Energy-Based Periodicity Mining With Deep Features for Action Repetition Counting in Unconstrained Videos

Jianqin Yin, Yanchun Wu, Chaoran Zhu, Zijin Yin, Huaping Liu, *Senior Member, IEEE*, Yonghao Dang, Zhiyi Liu, and Jun Liu, *Member, IEEE*

***Abstract*— Action repetition counting is to estimate the occurrence times of the repetitive motion in one action, which is a relatively new, significant, but challenging problem. To solve this problem, we propose a new method superior to the traditional ways in two aspects, without preprocessing and applicable for arbitrary periodicity actions. Without preprocessing, the proposed model makes our scheme convenient for real applications; processing the arbitrary periodicity action makes our model more suitable for the actual circumstance. In terms of methodology, firstly, we extract action features using ConvNets and then use Principal Component Analysis algorithm to generate the intuitive periodic information from the chaotic high-dimensional features; secondly, we propose an energy-based adaptive feature mode selection scheme to adaptively select proper deep feature mode according to the background of the video; thirdly,we construct the periodic waveform of the action based on the high-energy rules by filtering the irrelevant information. Finally, we detect the peaks to obtain the times of the action repetition. Our work features two-fold: 1) We give a significant insight that features extracted by ConvNets for action recognition can well model the self-similarity periodicity of the repetitive action. 2) A high-energy based periodicity mining rule using features from ConvNets is presented, which can process arbitrary actions without preprocessing. Experimental results show that our method achieves superior or comparable performance on the three benchmark datasets, i.e. YT_Segments, QUVA, and RARV.**

***Index Terms*— Action repetition counting, deep ConvNets.**

I. Introduction

VISUAL action repetition in real life appears in many applications, such as sports, music playing and manufacturing assembly. It is important to count the repetition of a specific motion in videos, which is of great application value in video question answering [1], action classification [2]–[4], segmentation [5]–[7], 3D reconstruction [8], [9], motion tracking [10] and motion planning of robots [11]. Due to the diversity of motion patterns and the limitations in video capturing (e.g., camera movement), the development of a universal solution for counting the repetitive actions remains under-explored.

To count the action repetition, early methods usually assumed that repetitive motions occurred in fixed scenes with regular periodicity. With this assumption, they usually used traditional features to analyze the action repetition, including human skeleton obtained by sensor device [8], [9], the wavelength spectrum [12]. However, the actions to be counted are usually captured in complex dynamic scenes and have variable periodicity over different periods, making the traditional features not suitable for the counting tasks. To tackle this complexity involved in real circumstances, two methods have been proposed in recent years. In [13], multiple repetitive motion modes are simulated to construct the periodicity of the repetitive actions to realize the counting. Because the simulated motion modes are fixed, the algorithm can handle the specified modes well. But the performance significantly decreases for actions with other repetitive modes [14]. In [14], a counting method based on the detection of the moving area is proposed to achieve improved performance. However, this method relies on additional preprocessing steps to detect the moving region. To sum up, the method based on the simulated action modes is ineffective to count the varied types of repetitive motions [13], and the scheme based on detecting moving regions is highly dependent on the preprocessing performance [14]. These problems motivate us to find an action repetition counting scheme that can work for various motion modes without relying on extra detection steps.

There are many challenges in action repetition counting for unconstrained videos, which can be summarized as follows. (1) Background noise & view changes & irrelevant motion. In the unconstrained videos, besides the information of interested repetitive action, there exists other information such as changes in the background, the changing viewpoint of moving cameras, actions of the false objects and other unrelated movements. How to distinguish the periodic actions from these

Manuscript received October 16, 2020; revised December 28, 2020; accepted January 19, 2021. Date of publication January 28, 2021; date of current version December 6, 2021. This work was supported in part by the National Natural Science Foundation of China under Grant 61673192 and Grant 62025304; and in part by the Fundamental Research Funds for the Central Universities under Grant 2020XD-A04-1. This article was recommended by Associate Editor M. Shehata. *(Corresponding authors: Jianqin Yin; Jun Liu.)*

Jianqin Yin, Yanchun Wu, Yonghao Dang, and Zhiyi Liu are with the School of Artificial Intelligence, Beijing University of Posts and Telecommunications, Beijing 100876, China (e-mail: jqyin@bupt.edu.cn).

Chaoran Zhu and Zijin Yin are with the International School, Beijing University of Posts and Telecommunications, Beijing 100876, China.

Huaping Liu is with the Department of Computer Science and Technology, Tsinghua University, Beijing 100084, China.

Jun Liu is with the Department of Mechanical Engineering, City University of Hong Kong, Hong Kong (e-mail: jun.liu@cityu.edu.hk).

Color versions of one or more figures in this article are available at https://doi.org/10.1109/TCSVT.2021.3055220.

Digital Object Identifier 10.1109/TCSVT.2021.3055220

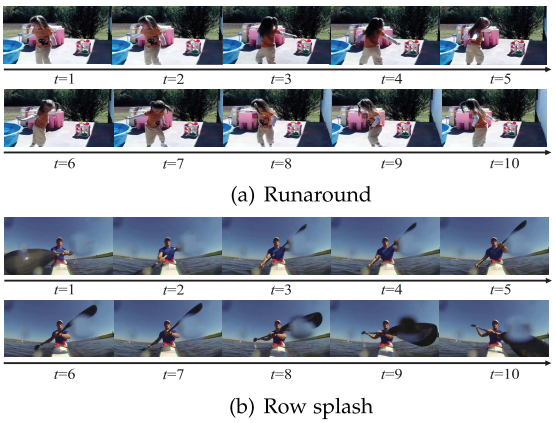


Fig. 1. The illustration of coupled motion. Here, "$t = *$" denotes the "$*$"-th frame of the video. In Fig. 1(a), the movement directions of the upper body are different than that of the feet of the girl; in Fig. 1(b), the motion from frames 1 to 5 is very similar to that of frames 6 to 10 in the same action.

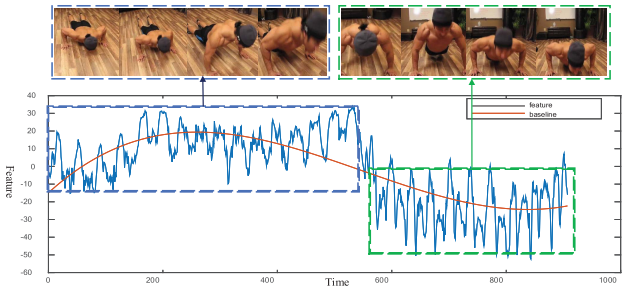


Fig. 2. An illustration of the signal trend. The overall amplitude of the signal changes with the variation of the background. We can describe this change using the signal represented by the red curve, and we name it as the signal trend.

various irrelevant signals is key to counting the repetitive action. (2) Various action repetition modes. Action repetition modes are very different in different actions. For example, the repetition mode can be rotation, swinging, translation, and other modes. How to discover the relationship between the periodicity and various repetition modes is another challenge. (3) Huge intra-class differences & coupled motion. In terms of intra-class differences, the amplitude and the frequency of the repetition within the same action can also be different; in terms of coupled motion, it is very easy to double-counting the number of action repetition and result in additional counting errors. As shown in Fig. 1, taking the action "Row splash" as an example, the sub-motions in the same complex action are similar; for the action "Runaround", different parts of the human body move with different frequencies so that the counting results tend to summarize the repetitive times of different parts. Therefore, even for the same action, how to design a scheme that is robust to the huge intra-class differences and coupled motion is an additional challenge.

Due to the above challenges, the research progress on repetition counting is relatively slow in recent years. In contrast, the development of action recognition using deep learning methods has achieved far advances [15], [16], [28], [30]–[32], which opens up new possibilities to propose new solutions for repetition counting. Deep ConvNets can capture versatile and robust action features for action recognition, which illustrates that the features from ConvNets contain rich information on the action. For repetition counting, there is an important clue that the self-similarity periodic dynamics are the key. Accordingly, we propose in this paper that the deep features are helpful for mining the self-similarity periodicity of the action and can be used to count the repetition. This insight relieves us from the preprocessing and helps in addressing the challenges of modelling the periodicity of the unconstrained videos.

On the one hand, the deep features contain the periodic rules and other irrelevant information for repetition counting, such as motion features for representing human motion. On the other hand, it is hard to mine the self-similarity action periodicity in a high-dimension feature space. To better mine the repetitive rules from the deep features, we use Principal Component Analysis (PCA) to extract the most significant component of the deep features, and then, we propose an energy-based action periodicity mining scheme that describes below to suppress the effect of irrelevant information for counting. Moreover, we empirically found that the non-stationary repetitive signals frequently appear in the first-dimensional principal component. This discovery is an important insight for modelling the action repetition. Using PCA, we convert the high-dimension deep features to a one-dimension waveform. In a word, deep features combining with PCA can generate the periodic signal.

Moreover, as shown in Fig. 2, we empirically show that the background noise and view changes are often reflected as the changing amplitude of the DC component in the one-dimensional waveform of the video in the time domain. Because this DC component reflects the trend of the time signals, we name the DC component as the signal trend. To overcome the effect of background noise and view changes, it is important to remove these DC components from the one-dimensional waveform of actions for counting repetition. In this paper, polynomial regression is applied to simulate these time-varying DC components and the signal trend can be removed based on the simulated signal.

For the coupled motion, it is sure that there is some relationship between the action and its sub-actions. In this paper, we convert the action into a signal, and then, we use Power Spectral Density (PSD) of the signal to mine the relationship of the action and its sub-actions, to detect the coupled and hence eliminate the irrelevant signal of its sub-actions.

For the video including the repetitive action, most of the energy of the video usually comes from the repetitive motions due to the repetition. From this perspective, once we obtain the waveform that contains the periodic motions, we can locate the periodic motion using the high energy rules. In this paper, frequency analysis is used to locate the signals with the high energy corresponding to the repetitive action automatically, and then we can finish counting based on the located signal.

As discussed above, a new action repetition counting method is proposed to solve the unconstrained action repetition counting in videos, and main contributions are summarized as follows. (1) We propose an energy-based action repetition

counting method without extra preprocessing, which can be used to effectively count the repetition of the action with arbitrary periodic motions and arbitrary viewpoints for unconstrained videos. (2) We give an important insight that the periodic self-similarity movement information can be well modelled by the deep features of the action used in action recognition. Specifically, we use the typical two-stream framework pre-trained on the task of action recognition to extract different types of deep features of the videos with the repetitive actions, and moreover, we provide an adaptive feature mode selection method to automatically select different feature modes for the videos with different backgrounds. (3) We propose a novel high-energy-based action periodicity reconstruction scheme to reconstruct the robust periodic waveform of the video with the repetitive action. On the one hand, the proposed scheme is robust to the noises by filtering the low-energy of the unrelated motion in the power spectrum of the video; on the other hand, the proposed scheme can adapt to arbitrary complex actions by automatically detecting the coupled and further removing the noise caused by its sub-actions, which can significantly improve the performance of repetition counting.

The remainder of the paper is organized as follows. The next section investigates the related work. Section 3 discusses our algorithm in detail. The datasets, evaluation criteria, experimental results and analysis are given in Section 4. We conclude our paper in Section 5.

II. Related Work

Action repetition counting is usually realized by converting the video into a one-dimensional waveform with the repetitive motion structures [8], [9] and then analyzing the spectral or frequency component by Fourier transform or wavelet analysis. Waveform analysis is also widely used in the periodic movement analysis [17], [18], [33], which is very related to the repetition counting. At the same time, as discussed in the Introduction, the action feature is another important problem for counting. Therefore, we will review the related work from the following two aspects: periodic movement analysis and deep features for action analysis.

A. Periodic Movement Analysis

The existing methods have achieved remarkable results in video action periodic analysis tasks. Burghouts *et al.* [12] proposed a spatiotemporal filter bank for online estimation of action repetition. But it was limited to the motion of stationary scenes, and the filter bank was manually adjusted. Laptev *et al.* [19] used a matching method for action counting, whose primary work is to detect and segment repetitive motions using the geometrical constraints generated by the same motion repeatedly when the viewpoint changes. Ormoneit *et al.* [10] used functional analysis to represent cyclic movement. Ribnick *et al.* [8], [20] found that it is possible to reconstruct accurately periodic movements in 3D from a single camera view. Based on this research, they applied 3D reconstruction to gait recognition. Ren *et al.* [21] and Li *et al.* [22] developed two autocorrelation counting systems based on matching visual descriptors. Although both systems completed the repetition counting, they are postprocessing methods, which are only applicable to specific domains of restricted video. Pogalin *et al.* [17] got the information of a certain part of the body by tracking the interesting object, then performed PCA and spectral analysis followed by detection and frequency measurement. But its purpose is to estimate the degree of periodic motion but not to count the repetition. Based on the human skeleton points captured by Kinect, Wang *et al.* [7] proposed an unsupervised repetitive motion segmentation algorithm based on the frequency analysis of the motion parameter, zero-velocity cross detection and adaptive k-means clustering. Although the above methods lay the foundation for the video repetitive counting task, they only realized the simple repetitive motion estimation of a fixed scene and had a poor performance on the diversity and non-stationary motion, which commonly exists in real applications.

In recent years, Kumdee *et al.* [3] used the image self-similarity measure as the input of the multi-layer perceptron neural network to determine whether the input video is a repetitive action. This method is relatively stable to image changes, noise, and low-resolution images. However, they focused on classifying that the video is a repetitive video or not but not counting. Levy *et al.* [13] proposed a method to count the repetitive action of the videos using the convolutional neural network. They used synthetic data to simulate four motion types for the periodic motion and carried out network training and prediction. In the test, the region of interest was calculated through the motion threshold for the test data. The motion cycle was classified through the classification network to complete the repetitive counting task. The method showed excellent performance on YT_Segments dataset. However, their algorithm decayed a lot when there are actions with different repetitive modes from the trained modes. The wavelet transform was presented in [14] to better deal with more complex and diverse video dynamics. From the flow field and its differentials, they derived different repetitive perceptions. Based on the gradient, curl and divergence, a motion foreground segmentation representation based on flow was realized, and remarkable results were obtained. In their recent work [33], they improved their foreground motion segmentation method by obtaining the segmentation results directly from the wavelet filter responses and obtained a significantly improved performance. However their methods need the foreground segmentation, which is also a difficult problem. Therefore, we propose a method without extra preprocessing.

B. Deep Features for Action Analysis

CNNs have been widely used in action recognition. Some of these CNNs use deep architectures with 2D convolutions to extract translation-invariant features in the video frames [15]. Specifically, Karpathy *et al.* [15] first introduced a CNN based method for action recognition and organized a large-scale sports video dataset (i.e., Sports-1M dataset) for training deep CNNs. To model the temporal information of the action, two-stream based CNN learning framework [16], [28] has been

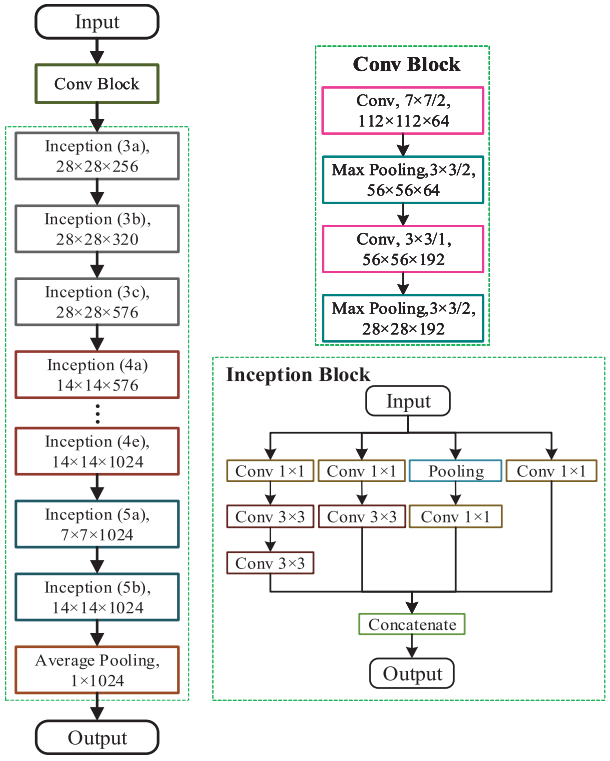


Fig. 3. The framework of BN-Inception Network.

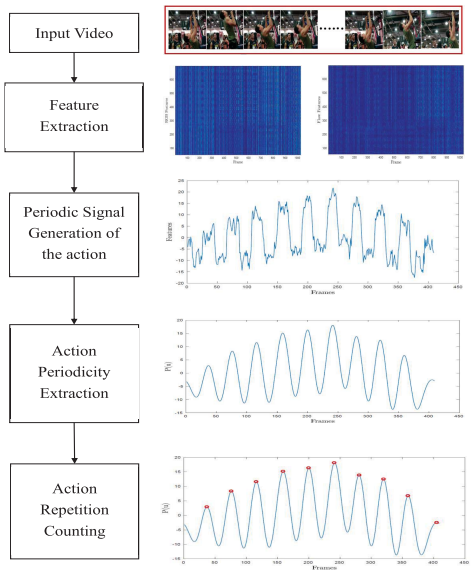


Fig. 4. The framework of action repetition counting.

proposed. The two streams mean the spatial stream represented by RGB values and the temporal stream represented by pre-computed optical flow features. Because of the excellent balance between efficiency and effectiveness, the BN-Inception Network [23] is used as the backbone of the framework. The prominent characteristic of BN-Inception network is the Inception module, which carries out multi-scale processing and fusion of image features to extract better feature representation. Moreover, it is well known that the 3×3 convolutional kernel has the best performance in VGG [24], and a very effective Batch Normalization (BN) method has been proposed to accelerate the learning of data distribution during training, making the accuracy of the classification improve significantly. In addition, the deep ConvNets can take pictures with any form as input to extract features. The training of deep ConvNets requires a large number of training samples to achieve good performance in action modeling. Nowadays, a large number of publicly available video datasets provide great convenience. Therefore, we extract the deep features of our experimental data using BN-Inception Network in this paper.

III. Action Repetition Counting

In this part, we will discuss the proposed algorithm. As shown in Fig. 4, our algorithm includes four steps. Firstly, deep features of the unconstrained videos are extracted using deep ConvNets. Secondly, based on the high-dimensional deep features, the periodic signal is generated using PCA, and a one-dimensional waveform can be obtained to reflect the repetitive changes of the videos. Thirdly, the action repetition rules are extracted to reconstruct the periodic waveform of the repetitive action using a signal trend removal scheme with polynomial regression and the high energy rules extracted from the 1D waveform that combines both Fourier analysis, power spectral analysis and the inverse Fourier transform. Finally, using the waveform, peak detection is used to count the repetition.

A. Deep Feature Extraction Based on BN-Inception ConvNets

The Inception v2 based on Batch Normalization network [15] is used to obtain the features of the action, as shown in Fig. 3. It was pre-trained on the large public Kinetics-400 dataset [25], which contains 300,000 clip videos from real scenes, including 400 action categories, which is a widely used dataset for action recognition. The pre-trained models used in this paper are provided in [28].

Two networks are used to extract robust action features, operating on two components, spatial and optical flow separately. The spatial flow network operates on the RGB image, which extracts spatial features describing the scene and object information. The optical flow network takes the pre-computed optical flow images as the input to extract the temporal features, which describe the motion information of the video. Robust spatiotemporal features are extracted by this method. Fig. 5 shows a diagram for feature extraction of the action.

Clipping and rotation of training images, to decrease the influence of the noise and increase the stability of features, are used to get the image set. In the process of feature extraction, we take the image set as network input and get the features in the avg_pool layer. Then the summation and the average features are computed in each dimension. Finally, the spatial features, denoted by f_s, and temporal features, denoted by f_t are extracted for the single image, respectively. We further get their fusion features f_f via concatenate operation by equation 3. Due to the designed structure of the network, the dimension of the spatial features and temporal features is 1024, as shown in equations 1 and 2, and therefore, the dimension

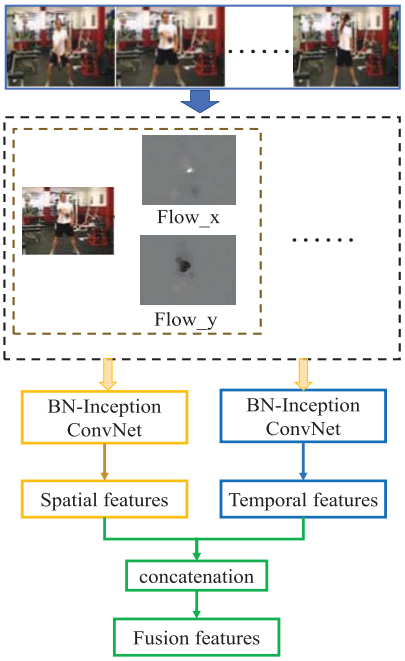


Fig. 5. The diagram of the deep feature extraction. The input modalities of BN-Inception are the RGB images and optical flow fields (x, y directions).

of the fusion features is 2048.

$$f_s = (s_1, s_2, \ldots, s_{1024}) \tag{1}$$

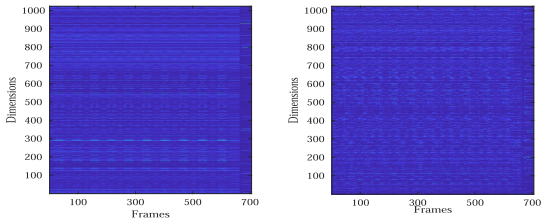
$$f_t = (t_1, t_2, \ldots, t_{1024}) \tag{2}$$

$$f_f = (s_1, s_2, \ldots, s_{1024}, t_1, t_2, \ldots, t_{1024}) \tag{3}$$

As discussed above, deep features with different modes can be obtained using the pre-trained models and we do not need to retrain the model using repetitive actions. The extracted deep features of the video with different modes can be represented by three 2D matrices with the shape of $N \times D$ (where N is the total number of the video frames and D (i.e. 1024 or 2048) is the dimension of the features). Taking the spatial features of the video as an example, it can be marked as $F_s = \left[f_{s0}, f_{s1} \cdots f_{s(N-1)}\right]$, where f_{si} denotes the spatial features of the $(i+1)$-th frame of the video according to equation 1. Similarly, the temporal features and fusion features of the video can be marked as $F_t = \left[f_{t0}, f_{t1} \cdots f_{t(N-1)}\right]$ and $F_f = \left[f_{f0}, f_{f1} \cdots f_{f(N-1)}\right]$, respectively. The spatial and temporal features are visualized in Fig. 6, where spatial features are given in Fig. 6(a), and temporal features are given in Fig. 6(b). From the visualization results, although we can find some specified patterns, it is difficult for us to find the periodic rules using this high-dimension features. Therefore, we need some other methods to mine the periodicity of the repetitive action.

B. Periodic Signal Generation

To extract the periodicity information, we mine the hidden periodic action rules from different features, including the spatial features F_s, the temporal features F_t, and the fusion features F_f. The mining method for these features is the same; therefore, we take spatial features F_s as an example to explain our extraction method.



(a) Spatial features (b) Temporal features

Fig. 6. The visualization of deep features.

Although deep features can be used to classify the actions well, the counting of the repetition of the action is totally different from the classification of the action. For classification, one action is considered as a whole, and it focuses on the difference between different actions. On the contrary, action repetition counting focuses on locating the repetition of the same motion pattern. Therefore, the deep features extracted directly for action recognition may not be suitable for counting. We transform the high-dimensional features into an intuitive waveform by extracting the primary component of its covariance matrix.

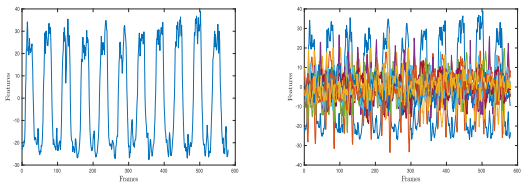
For the feature matrix $F_s = \left[f_{s0}, f_{s1} \cdots f_{s(N-1)}\right]$, we obtain its mean matrix $\bar{F}$, and construct the transformation matrix $\hat{F}$ using $\hat{F} = F_s - \bar{F}$. Then, the covariance matrix is calculated according to equation 4.

$$COV = \frac{1}{D}\hat{F}\hat{F}^T = V \Lambda V^T \tag{4}$$

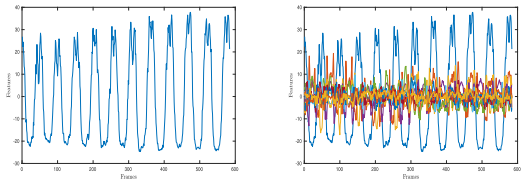
We also can compute the eigenvalues and eigenvectors of the covariance matrix. The corresponding results are separately denoted by their matrix form as $\Lambda = diag\,(\lambda_1, \lambda_2 \cdots \cdots \lambda_D)$ and $V = [\mu_1, \mu_2 \cdots \mu_D]$, where each μ_i is a vector with dimension D. We arrange Λ according to the value of its eigenvalue from large to small. And according to the new order of eigenvalues, we rearrange V to V' in columns. Then, we reserve the first eigenvector to get the transformation matrix V_1'. The size of V_1' is $D \times 1$. Therefore the mapped matrix $P = (p_0, p_1, p_2 \ldots p_{(N-1)})$ can be computed according to formula 5, where p_i is the principal component of $(i+1)$-th frame of the videos, $i = 0, 1, \ldots N-1$. The size of P is $N \times 1$, by which the high-dimensional video features are transformed into the new space constructed by 1D waveform, as shown in Fig. 7.

$$p_i = V_1'^T f_i \tag{5}$$

To analyze the effect of different principal components, we also compute the first 10-dimensional principal component transformation matrix V_{10}', the size is $D \times 10$. For each dimension, we get the mapped vector separately, the visualization results are shown in Fig. 7. From this figure, we can see that the first-dimensional feature includes more information on the motion characteristics of repetitive actions. Therefore, in this paper, the first-dimensional principal component is used to count the repetitive action.



(a) The first 1D of the component of the temporal features (b) The first 10D of the component of the temporal features



(c) The first 1D of the component of the spatial features (d) The first 10D of the component of the spatial features

Fig. 7. Periodicity visualization of the repetitive action.

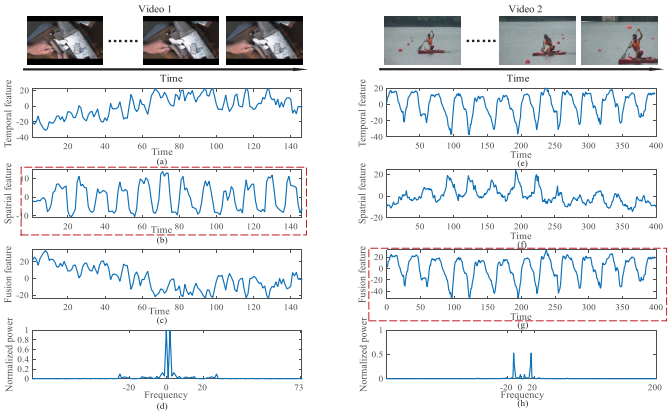


Fig. 8. Various feature modes under different circumstances. Video 1 represents a video with a relatively static background and Video 2 represents a video with a relatively dynamic background. The left one denotes the video 1 and the right one denotes the video 2. Taking the video 1 as an example, Fig. 8(a), Fig.8(b) and Fig.8(c) show the 1D waveform mined from the spatial, temporal and fusion features, respectively, which can be obtained by the method proposed in Section III-B. Fig. 8(d) show the Power Spectral Density(PSD) of the video 1.

C. Energy-Based Adaptive Feature Mode Selection

We found that different feature modes are robust to different types of videos. For example, the spatial features are usually robust to the video with a static background, and the temporal features or fusion features are usually robust to the video with a dynamic background. As shown in Fig. 8, for the video with a relatively static background (e.g. video 1), the 1D waveform mined from the spatial features can better reflect the periodicity of the repetition actions than that of the temporal feature, but for the video with a relatively dynamic background (e.g. video 2), the 1D waveform mined from the temporal features can better reflect the periodicity of the repetition actions than that of the spatial feature.

Besides, as shown in Fig. 8, we find that the high peaks of the video with a relatively static background usually lie in a relatively low frequency of their Power Spectral Density (PSD) while the high peaks of the video with a relatively dynamic background usually lie in a relatively high frequency of their PSD. Based on this insight, we propose an energy-based feature mode selection scheme using PSD to adaptively select feature mode, which can be described as follows.

Given a 1D waveform of a video as $x[t]$ $(t = 0, 1, \cdots, N-1)$ that can be mined from the fusion features, containing both the spatial features and temporal features, by the proposed method in Section III-B. Firstly, we obtain the PSD of the video. Specifically, the time-varying signal is decomposed into the superposition of the components in the frequency domain by Fourier transform. The vibration frequency of the waveform is separated to get the spectrum by equation 6. Secondly, the PSD of $x[t]$ can be calculated by equations 7 as $S[k]$. Based on the $S[k]$, high peak detection is applied to locate the frequency coordinate of the highest peak, and we mark it as p. Finally, the final features of this video can be obtained by equation 8.

$$X[k] = \sum_{t=0}^{N-1} x[t] e^{-j\frac{2\pi kt}{N}} (k = 0, 1, \cdots, N-1) \tag{6}$$

$$S[k] = \frac{(X[k])^2}{\max\limits_{0 \le k_1 \le N-1} \{(X[k_1])^2\}} \tag{7}$$

$$F_{final} = \begin{cases} F_s, & |p| < T_1 \\ F_t, & |p| > T_2 \\ F_f, & otherwise \end{cases} \tag{8}$$

where T_1 and T_2 are two thresholds that can be manually defined by the user. In experiments, we set the T_1 and T_2 to 2 and 20, respectively.

D. Action Periodicity Reconstruction and Repetition Counting

Due to the complexity and diversity of the videos captured in the real scene and the non-standardization during the action execution, the principal component contains lots of noises. As shown in Fig. 7(a), although there are some repetitive motion rules in the figure, the lower peak and the noises may lead to poor performance when counting. To locate the repetitive action, we need to distinguish the interesting actions from the unrelated noises. As discussed above, in the real challenging scenes, the video with repetitive actions usually has the following characteristics. (1) The noise caused by the background noise and view changes usually reflects as a time-varying DC component of the 1D waveform of the video. (2) The coupled motion is easily double counted due to the different motion frequencies or the self-similarity between the main action and its sub-actions, leading to large additional counting errors. (3) The interesting repetitive actions, including the main action and its sub-actions, usually carry more energy with relatively high frequency than the other unrelated movement. In the following sections, we first propose a signal trend removal scheme to eliminate the effect of the time-varying DC component, to a great extent, caused by the background noise and the view changes of the camera. Then, we propose a high-energy-based complex action detection scheme to extract the robust periodic signal of the repetitive

action by mining the relationship between the main action and its sub-actions.

1) Signal Trend Removal: The changing of the feature signal comes from two main sources: one is the changes of the original action itself, the other is the drifting or the changing of the background and the viewpoint of the cameras. The former variation is useful, but the latter one is the disturbance for the description of the action. The latter changing is often corresponding to the time-varying DC component, that is to say, it can reflect the general trend of the feature. Therefore, we name this changing as the signal trend, as shown in Fig. 2. The signal trend is often unrelated to the action. Therefore, we need to remove it from the feature signal. Here, we propose a signal trend removal scheme by subtracting this signal trend. Specifically, we first apply polynomial regression to simulate this signal trend as $b[t]$ by equation 9, and the parameters of the signal trend, W, can be learned by minimizing the loss L_b according to the equation 10. Then, we obtain the filtered signal $P_f[t]$ based on the simulated signal trend $b[t]$ by equation 11.

$$b[t] = b(t, W) = w_0 + w_1 t + w_2 t^2 + \cdots + w_M t^M = T_b W \tag{9}$$

$$L_b = \frac{1}{2} \sum_{t=0}^{N-1} (b[t] - P[t])^2 \tag{10}$$

$$P_f[t] = P[t] - b[t] \tag{11}$$

where $T_b = [1, t, \cdots, t^M]$, $W = [w_0, w_1, \cdots, w_M]^T$, $P[t]$ is the generated periodic waveform mined from the final feature F_{final} using the method proposed in Section III-B, N is the number of the video frames, and M is the order of the polynomial.

2) High-Energy-Based Action Periodicity Reconstruction and Repetitive Counting: The aims of this section are two-fold: (1) locate the interesting repetitive actions of the video, containing the main action and its sub-actions; (2) reconstruct the periodic waveform of the video.

a) Locate the Interesting Repetitive Actions of the Video: Because the main energy of the video with relatively high frequency comes from the interesting repetitive actions, we propose a two-stage threshold filter scheme to locate the interesting repetitive actions of the video, including the main action and its sub-actions, using the Fourier transform and the PSD of the corresponding video.

For the first stage, we propose a high-energy-based action location scheme to roughly locate the interesting action. Specifically, using the filtered waveform $P_f[t]$ of the video obtained in Section III-D.1, we first obtain their frequency spectrum and power spectrum by equations 6 and 7, respectively, marking as $X_i[k]$ and $S_i[k]$, respectively. Then, we set a threshold θ_1 to define the boundary between the interesting actions and the unrelated actions in the power spectrum of the video. Finally, as shown at the middle of Fig. 9, the filtered frequency spectrum $X'_f[k]$ and the power spectrum $S'_f[k]$ of the interesting actions can be located with the main-energy by filtering the low-energy power spectrum of the unrelated

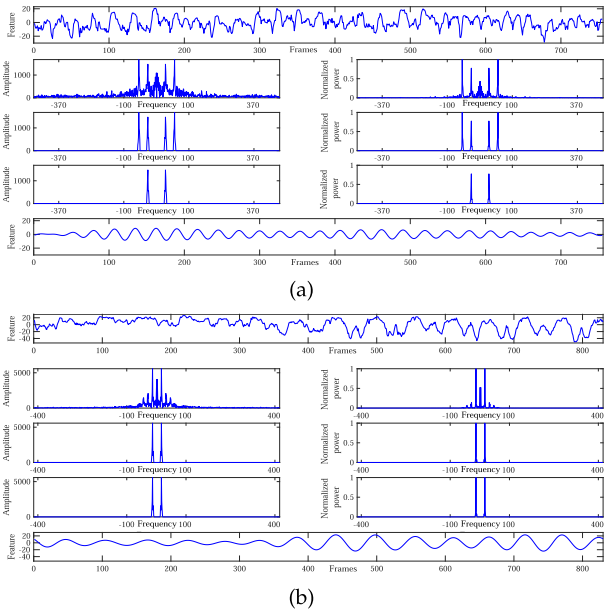


Fig. 9. The results of high-energy-based action periodicity reconstruction. In Fig. 9(a) or Fig. 9(b), we show the diagram of an example video with or without coupled motion, respectively. For each figure, we first show the waveform of the first 1D component of our adaptive features generated in Section III-B and its corresponding frequency spectrum and power spectrum at the top of the figure; then, we show the frequency spectrum and power spectrum of the located interesting repetitive actions and our reconstructed periodicity action waveform, respectively; finally, we show our final reconstructed periodic waveform.

motion according to equations 12 and 13.

$$X'_f[k] = \begin{cases} X_i[k], & S_i[k] \geq \theta_1 \\ 0, & S_i[k] < \theta_1 \end{cases} \tag{12}$$

$$S'_f[k] = \begin{cases} S_i[k], & S_i[k] \geq \theta_1 \\ 0, & S_i[k] < \theta_1 \end{cases} \tag{13}$$

$$X_f[k] = \begin{cases} X'_f[k], & |k| \geq \theta_2 \\ 0, & |k| < \theta_2 \end{cases} \tag{14}$$

$$S_f[k] = \begin{cases} S'_f[k], & |k| \geq \theta_2 \\ 0, & |k| < \theta_2 \end{cases} \tag{15}$$

where $k = 0, 1, \cdots, N-1$, N is the number of video frames, θ_1 is a threshold to define the low energy and θ_2 is a threshold to define the low frequency. In experiments, θ_1 and θ_2 are set to $0.5\% * (\max\{S_i\})$ and $0.15\% * \frac{N}{2}$, respectively.

For the second stage, based on the frequency spectrum $X'_f[k]$ and the power spectrum $S'_f[k]$, we further filter the noise with low energy and relative low frequency by equations 14 and 15.

b) Reconstruct the Periodic Waveform of the Video: Because the located interesting repetitive actions of the video contain both the main action and its sub-actions, and as discussed above, the sub-actions will cause miscounting errors to a great extent. Therefore, to reconstruct the accurate periodic waveform of the video, it is important to detect the sub-actions of the video and eliminate the information of the sub-actions.Therefore, using the frequency spectrum $X_f[k]$ and power spectrum $S_f[k]$ of the located interesting repetitive

actions above, the reconstruction detail of the periodic waveform of the video can be summarized as follows.

Coupled Motion Detection and counting(CMD). For one action including the coupled motion, the occurrence times of its sub-action usually are the multiple of the times of the main-action. Therefore, we can easily detect it using frequency analysis. Specifically, we first use spectral decomposition to extract the dominant frequencies. Then the inverse Fourier transform is used to obtain the temporal waveform of the corresponding frequency. Peak detection is adopted to compute the times of the signal with the corresponding frequency. If the times of the signal with one frequency is the multiple of the signal with the other frequency, then the original signal contains coupled motion. The least times is the occurrence times of the main action, and the repetition counting of the action can be obtained.

The detailed process is as follows. As shown in Fig. 10 (b), the power spectrum $S_f[k]$ contains multiple groups of high peaks. Firstly, as shown in Fig. 10 (c) and Fig. 10 (d), we use spectral decomposition to obtain the signals with the single group frequency $\{S_{fi}[k]\}$ and their corresponding frequency spectrum as $\{X_{fi}[k]\}$. Secondly, the Inverse Fourier Transform is applied to reconstruct the corresponding waveform of the single group of high peaks by equation 16, respectively, as shown in Fig. 10 (e) and Fig. 10 (f). Thirdly, peak detection is applied to calculate the number of periodicity of the waveform. Finally, we detect whether there exists the integer quantitative relationship among the periodicity numbers of different waveforms. For example, suppose the periodicity number of the waveform in Fig. 10 (e) and Fig. 10 (f) are N_1 and N_2, respectively, if the remainder of $\frac{N_1}{N_2}$ or $\frac{N_2}{N_1}$ is 0, the relationship between the Fig. 10 (e) and Fig. 10 (f) is the main action and the corresponding sub-actions, and the frequency spectrum and power spectrum of the reconstructed waveform can be obtained by removing the sub-actions, as shown at the bottom of Fig. 9(a). In this case, the final counting results are the periodicity number of the main action, and it can be obtained using the highest peak detection with the reconstructed waveform shown in Fig. 9(a).

$$z_i[t] = \frac{1}{N}\sum_{k=0}^{N-1} X_{fi}[k] e^{j\frac{2\pi kt}{N}} \tag{16}$$

where $z_i[t]$ is the waveform signal of the i-th single group of high peaks.

Counting for the action without the coupled motion. If the action does not consist of the coupled motion, the counting for the repetition is easy. In this case, since the dominant energy of the video comes from the repetitive action, we can use the highest energy rule to directly locate the repetitive action. Specifically, as shown in the middle of Fig. 9(b), we first decompose the spectrum of the video by detecting the group of the highest peak in their power spectrum. Then, we use Inverse Fourier Transform to reconstruct the periodic waveform of the video using the corresponding spectrum, as shown at the bottom Fig. 9(b). Finally, the high peak detection is applied to obtain the final counting results of the video.

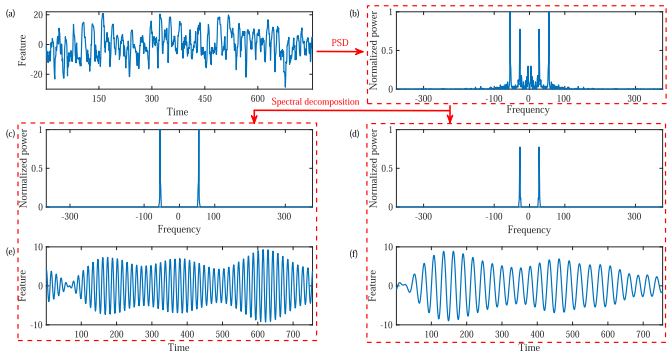


Fig. 10. The diagram of the detection and counting for the action with coupled motions. In Fig. 10(a), we give a signal features extracted from the video using deep features and principal component analysis. In Fig. 10(b), we give its PSD. In Fig. 10(c) and (d), using spectral decomposition, we can decompose the original PSD into two signals with frequencies of two groups. In Fig. 10(e) and (f), we give the inverse Fourier Frequency Transformation results of (c) and (d).

IV. Experiments

YT_Segments [13], QUVA [14], and Real-world Action Repetition Videos (RARV), a new proposed dataset, are used to evaluate our algorithm. These data are diverse and challenging, and they also include the movement of the camera and background. The repetitive actions have varying lengths and complicated appearance patterns.

A. Datasets

1) YT_Segments: It contains 100 videos with repetitive actions, including exercise, cooking, architecture, and so on. To create a clean benchmark, the videos are pre-split and only contain repetitive actions. The number of repetitive motion is pre-labelled. The smallest and largest numbers of the repetition are 4 and 50, respectively. The average duration of one video is 14.96s. Meanwhile, there are 30 videos with varying degrees of camera movement.

2) QUVA: It's also made up of 100 videos and shows various kinds of repetitive video dynamics, including swimming, stirring, cutting, and so on. Compared with the YT_Segments dataset, it has more challenges in cycle length, motion appearance, camera motion, and background complexity. Therefore, the dataset is a more realistic and challenging benchmark for estimating repetitive action.

3) RARV: It contains 200 videos in total with diverse backgrounds, including static backgroud, dynamic background, etc. The dataset is built by merging data from both YT_Segments and QUVA datasets. In the new dataset, we can comprehensively evaluate our model under the assumption of the unknown background in advance.

B. Evaluation Metrics and Baselines

1) Metrics: We use the same evaluation criteria [13] as those that used in the baselines as the metric for this task. For N videos, we calculate the Mean Relative Error(MRE) $\pm$ standard deviation (σ) [13] as the evaluation metrics, where

TABLE I
COMPARISONS WITH THE STATE-OF-THE-ART BASELINES

Methods \ Datasets	YT_Segments	QUVA	RARV
Pogalin et al. [17]	21.9 ± 30.1	38.5 ± 37.6	30.2 ± 33.9
Levy & Wolf [13]	**6.5** ± 9.2	48.2 ± 61.5	27.4 ± 35.4
Runia &Snoek [14]	10.3 ± 19.8	23.2 ± 34.4	16.8 ± 27.1
Runia et al. [33]	9.4 ± 17.4	26.1 ± 39.6	—
Ours	9.6 ± **8.6**	**19.9** ± **33.5**	**14.3** ± **18.9**

G is the ground truth and R is the predicted value

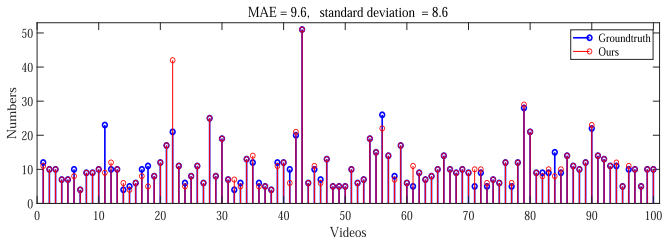
$$MRE = \frac{1}{N}\sum_{i=1}^{N}\frac{|G_i - R_i|}{G_i}$$
$$\sigma^2 = \frac{1}{N}\sum_{i=1}^{N}(G-R)^2 \tag{17}$$

2) Baselines: We compare our method with one classical method [17] and two recent methods in [13], [14]. When reporting the results, we directly make use of the results provided in the paper [14].

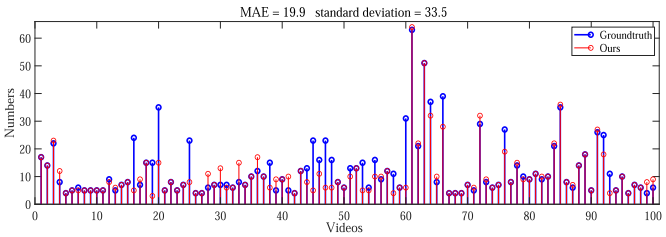
C. Comparisons With State-of-the-Art Baselines

1) Overall Results and Analysis: We compare the performance of our method with state-of-the-art baselines, and the results are presented in Table I. Compared with the baselines, our method achieves superior or comparable performance without extra preprocessing. (1) For the YT_Segments dataset, the method of [13] performs best with the MRE of 6.5. Our method is superior to articles [14], [17] with the MRE of 9.6, and we achieve the best standard error compared with the above methods, which illustrates that the worst performance of our method is the best in all the methods. Although the MRE of [33] is slightly better than that of our method, their standard errors are significantly worse than that of our method, showing the effectiveness of our method. The results show that our method can achieve good performance under the relatively static background. (2) In the more challenging QUVA dataset, our experimental results achieve the best performance at both the MRE and the standard error. The method [13] performed the worst with the MRE of 48.2, because their network considered only four types of action during training. The method of [17] was 38.5. In [14] and [33], the MRE was 23.2 and 26.1, respectively. These results show that our method can also adapt well to the dynamic backgrounds. (3) For the dataset with more diverse backgrounds (i.e. RARV), compared with the baselines, our method achieve the lowest MRE and standard error by a large margin. For example, compared with the the best baseline [14], the MRE of our method decreases by 2.5, and the standard error deceases by up to 8.2, showing the effectiveness of our proposed method powerfully.

In summary, compared to the above methods, we get the best standard error on the three public datasets. At the same time, we get the lowest MRE on most datasets. The results show that our method can achieve superior or comparable results in counting action repetition for unconstrained videos with a decent framework, not relying on preprocessing.



(a) The counting results on YT_Segments

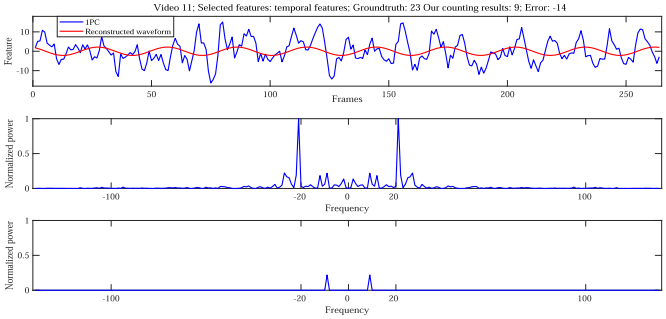


(b) The counting results on QUVA

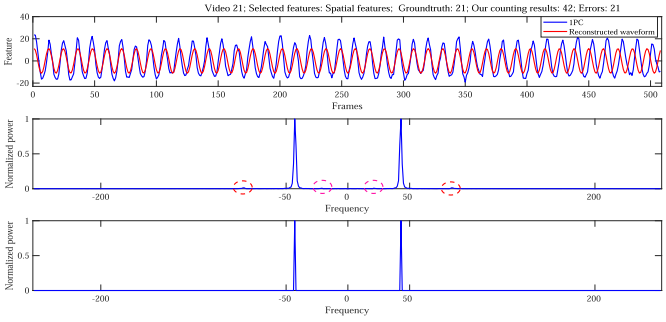
Fig. 11. The detailed counting results of our method.

2) Detailed Results and Analysis: To further validate our contribution, on YT_Segments and QUVA (Since the videos on RARV come from both the YT_Segments and QUVA datasets, we here only give the detailed analysis on YT_Segments and QUVA, respectively), we give the counting results in detail, as shown in Fig. 11. From Fig. 11(a), we can see that the counting results of most of the actions are very close to their groundtruth and the differences cluster around positive or negative 1. Our peak-detection counting scheme can well explain this. Because we use the number of the peaks as the counting results, it is slightly different from the repetitive case, where the repetition is, in fact, a cycle. Therefore, using more detailed cycle detection may solve this problem. In addition, there are some videos (e.g. video 11 and video 22) whose errors are relatively large. In video 11, as shown in Fig. 12(a), the energy of the reconstructed waveform is too low, and therefore this may be caused by some noise. The possible reason is: during the process of the coupled motion detection, the frequency of the noise is similar to that of the sub-action while there exists no sub-action actually, leading to error detection. In video 22, as shown in Fig. 12(b), this is a coupled motion, but the power spectrum of the action mainly contains one group of high peaks. The possible reasons are two-fold: (1) the high peaks of the main action and its sub-action are coupled together; (2) the high peaks of the main action are with low energy. During the reconstruction of action periodic waveform, the high peaks with low energy are removed, and the coupled high peaks can not split using our proposed method. In this case, we can not reconstruct the real periodic waveform of the main action, leading to double-counting.

From Fig. 11(b), besides the above mentioned negative and positive 1 difference problem, there are also other problems. We found the main error occurs on this challenging dataset becasue the interesting action is associated with multiple objects, and these objects move periodically with these actions. As shown in Fig. 13, their corresponding power spectrum



(a) Video 11

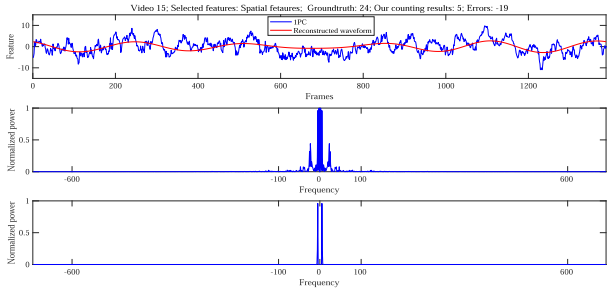


(b) Video 22

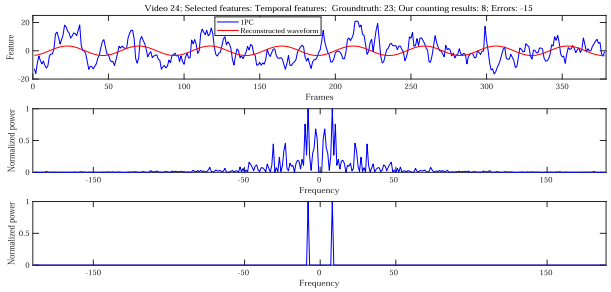
Fig. 12. Analysis of failed counting cases on YT_Segments. For each figure, the top one denotes the periodic waveforms before and after reconstruction using our proposed action periodic mining rules; the middle one denotes the power spectrum of the first 1D component of our adaptive features generated in Section III-B (1PC waveform); the bottom one denotes the power spectrum of our final reconstructed waveform, it is the same in Fig. 13.

contains multiple high peaks with high energy. After the waveform reconstruction with our proposed algorithms, some high-peaks will be removed, remaining only one group of high peaks. Therefore, much useful information has been filtered, leading to poor performance. For example, for a 'shoveling snow' action in video 16, the snow, the shovel, the feet of human, and the camera are periodic moving together. But after periodic mining with the proposed method, only one group of high peaks is left. In the future, we will focus on this.

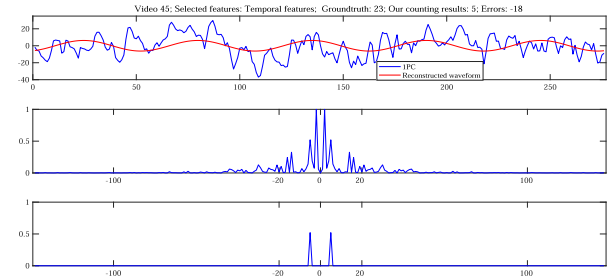
In summary, the main error of our method lies in two aspects. One is the difference between the peak and the cycle. The other is the coupled submovements in action and the movement with their associated objects.



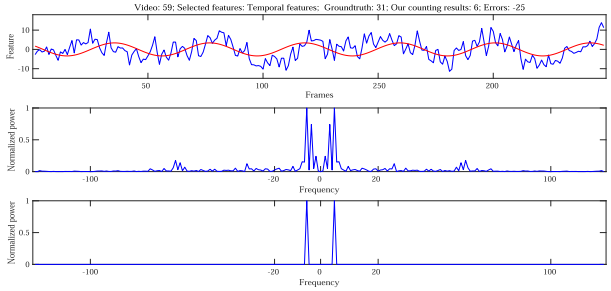
(a) Video 16



(b) Video 25



(c) Video 45



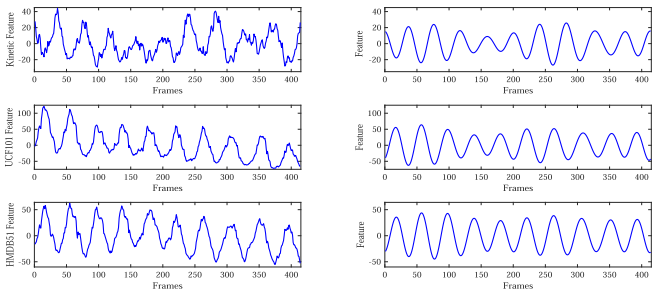
(d) Video 60

Fig. 13. Analysis of failed counting cases on QUVA.

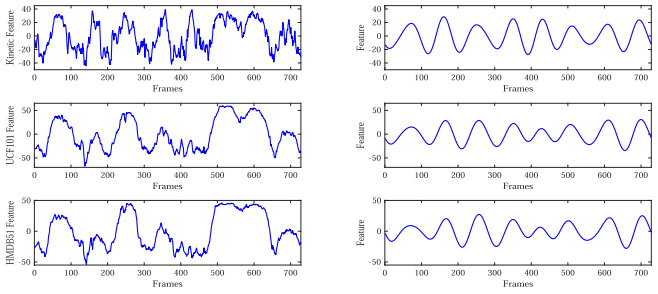
D. Ablative Analysis

In this section, we conduct extensive ablative experiments to show the effectiveness of our proposed method. Specifically, we first conduct a series of experiments to analyze what has a greater impact on our superior performance. Then, we verify the effectiveness of energy-based adaptive feature mode selection scheme (EAFS), signal trend removal scheme (STR), and coupled motion detection scheme (CMD), respectively.

To show what factors have a greater impact on our superior performance, we conduct two group experiments: (a) Evaluation of deep features: we extract deep features using the pre-trained models on different datasets (i.e. HMDB51, UCF101, and Kinetics-400). In this case, we show how do deep features influence the results of our model. The results are shown in Table II and Fig. 14. As shown in Table II, for the YT_Segments dataset, the deep features using the pre-trained model on UCF101 achieve the lowest MRE, but their standard error is higher than that of the deep features with Kinetics-400 pre-trained model; for both the QUVA and RARV datasets, the deep features using the pre-trained model on Kinetics-400 obtain both the lowest MRE and the lowest standard error. Although the features using the pre-trained models on the larger dataset such as Kinetic-400 can achieve superior performance, their performance difference is limited. This shows our proposed method does not heavily rely on different deep



(a) YT_Segments



(b) QUVA

Fig. 14. Visualized results of analyzing factors on our superior performance. In Fig. 14(a) and Fig. 14(b), the left one denotes the waveforms of the first 1D component of our adaptive features generated in Section III-B, the right one denotes the corresponding waveforms reconstructed with our proposed method in Section III-D.2.

TABLE II

EVALUATION OF DEEP FEATURES

Features \ Datasets	YT_Segments	QUVA	RARV
HMDB51 features	10.0±13.1	20.5±51.2	15.3±32.1
UCF101 features	**9.0**±12.7	21.0±43.9	14.4±26.2
Kinetic-400 features	9.6± **8.6**	**19.9**±**33.5**	**14.3**± **18.9**

features, and the superior performance benefits from our proposed action periodic mining rules to a great extent. Moreover, as shown in Fig. 14, the reconstructed waveforms using the deep features with various pre-trained models are very similar, showing the effectiveness of our proposed method again.

(b) Evaluation of our proposed action periodic mining rules (APMR): based on the generated waveform in Section III-B, we obtain the counting results with the high peak detection, and the results are reported in Table III and Fig. 14. As shown in Table III, compared with the results of "w/o APMR", the errors of our method decrease significantly, demonstrating the effectiveness of our proposed method again. The reasons are: as shown at the left of Fig. 14, the generated waveforms extracted from the deep features contain lots of noises. Therefore, it is not suitable for counting directly, and it needs more robust algorithms to mine the action periodic rules. By contrast, as shown at the right of Fig 14, the reconstructed waveforms with our proposed method are smooth and can better reflect the periodicity of repetitive action.

To show the effectiveness of the energy-based adaptive feature mode selection scheme (EAFS), we conduct three experiments by using spatial features, temporal features and

TABLE III

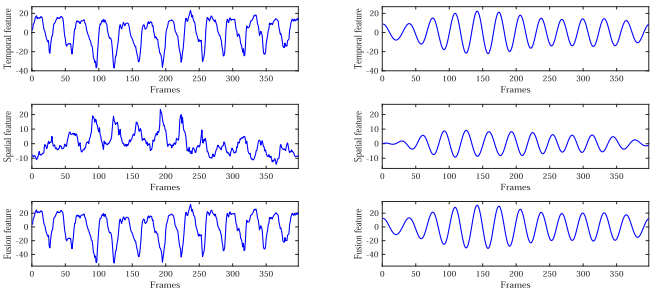
EVALUATION OF OUR PROPOSED ACTION PERIODIC MINING RULES, WHERE "HFEAT", "UFEAT" AND "KFEAT" DENOTE THE FEATURES EXTRACTED USING THE PRE-TRAINED MODEL ON HMDB51, UCF101 AND KINETICS-400 DATASETS, RESPECTIVELY

Features \ Datasets	YT_Segments	QUVA	RARV
w/o APMR (Hfeat)	94.5±3483.3	100±13164.8	95.4±7007.0
Ours (Hfeat)	**8.2**± **7.7**	**21.5**±**46.8**	**14.9**± **27.2**
w/o APMR (Ufeat)	96.0±3296	97.0±10346.9	96.5±6821.7
Ours (Ufeat)	**7.6** ±**7.4**	**20.2**±**37.9**	**13.9**±**22.7**
w/o APMR (Kfeat)	98.2± 4841.7	96.2± 10530.6	99.1±9003.2
Ours (Kfeat)	**9.6**± **8.6**	**19.9**±**33.5**	**14.3**± **18.9**

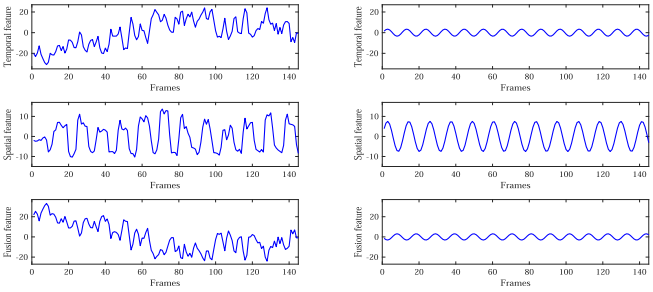
TABLE IV

EVALUATION OF ADAPTIVE FEATURE MODE SELECTION

Features \ Datasets	YT_Segments	QUVA	RARV
w/o EAFS (F_s)	11.6±12.2	24.3±52.9	21.0±35.6
w/o EAFS (F_t)	12.7±10.9	20.0±34.9	16.3±22.9
w/o EAFS (F_f)	11.4±14.6	20.0±34.9	15.7±24.7
Ours	**9.6**±**8.6**	**19.9**±**33.5**	**14.3**± **18.9**



(a) YT_Segments



(b) QUVA

Fig. 15. Visualized results of adaptive feature mode selection, where the left one of Fig. 15(a) and Fig. 15(b) denotes the waveforms of the first 1D component of our adaptive features generated in Section III-B, the right one denotes the corresponding reconstructed waveforms using our proposed method in Section III-D.2.

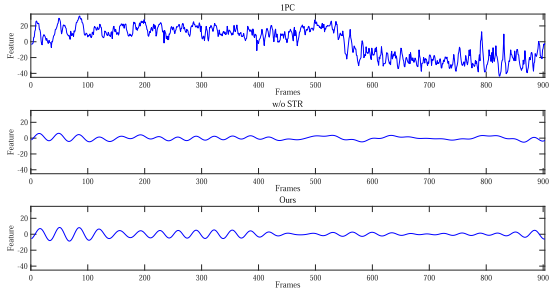
fusion features, respectively. The experimental results are reported in Table IV and Fig. 15. As shown in Table IV, without considering the background and simply using one of the feature modes for all videos will decline the performance of our model, showing the effectiveness of our proposed EAFS scheme that adaptively selects the proper feature mode according to the background of the video. As shown in Fig. 15, the reconstructed periodic waveforms using different features are very different. For example, for the video of Fig. 15(a), the temporal features or fusion features may be better than the spatial features for accurate counting, but for the video of

TABLE V
EVALUATION OF SIGNAL TREND REMOVAL

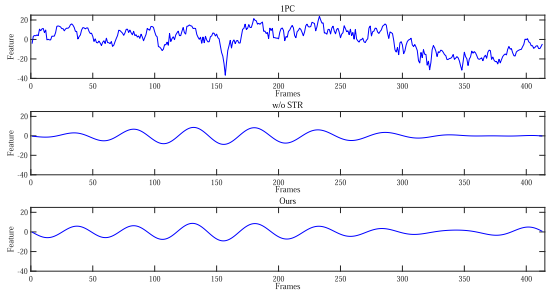
Features \ Datasets	YT_Segments	QUVA	RARV
w/o STR	10.9 ± 13.2	22.8 ± 56.0	16.8±34.6
Ours	**9.6± 8.6**	**19.9±33.5**	**14.3± 18.9**

TABLE VI
EVALUATION OF COUPLED MOTION DETECTION

Features \ Datasets	YT_Segments	QUVA	RARV
w/o CMD	10.8 ± 9.4	33.9± 146.4	21.5±77.9
Ours	**9.6±8.6**	**19.9±33.5**	**14.3± 18.9**

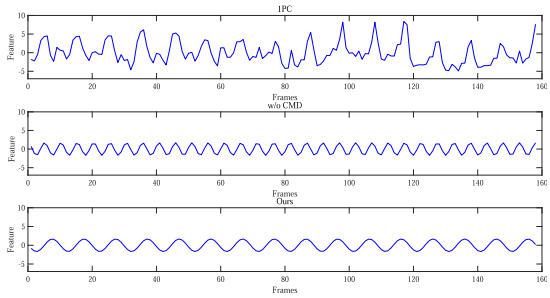


(a) YT_Segments

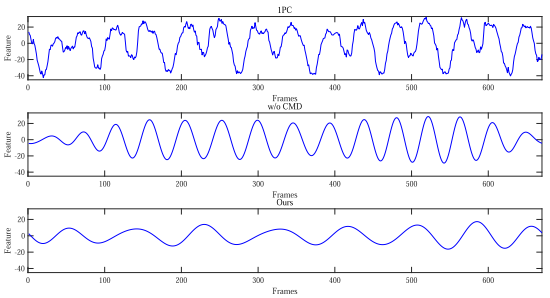


(b) QUVA

Fig. 16. Visualized results of signal trend removal. In Fig. 16(a) or Fig. 16(b), from top to bottom, it denotes the waveform of the first 1D component of our adaptive features generated in Section III-B, the reconstructed waveform of "w/o STR" and the reconstructed waveform our proposed method, respectively.



(a) YT_Segments



(b) QUVA

Fig. 17. Visualized results of coupled motion detection.

Fig. 15(b), the spatial features may be better than the temporal or fusion features for counting. The quantitative and qualitative results further show the importance of adaptive feature mode selection for accurate counting.

To evaluate the effectiveness of the signal trend removal (STR) scheme, we conduct the experiments by removing this scheme from our methodology, and the results are shown in Table V and Fig. 16. As shown in Table V, without removing the signal trend caused by the background noise and changing views of the camera, the errors on all datasets increase greatly, showing the effectiveness of our signal trend removal scheme. As shown in Fig. 16, without our proposed STR scheme, the reconstructed periodic waveforms are poor at the later time-steps of the video, and it is hard to achieve accurate counting using such waveforms. By contrast, with the proposed STR scheme, the reconstructed waveforms are better, showing the importance of removing the time-varying DC component caused by the background noise and the camera.

To evaluate the effectiveness of the coupled motion detection scheme (CMD), we remove this scheme of our methodology. In this case, we obtain the counting results using the highest energy rules as described in the Section III-D.2. The results are reported in Table VI and Fig. 17. As shown in Table VI, the errors of "w/o CMD" decrease on all datasets, especially on the more challenging datasets such as QUVA and RARV. Because of the more challenging datasets, human action is very complex, there may exist many actions with coupled motion characteristics, as shown in Fig. 1. Therefore, it easily suffers from miscounting on these datasets, leading to poor performance. As shown in Fig. 17(a) and Fig. 17(b), without removing the corresponding sub-actions, the frequencies of the reconstructed action periodic waveforms are usually double, and therefore it easily causes double-counting, showing the importance of CMD and removing the corresponding sub-actions.

V. CONCLUSION

We propose an important insight that the periodicity of the action can be well modelled by the deep features extracted from the action recognition task. We think this insight is very important for repetition counting due to two reasons. On the one hand, the repetition counting method can borrow the state-of-the-art results or experiences from action recognition, which decreases the gap between the development of action recognition and the repetition counting. On the other hand, this insight can simplify the repetition counting task from the trivial preprocessing or synthetic mode generation.

Based on this insight, we propose a new counting method using high energy rules for unconstrained videos. In detail, using the pre-trained model, we extract deep features, including the temporal evolution characteristics of video actions and

the unique appearance and spatiotemporal characteristics of motion patterns, by deep ConvNets, and then the periodic movement information is obtained by the PCA based on the deep features. Besides, we propose a novel scheme to adaptively select proper features for the videos with the different backgrounds, making it robust in the real complex scenes. Furthermore, we compute the frequency spectrum and power spectrum based on Fourier transform to remove noise information by action periodicity reconstruction scheme. Finally, the time sequence waveform is smoothed, and the action repetition counting task is completed according to peak detection. Extensive experimental results show the effectiveness of our method.

Compared with the existing methods, our method is simple and flexible without preprocessing. However, it still has poor performance when there is interference or chaotic background in the motion, especially when there are main actions with low energy or the repetitive actions associated with multiple objects simultaneously. These interferences make it hard to analyze the motion characteristics of the target object accurately. We will focus on these problems in the future.

REFERENCES

[1] Y. Jang, Y. Song, C. D. Kim, Y. Yu, Y. Kim, and G. Kim, "Video question answering with spatio-temporal reasoning," *Int. J. Comput. Vis.*, vol. 127, no. 10, pp. 1385–1412, Oct. 2019.

[2] S. M. Seitz and C. R. Dyer, "View-invariant analysis of cyclic motion," *Int. J. Comput. Vis.*, vol. 25, no. 3, pp. 231–251, 1997.

[3] O. Kumdee and P. Ritthipravat, "Repetitive motion detection for human behavior understanding from video images," in *Proc. IEEE Int. Symp. Signal Process. Inf. Technol. (ISSPIT)*, Dec. 2015, pp. 484–489.

[4] O. Kihl, D. Picard, and P.-H. Gosselin, "Local polynomial space–time descriptors for action classification," *Mach. Vis. Appl.*, vol. 27, no. 3, pp. 351–361, Apr. 2016.

[5] C. Lu and N. J. Ferrier, "Repetitive motion analysis: Segmentation and event classification," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 26, no. 2, pp. 258–263, Feb. 2004.

[6] A. B. Albu, R. Bergevin, and S. Quirion, "Generic temporal segmentation of cyclic human motion," *Pattern Recognit.*, vol. 41, no. 1, pp. 6–21, Jan. 2008.

[7] Q. Wang, G. Kurillo, F. Ofli, and R. Bajcsy, "Unsupervised temporal segmentation of repetitive human actions based on kinematic modeling and frequency analysis," in *Proc. Int. Conf. 3D Vis.*, Oct. 2015, pp. 562–570.

[8] E. Ribnick, R. Sivalingam, N. Papanikolopoulos, and K. Daniilidis, "Reconstructing and analyzing periodic human motion from stationary monocular views," *Comput. Vis. Image Understand.*, vol. 116, no. 7, pp. 815–826, Jul. 2012.

[9] B. Wandt, H. Ackermann, and B. Rosenhahn, "3D reconstruction of human motion from monocular image sequences," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 38, no. 8, pp. 1505–1516, Aug. 2016.

[10] D. Ormoneit, M. J. Black, T. Hastie, and H. Kjellström, "Representing cyclic human motion using functional analysis," *Image Vis. Comput.*, vol. 23, no. 14, pp. 1264–1276, Dec. 2005.

[11] T. F. Iversen and L.-P. Ellekilde, "Kernel density estimation based self-learning sampling strategy for motion planning of repetitive tasks," in *Proc. IEEE/RSJ Int. Conf. Intell. Robots Syst. (IROS)*, Oct. 2016, pp. 1380–1387.

[12] G. J. Burghouts and J.-M. Geusebroek, "Quasi-periodic spatiotemporal filtering," *IEEE Trans. Image Process.*, vol. 15, no. 6, pp. 1572–1582, Jun. 2006.

[13] O. Levy and L. Wolf, "Live repetition counting," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Dec. 2015, pp. 3020–3028.

[14] T. F. H. Runia, C. G. M. Snoek, and A. W. M. Smeulders, "Real-world repetition estimation by div, grad and curl," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2018, pp. 9009–9017.

[15] A. Krizhevsky, I. Sutskever, and G. Hinton, "ImageNet classification with deep convolutional neural networks," in *Proc. Adv. Neural Inf. Process. Syst. (NIPS)*, 2012, pp. 1097–1105.

[16] K. Simonyan and A. Zisserman, "Two-stream convolutional networks for action recognition in videos," in *Proc. Adv. Neural Inf. Process. Syst. (NIPS)*, 2014, pp. 568–576.

[17] E. Pogalin, A. W. M. Smeulders, and A. H. Thean, "Visual quasi-periodicity," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2008, pp. 1–8.

[18] R. Cutler and L. S. Davis, "Robust real-time periodic motion detection, analysis, and applications," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 22, no. 8, pp. 781–796, Aug. 2000.

[19] I. Laptev, S. J. Belongie, P. Perez, and J. Wills, "Periodic motion detection and segmentation via approximate sequence alignment," in *Proc. 10th IEEE Int. Conf. Comput. Vis. (ICCV)*, vol. 1, Dec. 2005, pp. 816–823.

[20] E. Ribnick and N. Papanikolopoulos, "3D reconstruction of periodic motion from a single view," *Int. J. Comput. Vis.*, vol. 90, no. 1, pp. 28–44, Oct. 2010.

[21] Y. Ren *et al.*, "An efficient framework for analyzing periodical activities in sports videos," in *Proc. 4th Int. Congr. Image Signal Process.*, Oct. 2011, pp. 502–506.

[22] G. Li, X. Han, W. Lin, and H. Wei, "Periodic motion detection with ROI-based similarity measure and extrema-based reference selection," *IEEE Trans. Consum. Electron.*, vol. 58, no. 3, pp. 947–954, Aug. 2012.

[23] S. Ioffe and C. Szegedy, "Batch normalization: Accelerating deep network training by reducing internal covariate shift," 2015, *arXiv:1502.03167*. [Online]. Available: http://arxiv.org/abs/1502.03167

[24] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," 2014, *arXiv:1409.1556*. [Online]. Available: http://arxiv.org/abs/1409.1556

[25] W. Kay *et al.*, "The kinetics human action video dataset," 2017, *arXiv:1705.06950*. [Online]. Available: http://arxiv.org/abs/1705.06950

[26] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*. Piscataway, NJ, USA: IEEE, Jun. 2016, pp. 770–778.

[27] N. Xiao, P. Yang, Y. Yan, H. Zhou, X.-Y. Li, and H. Du, "From communication to sensing : Recognizing and counting repetitive motions with wireless backscattering," 2018, *arXiv:1810.11707*. [Online]. Available: http://arxiv.org/abs/1810.11707

[28] L. Wang *et al.*, "Temporal segment networks: Towards good practices for deep action recognition," in *Proc. Eur. Conf. Comput. Vis. (ECCV)*. Cham, Switzerland: Springer, 2016, pp. 20–36.

[29] R. Polana and R. C. Nelson, "Detection and recognition of periodic, nonrigid motion," *Int. J. Comput. Vis.*, vol. 23, no. 3, pp. 261–282, 1997.

[30] B. Jiang, M. Wang, W. Gan, W. Wu, and J. Yan, "STM: SpatioTemporal and motion encoding for action recognition," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2019, pp. 2000–2009.

[31] J. Lin, C. Gan, and S. Han, "TSM: Temporal shift module for efficient video understanding," in *Proc. IEEE/CVF Int. Conf. Comput. Vis.*, Oct. 2019, pp. 7083–7093.

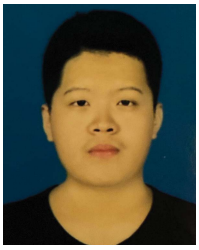
[32] Y. Li, B. Ji, X. Shi, J. Zhang, B. Kang, and L. Wang, "TEA: Temporal excitation and aggregation for action recognition," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2020, pp. 909–918.

[33] T. Runia, C. Snoek, and A. Smeulders, "Repetition estimation," *Int. J. Comput. Vis.*, vol. 127, no. 9, pp. 1361–1383, 2019.

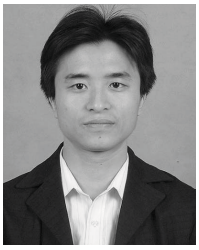
Jianqin Yin received the Ph.D. degree from Shandong University, Jinan, China, in 2013. She is currently a Professor with the School of Artificial Intelligence, Beijing University of Posts and Telecommunications, Beijing, China. Her research interests include service robot, pattern recognition, machine learning, and image processing.

Yanchun Wu received the master's degree from the School of Information Science and Engineering, University of Jinan, China. She is currently a Visiting Student with the School of Artificial Intelligence, Beijing University of Posts and Telecommunications, Beijing, China. Her main research fields include pattern recognition and machine learning.

Chaoran Zhu is currently pursuing the B.S. degree with the International School, Beijing University of Posts and Telecommunications, Beijing, China. His research interests include computer vision and machine learning.



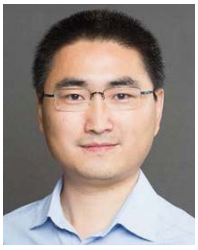
Zijin Yin is currently pursuing the B.S. degree with the International School, Beijing University of Posts and Telecommunications, Beijing, China. His research interests include computer vision and machine learning.



Huaping Liu (Senior Member, IEEE) received the Ph.D. degree from Tsinghua University, Beijing, China, in 2004. He currently is an Associate Professor with the Department of Computer Science and Technology, Tsinghua University. His research interest includes robot perception and learning.

Yonghao Dang received the bachelor's degree from the University of Jinan, Jinan, China, in 2018. He is currently pursuing the Ph.D. degree with the School of Artificial Intelligence, Beijing University of Posts and Telecommunications, Beijing, China. His research interests include computer vision and machine learning.

Zhiyi Liu is currently pursuing the degree with the Affiliated High School, Peking University. She is also a Visiting Student with the Automation School, Beijing University of Posts and Telecommunications. Her research interest includes image recognition based on computer vision.



Jun Liu (Member, IEEE) received the Ph.D. degree from the University of Toronto, Toronto, Canada, in 2016. From 2017 to 2019, he was a Post-Doctoral Fellow with the Dalio Institute of Cardiovascular Imaging, Cornell University, Ithaca, NY, USA. He is currently a Teacher with the Department of Mechanical Engineering, City University of Hong Kong, Hong Kong. His research interests include micro-/nanorobotics and medical image analysis and interaction.