\nolimits_{j=1}^{J}\left\|b(X_j)^{\mathrm{T}}(u+v^j)-Y_j\right\|_2^2+\right.$$
$$\left.\frac{\rho_1}{t}\left\|b\right\|_1+\frac{\rho_2}{t}\left\|be-1\right\|_2^2\right) \tag{28}$$

$$\left\langle \mathrm{b}\right\rangle=\arg\min{}_{(\mathrm{b})}\left[\sum\nolimits_{j=1}^{J}\left(\left\|bX-Z\right\|_2^2+\frac{\rho_1}{t}\left\|b\right\|_1\right)\right] \tag{29}$$

where $X=[(X_j)^{\mathrm{T}}(u+v^j);\rho_2\, {}^{e}\!/\!_{t}]$ and $Z=[Y_j;\rho_2\, {}^{e}\!/\!_{t}]$. By some L_1-minimization methods including gradient descent algorithm and Least angle regression [30], the problem in Eq. (29) can be easily solved.

4.3 Human interactive behavior recognition

The corresponding model $w^j=u+v^J$ and b can be established by graph shared multi-task learning function. In the process of prediction and recognition, the J prediction label, which is obtained by multiplying the test sample X_{Test}^{*}, and each task model parameter $w^j=u+v^J$ subtract the real labels of test samples and then choose the minimum interactive behavior label as the label of test samples. Finally, to judge whether the predicted labels are consistent with real labels, make use of the final recognition rate which is the result of the figure for correctly predicted samples divided by the total number of predicted samples.

$$\left\langle L_j\right\rangle=\mathrm{argmin}_j\left\|b(X_{\mathrm{Test}}^{*})^{\mathrm{T}}(u+v^j)-Y_j\right\|_2^2 \tag{30}$$

5 Experiment

To verify the effectiveness of the proposed method, we conducted experiments on the CASIA dataset and the i3Dpose dataset. Considering the limitation of video quantity in these two datasets, this work adopts leave-one-out cross validation strategy to evaluate the performance.

5.1 CASIA dataset

The CASIA behavior analysis database [31] has a total of 1446 video data, which is captured by cameras in horizontal, angle and top down view in outdoor environment, providing experimental data for behavioral analysis. The data are divided into single-person behaviors and interactive behaviors. Single-person behavior includes walking(1), running(2), bending(3), jumping(4), crouching(5), fainting(6), wandering(7) and punching(8). There are 24 people involved in each type of behavior, about 4 times per person. Interactive behaviors consist of robing(9), fighting(10), following(11), follow and gathering(12), meeting and parting(13), meeting and gathering(14) and overtaking(15). The frame rate is 25fps and the resolution is 320× 240. The operation example is shown in Figure 3.

5.2 i3Dpose dataset

The i3Dpose dataset [32] contains 12 different behaviors, including 6 single behaviors: walking, running, jumping forward, jumping in place, bending, and one hand waving. Four hybrid activities are sitting down-standing up, walking-siting down, running-falling and running-jumping-walking. Two interactive activities are two persons handshaking and one person pulling another. Camera setups are labeled on Figures 4(a)−(h). Eight amateurs (2 females and 6 males) are participated 13 times for every action. We only test interactive behaviors.

5.3 Self-built dataset

Because there are few data sets for interactive behavior under multiple views, we have built a multi-view database. The self-built data set was shot and established by us on the lawn of Central South University, with a total of 1200 pieces of video data. All video was shot simultaneously from two different angles by two un-calibrated static cameras, with a frame rate of 25fps and a video image resolution of 640×480. The data are divided into single-person behavior and multi-person behavior. Single-person behaviors include walking, running, jumping, squatting, fainting, wandering. For each type of behavior, 10 people are separately photographed, about 4 times per person. Multi-person behaviors include robing, fighting, tailing, many people walking, chatting and playing badminton. Each behavior is filmed 5−6 times. Figure 5 shows some examples of the behavior for a self-built data set.

5.4 Selection of parameters

Firstly, the hierarchical L and the codebook size K of the temporal pyramid bag-of-word are determined by experiments on the i3Dpose dataset. N is the video frames size, and the fixed step size is set to 2 when extracting the local descriptor of self-similarity matrix. Figure 6 shows the testing results on the i3Dpose database. It can be seen from the graph that interactive behavior recognition rate is the highest which ups to 95.12% when L=2 and K=400. In the subsequent experiments, we all set with this parameter.

5.5 Comparison of different interaction features

Our method is compared with other advanced interactive feature extraction methods. SAEID et al [33] represented interactions by forming temporal

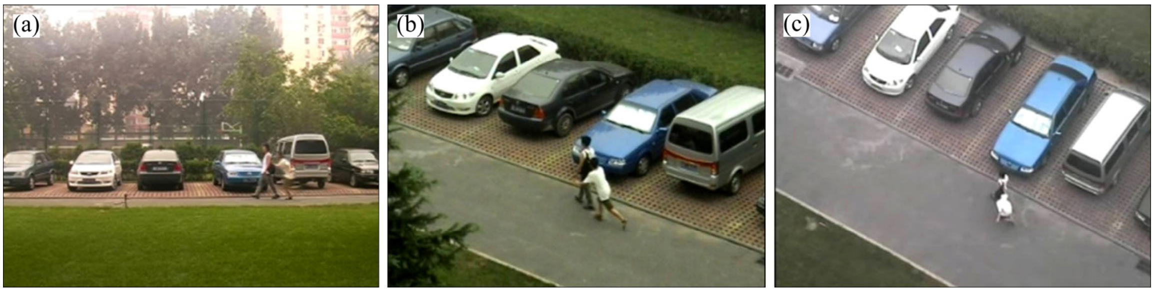

Figure 3 Samples on CASIA dataset(robing, punching a car): (a) Horizontal view (HV); (b) Angle view (AV); (c) Top down view (TV)

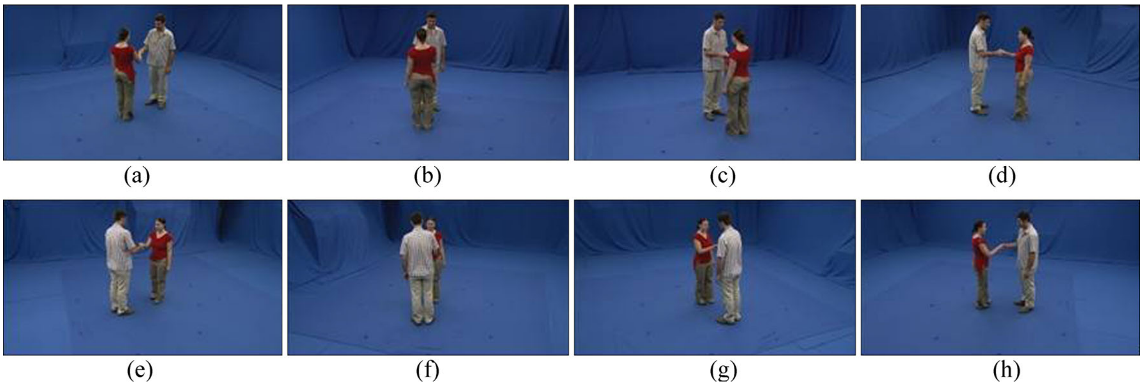

Figure 4 Samples on i3Dpose dataset of shaking hands

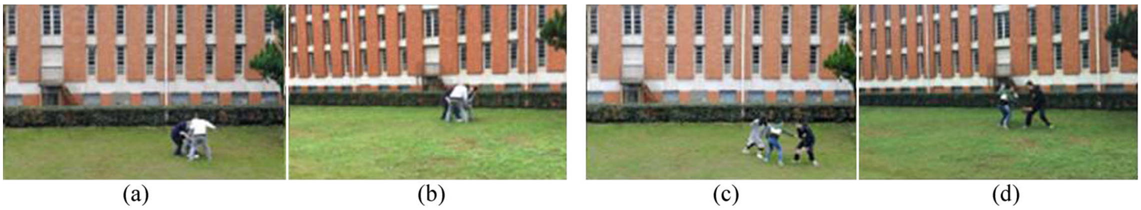

Figure 5 Samples on self-built dataset of fighting (a, b) and robing (c, d)

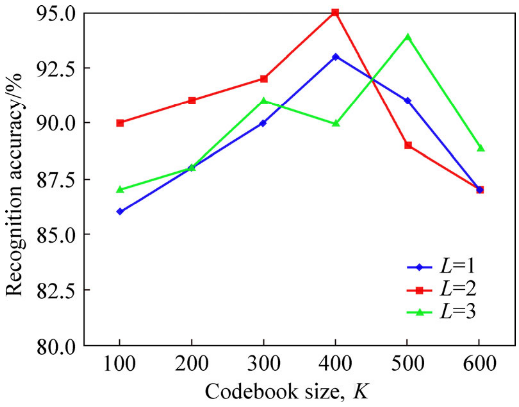

Figure 6 Parameter setting of temporal pyramid bag-of-word model

trajectories, coupling together the body motion of each individual and their proximity relationships with others; In Ref. [34], a novel representation based on hierarchical histogram of local feature sequences was proposed for human interaction recognition; KONG et al [22] proposed a discriminative model to encode interactive phrases based on the latent SVM formulation; JI et al [35] proposed a contrastive feature distribution model (CFDM) for interaction recognition. A novel feature descriptor based on spatial relationship and semantic motion trend similarity between body parts was proposed for human-human interaction recognition in Ref. [36]. The comparison results are shown in Table 1. The proposed compositional interactive feature descriptors contain inter-class and intra-class variations, representing motion relationships at the individual and global levels, which are highly discriminatory. Secondly, i3Dpose datasets are collected indoors and CASIA are collected outdoors result in complex background, which increase the difficulty in recognition accuracy. Therefore, the recognition results on i3Dpose datasets are better than those on CASIA datasets.

Table 1 Recognition accuracy of different interaction features (%)

Ref.	CASIA	i3Dpose	Self-built
[33]	80.32	93.54	71.21
[34]	81.23	92.82	70.61
[22]	70.02	90.61	65.32
[35]	76.90	92.73	69.81
[36]	79.53	93.80	67.32
This work	86.21	95.12	73.30

5.6 Influence of different view fusion methods on recognition

We compared the proposed method with some other up-to-date multi-view information fusion methods. FV is combined with sparse coding, and finally SRMTL is used to classify human behavior [17]; GM-GS-DSDL is a multi-view discriminant structured dictionary learning based on group sparse and graph model [15], which is used to fuse different views and recognize human actions human behavior; MDVSD aims to learn a structured dictionary shared by all views and multiple view-specific structured dictionaries with each corresponding to a dictionary [37]. The results are shown in Tables 2−4. The optimal recognition rates are 86.21%, 95.12% and 73.3% respectively. The reason is that the more the views are used as training data, the richer the view information contained in action recognition model is. Our method is superior to FV with sparse coding. Because a specific action was recorded from various camera views, the appearance of the action would be completely different, and finally the completely different FV underlying features will be obtained, which will reduce the human action recognition rate. However, the self-similarity matrix consists of the Euclidean distances between the compositional interactive feature descriptor, which changes less with the variation of the viewing angle. As shown in Tables 2−4, the combination interaction feature self-similar matrix has good robustness to the change of the viewing angle, and performs well on interactive human behavior recognition under multiple views, with the recognition rate as high as 86.21%, 95.12% and 73.3%.

5.7 Comparison of different multi-task functions

We also used GSMTL, Lasso, L21 and SRMTL to learn the action recognition model to verify the superiority of the GSMTL method. The combination of perspectives with the highest

Table 2 Recognition accuracy of different view fusion methods on CASIA (%)

View	FV+sparse coding	GM-GS-DSDL	MDVSD	This work
HV	65.22	70.76	73.19	73.68
HV _AV	71.76	75.07	79.89	80.56
HV _AV _TV	77.19	80.76	85.39	86.21

Table 3 Recognition accuracy of different view fusion methods on i3Dpose (%)

View	FV+sparse coding	GM-GS-DSDL	MDVSD	This work
C4	81.41	83.24	87.14	87.25
C4_ C8	79.54	84.91	88.21	88.59
C4_ C8_ C5	82.17	86.39	89.36	90.36
C4_ C8_ C5_C1	84.47	88.36	90.14	91.57
C4_ C8_ C5_C1_C3	84.91	88.12	92.21	92.69
C4_ C8_ C5_C1_C3_C7	86.07	89.39	93.24	93.98
C4_ C8_C5_ C1_C3_C7_C2	86.59	90.24	93.61	94.65
All views	87.19	91.95	94.34	95.12

Table 4 Recognition accuracy of different view fusion methods on self-built database

View	FV+sparse coding	GM-GS-DSDL	MDVSD	Overtake
C1	59.13	63.27	68.37	70.6
C1_C2	61.45	65.56	71.54	73.3

recognition rates in Section 5.6 were viewed as training data. Lasso method was introduced sparsity into the multi-task learning model and aims to reduce the complexity of the model and feature learning. The L21 norm regularization approach captured the common problems of multiple related tasks by limiting all models and shared the same feature set. The SRMTL assumes that all tasks are related and the model of each task is close to the average of all task models. As shown in Table 5, the performance of the GSMTL method is slightly higher than that of other methods because GSMTL considers not only the common information of all tasks, but also the characteristic information of each task, making the model parameters of each task more discriminant.

Table 5 Recognition accuracy of different multi-task functions (%)

Database	Lasso	L21	SRMTL	GSMTL
CASIA	60.34	64.18	72.36	86.21
i3Dpose	71.95	74.29	82.24	95.12
Self-built	52.69	58.36	65.91	73.30

5.8 Comparison of different methods

Tables 6 and 7 show the confusion matrix on CASIA data set and self-built database, from which we could find the overall accuracy rates were 86.21% and 73.3%.

Table 6 Confusion matrix on CASIA

Action	Rob	Fight	Follow	Follow and gather	Meet and part	Meet and gather	Overtake
Rob	88.1	8			3.9		
Fight	11.5	86.4				2.1	
Follow			85.1	10.1	4.8		
Follow and gather	2.2		5.9	88.2			3.7
Meet and part			2.0		84.2	11.8	2.0
Meet and gather				8.2		87.9	3.9
Overtake			4.1	5.6	6.7		83.6

Table 7 Confusion matrix on self-built database

Action	Rob	Fight	Tail	Walk	Chat	Badminton
Rob	59.1	26.3			14.6	
Fight	15.8	67.3			16.9	
Tail			78.2	16.4	5.4	
Walk			15.9	77.6		6.5
Chat	14.3	14.2			71.5	
Badminton			5.7	6.2		86.1

The proposed method had also been compared with other advanced methods. From Tables 6 and 7, we can clearly observe that the overall recognition rate of the proposed method is higher than that of other multi-view methods with an accuracy of 86.21% and 73.3%. Local motion and context information of interest points collected by Harris-Laplace algorithm are encoded at individual and global levels, describing the interaction behavior accurately. Local descriptors abstracted from self-similarity matrix are robust to the changing views. The graph shared multi-task learning explores the potential correlation between different actions and retain specific information of each task, making human behavior recognition more discriminatory. Therefore, our proposed method improves the accuracy of interactive behavior recognition under the change of views.

Then, we tested the computational complexity on the CASIA dataset, and calculated the average of the six methods respectively. It can be seen from the Table 8 that the calculation time of method in Ref. [15] is long because the GM-GS-DSDL used the graph algorithm and constructed the

Table 8 Comparison of different methods

Ref.	Accuracy/%	Time/s
[14]	65.70	0.71
[15]	80.76	1.45
[16]	75.46	0.98
[17]	77.19	1.23
[37]	85.39	1.67
This work	86.21	1.14

discriminative structured dictionary, which consumed a lot of time. In Ref. [17], the computational complexity of fisher vector is relatively large but sacrifices the recognition rate compared with GSMTL, CMTL decreases time. Our method used temporal-pyramid BOW to describe the local feature of self-similarity matrix and utilized the accelerated proximal method to optimize the graph shared multi-task learning function, so the calculation speed is faster.

6 Conclusions

This paper suggests a novel multi-view interactive behavior recognition method based on self-similarity matrix and graph shared multi-task learning function. The main work is as follows:

1) The proposed composite interactive feature describes the interaction motion at individual and global levels.

2) The self-similarity matrix represented by local feature descriptors is constructed to reduce the difference of the same interactive behavior caused by the changing view. Compared with other multi-view methods, the proposed method does not need to reconstruct the 3D model and not to calculate the relationship between different views.

3) In order to discover latent correlation among different interactive behaviors and retain specific information of each task, the corresponding interactive behavior recognition model is learned by using the graph shared multi-task learning function, and the minimum reconstruction of tag information error is used to classify human behavior.

4) The proposed method can achieve competing performance against the state-of-the-art methods for interactive behavior recognition on CASIA, i3Dpose and self-built database.

References

[1] FERNANDO I, RICARDO P. Human actions recognition in video scenes from multiple camera viewpoints [J]. Cognitive Systems Research, 2019, 56: 223−232. DOI: 10.1016/ j.cogsys.2019.03.010.

[2] LIN Bo, FANG Bin, YANG Wei-bin. Human action recognition based on spatio-temporal three-dimensional scattering transform descriptor and an improved VLAD feature encoding algorithm [J]. Neurocomputing, 2019, 348: 145−157. DOI: 10.1016/j.neucom.2018.05. 121.

[3] HAN Fei, REILY B, HOFF W, ZHANG Hao. Space-time representation of people based on 3D skeletal data: A review [J]. Computer Vision & Image Understanding, 2017, 158(3): 85−105. DOI: 10.1016/j.cviu.2017.01.011.

[4] LIU J, WANG G. Skeleton based human action recognition with global context-aware attention LSTM networks [J]. IEEE Transactions on Image Processing, 2018, 27(4): 1586−1599.

[5] AMIRA B M, EZZEDDINE Z. Abnormal behavior recognition for intelligent video surveillance systems: A review [J]. Expert Systems with Applications, 2018, 91: 480−491. DOI: 10.1016/j.eswa.2017.09.029.

[6] LUVIZON D C, TABIA H, PICARD D. Learning features combination for human action recognition from skeleton sequences [J]. Pattern Recognition Letters, 2017, 99: 13−20. DOI: 10.1016/j.patrec.2017.02.001.

[7] HUANG Z W, WAN C D. Deep learning on lie groups for skeleton-based action recognition [C]// IEEE Conference on Computer Vision and Pattern Recognition. Honolulu, HI, USA：IEEE, 2017: 1243−1252.

[8] SHAHROUDY A , LIU J. NTU RGB+D: A large scale dataset for 3D human activity analysis [C]// IEEE Conference on Computer Vision and Pattern Recognition. Seattle, WA: IEEE, 2016: 1010−1019.

[9] SHEN Y P, FOROOSH H. View-invariant action recognition from point triplets[J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2009, 31(10): 1898−1905.

[10] LI Jin-xing, ZHANG Bob, ZHANG David. Generative multi-view and multi-feature learning for classification [J]. Information Fusion, 2018, 45: 215−226. DOI: 10.1016/ j.inffus.2018.02.005.

[11] LI R, TIAN T, SCLAROFF S. Simultaneous learning of nonlinear manifold and dynamical models for high-dimensional time series [C]// 11th IEEE International Conference on Computer Vision. Rio de Janeiro, BRAZIL: IEEE, 2007:1687−1694.

[12] ZHENG J J, JIANG Z L, PHILLIPS J. Cross-view action recognition via a transferable dictionary pair [C]// 23rd British Machine Vision Conference. Guildford, England: Springer-Verlag, 2012: 1−10.

[13] LIU J G, SHAH M, KUIPERS B. Cross-view action recognition via view knowledge transfer [C]// 2011 IEEE Conference on Computer Vision and Pattern Recognition. NJ, USA : IEEE, 2011: 3209−3216.

[14] JUNEJO I N, DEXTER E, LAPTEV I. Cross-view action recognition from temporal self-similarities [C]// The

European Conference on Computer Vision. Marseille, France: Springer, 2008: 293−306.

[15] GAO Zan, ZHANG Hua, XU Guang-ping. Multi-view discriminative and structured dictionary learning with group sparsity for human action recognition [J]. Signal Processing, 2015, 112: 83−97. DOI: 10.1016/j.sigpro. 2014.08.034.

[16] HSU Yen-pin, LIU Cheng-yin, CHEN Tzu-yang. Online view-invariant human action recognition using rgb-d spatio-temporal matrix [J]. Pattern Recognition, 2016, 60: 215−226. DOI: 10.1016/j.patcog.2016.05. 010.

[17] HAO Tong, WU Dan, WANG Qian, SU Jin-sheng. Multi-view representation learning for multi-view action recognition [J]. J Vis Commun Image R, 2017, 48: 453−460. DOI: 10.1016/j. jvcir.2017.01.019.

[18] YU T H, KIM T K, CIPOLLA R. Real-time action recognition by spatiotemporal semantic and structural forests [C]// Proceedings of the 21st British Mac hine Vision Conference. United Kingdom: Springer-Verlag, 2010: 1−12. DOI: 10.5244/C.24.52.

[19] YUAN F, SAHBI H, PRINET V. Spatio-temporal context kernel for activity recognition [C]// Proceedings of the 1st Asian Conference on Pattern Recognition. Beijing, China: IEEE, 2011: 436−440.

[20] BURGHOUTS G J, SCHUTTE K. Spatio- temporal layout of human actions for improved bag-of-words action detection [J]. Pattern Recognition Letters, 2013, 34(15): 1861−1869. DOI: 10.1016/j.patrec.2013.01. 024.

[21] LI N J, CHENG X, GUO H Y, WU Z Y. A hybrid method for human interaction recognition using spatio-temporal interest points [C]// The 22nd International Conference on Pattern Recognition. Stockholm, Sweden: IEEE, 2014: 2513−2518.

[22] KONG Yu, JIA Yun-de, FU Yun. Interactive phrases: semantic descriptions for human interaction recognition [J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2014, 36(9): 1775−1788. DOI: 10.1109/ TPAMI.2014.230 3090.

[23] SLIMANI K, BENEZETH Y, SOUAMI F. Human interaction recognition based on the co-occurrence of visual words [C]// Proceedings of IEEE Conference on Computer Vision and Pattern Recognition Workshops. Columbus, Ohio, USA: IEEE, 2014: 461−469. DOI: 10.1109/CVPRW. 2014.74.

[24] CHO N G, PARK S H, PARK J S. Compositional interaction descriptor for human interaction recognition [J]. Neurocomputing, 2017, 267: 169−181. DOI: 10.1016/j.neucom.2017.06.009.

[25] HARRIS C, STEPHENS M J. A combined corner and edge detector [C]// Proceedings of Fourth Alvey Vision Conference. Manchester, England: IEEE, 1988: 147−151. DOI: 10.5244/C.2.23.

[26] JUNEJO I, DEXTER E, LAPTEV I. Cross-view action recognition from temporal self-similarities [C]// European Conference on Computer Vision. Berlin: Springer-Verlag, 2008: 293−306. DOI: 10.1007/978-3-540- 88688-4_22.

[27] LAPTEV I, MARSZALEK M, SCHMID C. Learning realistic human actions from movies [C]// Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. Anchorage, AK: 2008: 1−8. DOI: 10.1109/ CVPR.2008. 4587756.

[28] HU YQ, AJMAL S, ROBYN O. Sparse approximated nearest points for image set classification [C]// IEEE Conference on Computer Vision and Pattern Recognition. Colorado Springs, CO: IEEE, 2011: 121−128.

[29] WRIGHT J, YANG A, GANESH A. Robust face recognition via sparse representation [J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2009, 31(2): 210−227.

[30] HUANG Zhi-wu, WANG Rui-ping, SHAN Shi-guang. Face recognition on large-scale video in the wild with hybrid Euclidean-and-Riemannian metric learning [J]. Mixture Research Article Pattern Recognition, 2015, 48(10): 3113−3124. DOI: 10.1016/j.patcog.2015.03. 011.

[31] ZHANG Z, HUANG K Q, TAN T N. Multi-thread parsing for recognizing complex events in videos [C]// European Conference on Computer Vision. Marseille, France: Springer,2008: 738−751.

[32] NIKOLAOS G, HANSUNG K, ADRIAN H. The i3DPost multi-view and 3D human action/interaction database [C]// 2009 Conference for Visual Media Production. London, England: IEEE, 2009:159−168.

[33] SAEID M, FARZAD S, RANYA A. Online human interaction detection and recognition with multiple cameras[J]. IEEE Transactions on Circuits and Systems for Video Technology, 2017, 27(3): 649−663.

[34] CAVENT A, IKIZLER N. Histograms of sequences: A novel representation for human interaction recognition [J]. IET Computer Vision, 2018, 12(6): 844−854.

[35] JI Yan-li, CHENG Hong, ZHENG Ya-li, LI Hao-xin. Learning contrastive feature distribution model for interaction recognition [J]. Journal of Visual Communication and Image Representation, 2015, 33: 340−349. DOI: 10.1016/j.jvcir.2015.10.001.

[36] LIU B L, CAI H B, JI X F, LIU H H. Human-human interaction recognition based on spatial and motion trend feature [C]// IEEE International Conference on Image Processing. Beijing, China: IEEE, 2017: 4547−4551.

[37] WU Fei, JING Xiao-yuan, YUE Dong. Multi-view discriminant dictionary learning via learning view-specific and shared structured dictionaries for image classification [J]. Neural Process Lett, 2017, 45(2): 649−666. DOI: 10.1007/ s11063-016-9545-7.

(Edited by ZHENG Yu-tong)

中文导读

多视角下的交互行为识别

摘要：本文提出了一种基于局部自相似描述符和图共享多任务学习的多视角交互行为识别方法。首先，提出了一种复合交互特征表示方法，该方法对兴趣点局部运动的空间分布及其上下文进行编码。其次，为了减小观测角度变化对识别的影响并保留时序信息，用时间金字塔词袋模型对局部自相似描述符进行表示。为了从不同的视角探索不同交互行为之间的潜在关联，并保留每种交互行为的特定信息，采用图共享多任务学习学习相应的交互行为识别模型。结果表明，该方法在 CASIA、i3Dpose 公共数据集和自建交互行为识别数据库上相比其他方法识别率更高。

关键词：局部自相似描述符；图共享多任务学习；复合交互特征；时间金字塔词袋模型