EV-Action: Electromyography-Vision Multi-Modal Action Dataset

Lichen Wang[1], Bin Sun[1], Joseph Robinson[1], Taotao Jing[1], and Yun Fu[1,2]
[1] Department of Electrical and Computer Engineering, Northeastern University, USA
[2] Khoury College of Computer Science, Northeastern University, USA

Abstract— **Multi-modal human action analysis is a critical and attractive research topic. However, the majority of the existing datasets only provide visual modalities (*i.e.*, RGB, depth and skeleton). To make up this, we introduce a new, large-scale EV-Action dataset in this work, which consists of RGB, depth, electromyography (EMG), and two skeleton modalities. Compared with the conventional datasets, EV-Action dataset has two major improvements: (1) we deploy a motion capturing system to obtain high quality skeleton modality, which provides more comprehensive motion information including skeleton, trajectory, acceleration with higher accuracy, sampling frequency, and more skeleton markers. (2) we introduce an EMG modality which is usually used as an effective indicator in the biomechanics area, also it has yet to be well explored in motion related research. To the best of our knowledge, this is the first action dataset with EMG modality. The details of EV-Action dataset are clarified, meanwhile, a simple yet effective framework for EMG-based action recognition is proposed. Moreover, state-of-the-art baselines are applied to evaluate the effectiveness of all the modalities. The obtained result clearly shows the validity of EMG modality in human action analysis tasks. We hope this dataset can make significant contributions to human motion analysis, computer vision, machine learning, biomechanics, and other interdisciplinary fields.**

I. INTRODUCTION

There are a wide range of applications for human motion analysis (*e.g.*, event detection, behavior prediction, gait analysis, joint mechanics, prosthetic designs, sports medicines [36], [18], [50], [41], [48], [49], [47], [51]). The availability of datasets tends to directly impact the progress of research. From the start, action datasets only consisted RGB modality [36]. Later on, as 3D sensors became more accessible, several datasets included the depth modality [58], [24], [27], [23]. This paved a way for researchers to propose more effective approaches in terms of multi-modal methods [18], [50], [41]. After that, skeleton data was introduced by some works [42], [55]. However, most skeleton information of these datasets was directly obtained from Kinect sensors [30], resulting in low localization accuracy. Skeleton modal captured by more accurate devices was released [9], while RGB-D modals were not included.

We introduce EV-Action dataset, which includes all visual modalities mentioned above (*i.e.*, RGB, depth, and two skeleton modalities). An optical tracking-based Vicon system [31] is deployed to capture high-quality skeleton motion information. Compared with Kinect, Vicon achieves significantly higher sampling rate (100 vs. 30 fps), higher localization accuracy, and more skeleton markers (39 vs. 26). It provides more comprehensive skeleton motion information in terms of location, trajectory, velocity, and acceleration. We further collected Electromyography (EMG) signals to measure the electrical activity of human skeletal muscles as a function of the intensity of force [10]. EMG is regularly used in medical and biomechanics fields. It has not yet been well explored in the fields of human motion analysis. In EV-Action, all modalities are captured simultaneously with action labels frame by frame. The goal of EV-Action is exploring the latent correlation across different modalities and improving the performances of action analytic tasks. EV-Action could contribute significantly to the research fields of human motion analysis, multimedia, computer vision, machine learning, biomechanics, and other interdisciplinary sub-fields. The contributions of our paper are shown below:

1) We designed and constructed a data collection center with optical tracking system and Kinect-V2 systems. This allowed us to capture the four visual modalities (*i.e.*, RGB, Depth, Skeleton-K, and Skeleton-V).
2) EMG signal from skeletal muscles is extracted. This is the first action dataset including EMG, which provides complimentary information and reveals valuable correlations between visual and non-visual modalities.
3) A simple yet effective EMG recognition framework is proposed which achieves highest performance and reveals unique characteristics of EMG in human actions.
4) We defined experimental settings and provided the state-of-the-art benchmarks for each modality. EMG is merged with other modalities which further demonstrates the complementary of the EMG modal.

II. RELATED WORKS

A. RGB/D and Skeleton Datasets

Small-scale datasets included tens of action classes (*e.g.*, Weizmann [17]) are initially deployed for action analytic tasks [58]. Upon the arrival of deep learning, large-scale RGB datasets were introduced (*e.g.*, UCF101 [40] and Kinetics [22]). Later on, RGB-D datasets were released (*e.g.*, MSR-Action3D [24], RGBD-HuDaAct [27]). Due to the space and budgeting constraints, most RGB-D datasets were collected using low-cost Kinect sensors [30], [38]. In addition, Kinect sensors can extract skeleton data, as introduced in MAD [20], UCF-Kinect [14] and NTU-RGBD [37]. However, the accuracy and stability of Kinect are low, which limits the potential research of action analysis.

We thank our volunteers, Allyson Vakhovskaya, Daniel J. Peluso, Emily Freed, Kasey Lee, Yue Bai, and Yunyu Liu from Northeastern University for their substantial contributions to our project in data collection, labeling, and analytical procedures.

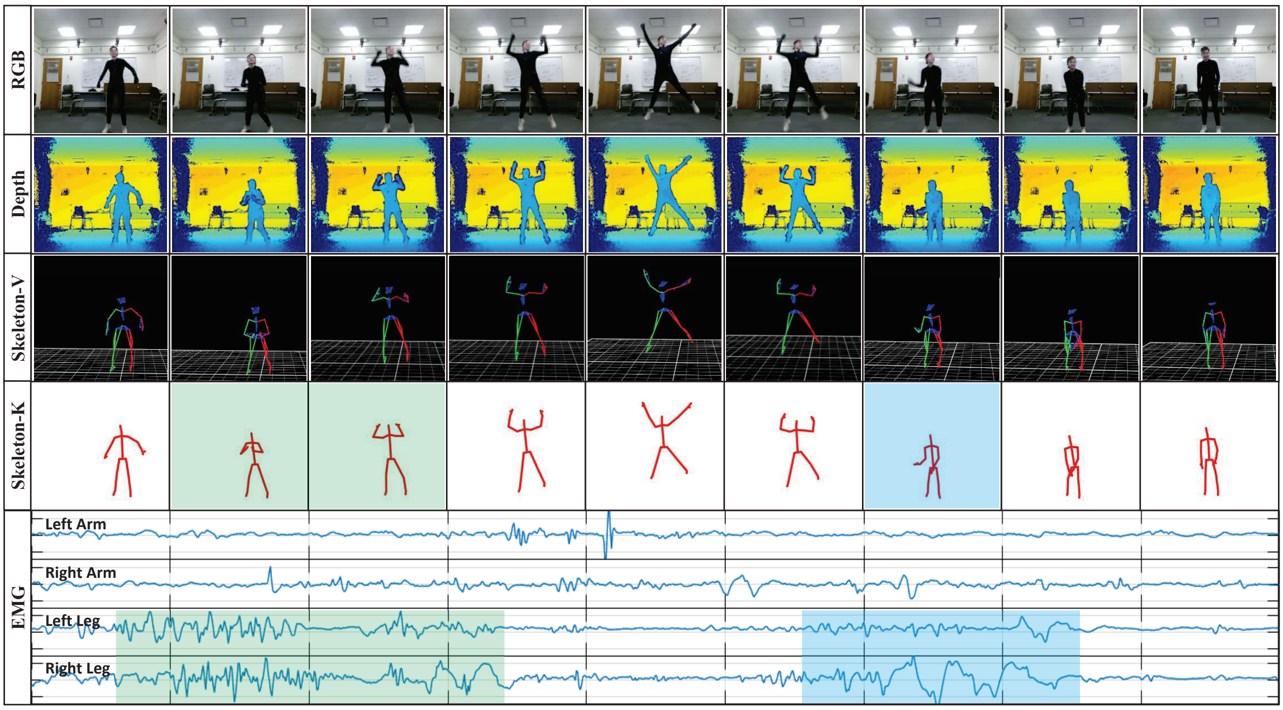


Fig. 1: Visualization of sample frames in EV-Action dataset. Colored boxes show the correlations between visual modalities and EMG (*i.e.*, *Take Off* and *Touch Down*). We can clearly observe that EMG responds early and last longer than visual modalities which provides unique view for action analysis. All modalities were well aligned and labeled.

TABLE I: Technical specifications of the sensors used in EV-Action dataset.

Sensors	Modality	Resolution	Frame Rate (fps/Hz)	Skeleton Joints	Field of View	Sensor Number	Range	Sensitivity
	RGB	$1,920 \times 1,080$	30	-	$84.1^\circ \times 53.8^\circ$	1	-	8-bit
Kinect-V2	Depth	512×424	30	-	$70.6^\circ \times 60.0^\circ$	1	0.5-4.5 m	16-bit
	Skeleton-K	-	30	26	-	-	-	-
Vicon-T40s	Skeleton-V	$2,336 \times 1,728$	100	39	$98.1^\circ \times 50.1^\circ$	8	12 m	10-bit
Delsys-Trigno	EMG	-	1,000	-	-	4	± 22 mV	16-bit

B. Multi-Modal Action Datasets

We consider the dataset containing more than RGB-D modalities as multi-modal dataset. Currently, only a few datasets provide additional modalities. NTU-RGBD [37] and PKU-MMD [8] contained infrared frames captured by Kinect sensors. CMU-MMAC [9] utilized an optical tracking technique to capture action sequences. UTD-MHAD [5] utilized a single wearable inertial sensor to capture inertial signals. However, the modality is severely limited and sporadic due to the inconsistent collection manner. Our EV-Action dataset utilizes 39 markers to capture precise location, trajectory and acceleration information at a high frame rate (100 fps). To this end, EV-Action is the **most accurate and comprehensive** dataset of this kind.

C. EMG Signal

Electromyography (EMG) is an electrodiagnostic technique to evaluate the electrical activity produced by skeletal muscles [33], [10]. Typically, EMG is used in neural science, biomechanics, and signal processing fields (*e.g.*, hand gesture [6], robot arm control [15], face expression [34]. Since EMG activates before visual signal which could foresee potential information such as intention, force, and even mental activities information that cannot be recognized in visual domain. To this end, we consider EMG as another crucial clue for exploring actions. To the best of our knowledge, no existing work associates EMG with other action modalities. Considering the potential applications that could be had, we generated EV-Action with EMG as one of the critical modalities for human action analysis.

III. EV-ACTION DATASET

A. Sensors and Setup

There are 1 Kinect-V2 sensor [1], 4 wireless EMG sensors, and 8 Vicon-T40s cameras in the data collection system.

Kinect [1] captures RGB-D modalities from subjects. Skeleton information is further extracted from the depth image. We used a second generation Kinect [1], [30] (Kinect-V2) which has a high resolution camera (1,920×1,080) at 30 fps with a wide field of view (70°×60°). Moreover, the resolution of the depth sensor is 512×424. It is more robust and efficient for pose estimation with reference to 26 joints (Figure 3(a)). In the collection procedure, a Kinect-V2 captures the subjects in the front view (Figure 3(b)).

Vicon System utilizes optical tracking-based technology to capture skeleton data with more accurate and comprehensive motion information [31]. We deploy 8 Vicon-T40s infrared cameras to capture the stickup marks on each subject (Figure 3(a)). The cameras sample data points as 10-bit grayscale frames at 100 fps and with a resolution of 2336×1728.

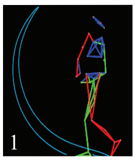
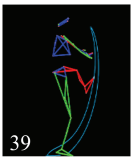
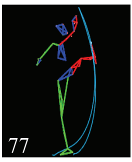
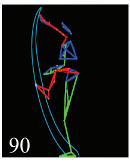
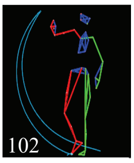
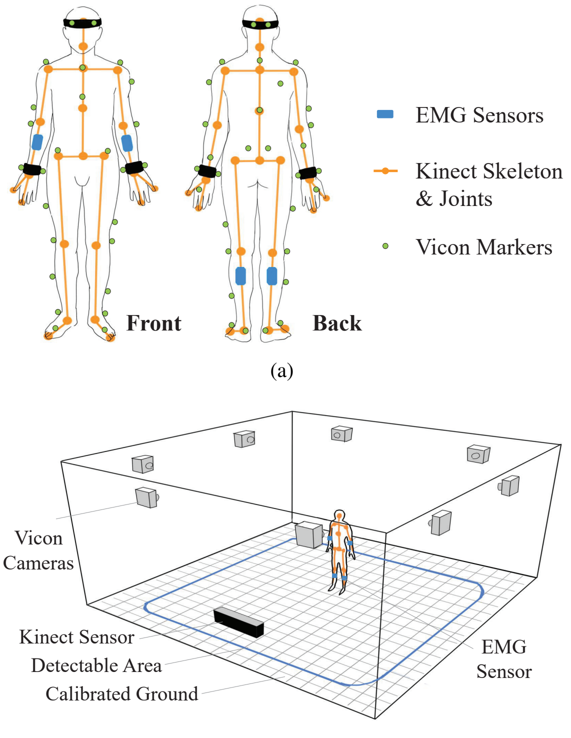
Vicon SK	 1	 14	 25	 39	 52	 64	 77	 90	 102
Angle	0°	45°	90°	135°	180°	225°	270°	315°	360°
Time(s)	0.00	0.13	0.25	0.40	0.53	0.65	0.78	0.91	1.04

Fig. 2: Visualization of a subject performing a kicking action across view angles and time. The blue curve highlights the trajectory of a marker. Clearly, EV-Action contains the precise detailed motion information of the actions. Frame numbers are shown in left bottom which indicate the high sampling rate of Vicon system.

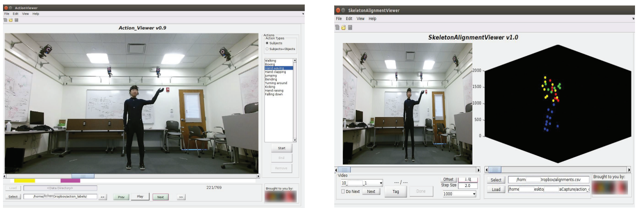


(a)

(b)

Fig. 3: (a) Sensor placement schemes. Orange lines and spots indicate Kinect skeleton with 26 joints. Small gray points denote Vicon markers. And blue blocks indicate the EMG sensors. (b) Data collection center environment setup.

Then, the frames were calibrated and labeled to obtain skeleton information. We follow the standard scheme [43] by placing 39 markers around human body (Figure 3(a)). It captures precise and comprehensive motion information, such as the second bounce in the *Fall Down* action class. Also, due to the high frame rate and accuracy, high quality trajectories and accelerations were obtainable in reference to ground coordinates. Figure 2 shows the *Kick* action viewed across time and at different angles, with the blue curve indicating the trajectory of the toe marker. No other action datasets provides such detailed information.



(a)

(b)

Fig. 4: (a) Video labeling tool used to label action sequences, and (b) tool used to precisely align modalities.

EMG Sensor captures EMG signals from human muscles. We deploy wireless EMG sensors which captures 16-bit EMG signal at 1000 Hz. This enables the sensors to cover the whole frequency spectrum of skeletal EMG (*i.e.*, 20-450 Hz) signal. We attached 4 sensors to each subject: the middle of each forearm and the shank muscles (Figure 3(a)). There are 3 reasons: 1), the common actions usually utilize arms and legs; 2), the location of each muscle (mid-line of the muscle in the belly that is between the myotendinous junction and the nearest innervation zone) gives off a signal of highest amplitude, which makes the signal most responsive to the corresponding action [10]; 3), the *crosstalk* noise generated by neighboring muscles has the potential to get misinterpreted for originating from a muscle of interest, and placing the sensor mid-line makes it less susceptible to this noise.

Data Collection Center consists of 8 the Vicon cameras placed around the parameter of a 4.6m × 4.6m room which has a detectable area of 3m × 3m. All traceable markers fell in the Vicon cameras field of view. There was a single Kinect sensor centered facing front, and each action was performed with the face of this Kinect sensor as the front. 4 EMG sensors were connected to each subject (Figure 3(b)).

B. *Dataset Description*

Completeness, comprehensiveness, and diversity were highly considered when building EV-Action. To make it practical and generalizable, we included 20 common actions (Table III), 10 were done by a single subject and the other 10 were done by that same subject interacting with different objects. The dataset includes 70 subjects performing the actions 5 times (*i.e.*, 100 action clips per subject). To introduce diversities, the subjects intentionally perform slightly different style in each loop. All-in-all, resulting in 7000 action clips at multiple views. Table II summarizes these statistics compared with recent and popular multi-modal action datasets. It is clear that EV-Action is one of the largest multi-modal datasets, as it significantly surpasses other datasets in terms of modal diversity, number of subjects, and number of samples. And it includes non-visual EMG signal for the first

TABLE II: Comparison between EV-Action dataset and other popular multi-modal datasets. EV-Action is one of the largest multi-modal datasets and significantly outperforms other datasets in modal diversity, subject numbers, and sample clips.

Datasets	Samples	Classes	Subjects	Framerate (fps)	Sensors	Modalities
RGBD-HUDA [27]	1189	13	30	30	KinectV1	RGB+D
MSR-Action3D [24]	567	20	10	30	RGB-Cam	D+SK
CAD-60 [41]	60	12	4	30	KinectV1	RGB+D+SK
Action4^2 [7]	6844	14	24	30	KinectV1	RGB+D
CAD-120 [23]	120	20	4	30	KinectV1	RGB+D+SK
Multiview 3D Event [54]	3815	8	8	30	KinectV1	RGB+D+SK
Online RGB+D Action [57]	336	7	24	30	KinectV1	RGB+D+SK
Northwestern-UCLA [45]	1475	10	10	30	KinectV1	RGB+D+SK
UWA3D Multiview [32]	900	30	10	30	KinectV1	RGB+D+SK
Office Activity [46]	1180	20	10	30	KinectV1	RGB+D
UTD-MHAD [5]	861	27	8	30+50	KinectV1+WIS	RGB+D+SK
3D Action Pairs [29]	360	12	10	30	KinectV1	RGB+D+SK
UWA3D Multiview II [28]	1075	**30**	10	30	KinectV1	RGB+D+SK
EV-Action (Ours)	**7000**	20	**70**	**30+100+1000**	**KinectV2+Vicon+EMG**	**RGB+D+SKK+SKV+EMG**

TABLE III: A list of the 20 actions included in EV-Action.

Single Person Actions		Person-Objects Actions	
1. Walk	6. Bend Over	1. Answer Phone	6. Throw Ball
2. Boxing	7. Turn Around	2. Check Watch	7. Drink Water
3. Wave Hands	8. Kick	3. Stand Up	8. Tie Shoes
4. Clap Hands	9. Raise Hand	4. Sit Down	9. Read Book
5. Jump	10. Fall Down	5. Grab Bag	10. Move Table

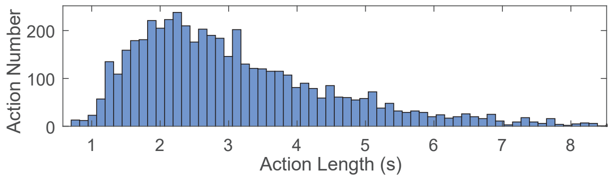


Fig. 5: Histogram of the length distribution of all videos.

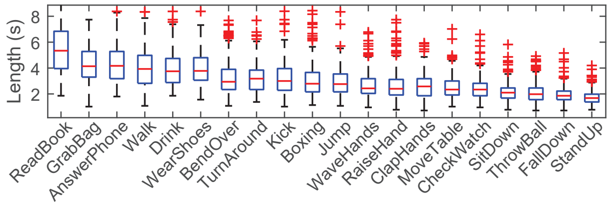


Fig. 6: Video length distribution of each action.

time. Referencing figure 1, we notice that the EMG signal activates prior to the actions (*i.e.*, *Take Off* and *Touch Down*). We also notice that the duration of EMG are typically longer than the visual modalities. These patterns are unrecognizable from any visual modal. It demonstrates that EMG does provide unique and complementary information for more deep and sophisticated action analytical research.

C. Data Labeling

There are two steps for data labeling: 1) annotating the actions in RGB modality, and 2) aligning the Skeleton-K and Skeleton-V. Since other modalities are captured synchronously with either Kinect or Vicon, thus, the rest of the modalities are also well aligned automatically. **Video Anotation:** We built a MATLAB labeling tool to facilitate the labeling process (Figure 4(a)). The tool displays a video sample for a human labeler to tag start and end frames of actions selected from the predefined lists shown in Table III. **Data Alignment:** We then align the clip with the clip captured by the Vicon system. We develop another MATLAB tool that allows us to visually align single frame of the RGB and the skeleton from the Kinect and the Vicon systems, respectively (Figure 4(b)).

IV. Data Analysis

Histograms of all video length are shown in Figure 5, and box plots (Figure 6) depict action-specific statistics from longest to shortest. We observe that the video lengths for different actions varied. For instance, *Read Book* tends to be the longest, with *Stand Up* the shortest. Moreover, variation in video length exists for the same action across different subjects, which is especially true for repetitive actions. For instance, when preforming actions such as *Boxing* or *Jump*, subjects prefer to choose the exact number of reps. For non-repetitive actions there tended to be less variation across different subjects. Since subjects perform different actions continuously without intentional pause during collection; another observation is that subjects tended to move faster between actions. This sometimes results in the end of the current action and the beginning of the next getting mixed across frames in between (*i.e.*, overlap between actions). For example, when a subject *Put Down Phone* and then *Check Watch* immediately, there might be an overlap. These situations make more of challenges while make it better suited for more practical research.

Root Mean Squared (RMS) is an effective method to preprocess EMG data [10]. We obtain the average of action RMS and surprisingly notice that the shank muscles have significantly higher (2 times) amplitude than forearm muscles since the stronger and bigger muscles around the shank region. We separately illustrated these four channels in Figure 7 and found more interesting observations. For instances, most subjects utilize right hand for *Throwing Ball*, while they are also utilizing their right legs simultaneously (might for balance requirement). Moreover, subjects use left legs even for *Check Watch* (might for body balance; in order to rise left arm to check watch, they should hold/balance their body by left leg to take over left arm). These observations are unique and valuable for action understanding but cannot be obtained in any visual modality. We wish more interesting discoveries could be revealed by exploring EMG modal.

Since Vicon tracks the markers pasted on subjects, there were situations that several markers were obstructed to the cameras, such as *Fall Down* and *Sit Down*. Once the occluded marker is again detected, Vicon could re-localized the re-

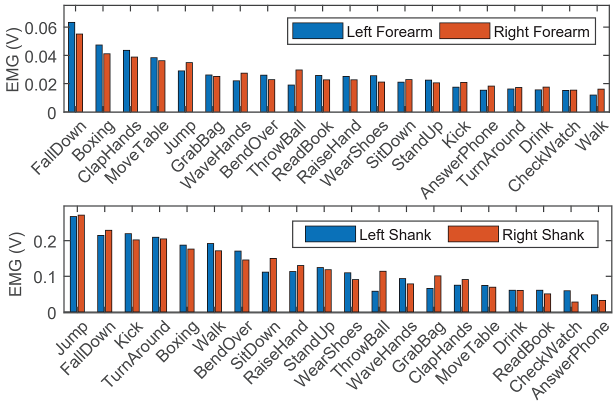


Fig. 7: The average of Root Mean Square (RMS) value of the EMG recordings in different actions. We separate the value of upper body (left and right forearm) and lower body (left and right shank) for better discussion.

spective point. In response for the missing situation, we split the data as two types, unlabeled marker locations data and labeled skeleton data. Thus, advanced skeleton reconstruction methods or label independent research can also explore EV-Action. This also leads to believe that more sophisticated algorithms are needed to achieve higher performance rating (*e.g.*, missing-modality and multi-view algorithms). The rest modalities (*i.e.*, RGB, Depth, Skeleton-K and EMG) are stable across all actions without noticeable errors.

V. EXPERIMENTS

State-of-the-art approaches were used to benchmark the different modalities. Specifically, single-modal benchmarks using RGB, Skeleton-K, Skeleton-V were done. In the multi-modal scenario, RGB-D, Skeleton-K + EMG, and Skeleton-V + EMG were conducted. We achieved considerable performance improvements by employing a simple, yet effective fusion technique (*i.e.*, fused at the feature-level). This project is the first to model the non-visual EMG signal for action recognition. This is also a great promise for further improvement by providing more sophisticated learning frameworks and fusion techniques. Considering the information captured in an EMG signal, it is capable of discriminating between action types in itself, thus, it is complimentary to visual evidence. Thus, the EMG modality could both improve our current action recognition capabilities and serve as a necessity for certain applications.

A. Experimental Settings

Benchmarks on EV-Action followed conventional classification settings. The action clips from 56 subjects were used during training (*i.e.*, 5600 clips), while the other 14 subjects were set aside for testing (*i.e.*, 1400 clips). All experiments were evaluated in terms of classification accuracy.

B. EMG Signal

Signal processing methods associated with hand-crafted features are usually deployed for EMG analysis. We design a novel deep-structure framework for EMG recognition.

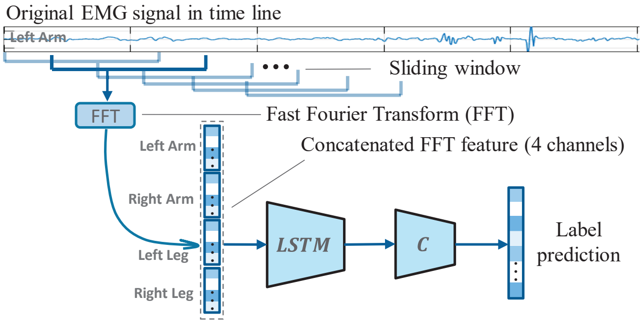

Original EMG signal in time line
Left Arm
Sliding window
FFT
Fast Fourier Transform (FFT)
Concatenated FFT feature (4 channels)
Left Arm
Right Arm
Left Leg
Right Leg
LSTM
C
Label prediction

Fig. 8: EMG modal based action recognition framework.

We first introduce the conventional procedures for EMG classification. As noise of raw EMG signals occurred during collecting period, choosing the best way to extract features and reduce the dimension are crucial to achieve high classification performance. Butterworth filter [4] yields a flat frequency response which is effective to filter out EMG noise. The generalized equation is: $H_{j\omega} = (\sqrt{1+\varepsilon^2(\frac{\omega}{\omega_p})^{2n}})^{-1}$, where n is the filter order and we received the best classification performance when $n = 5$. ω is the radian frequency and $\omega = 2\pi f$. ω_p is the pass band frequency and ε is the maximum pass band. H is the frequency response. The low frequency cutoff value is set to 10 Hz, which removes the static electricity variation caused by friction and movement. We set high frequency cutoff value at 500 Hz, since it is the highest frequency of the EMG signal. RMS is further deployed. The expression of RMS is $R_k = (\frac{1}{N}\sum_{i=1}^{N} x_i^2)^{\frac{1}{2}}$, where R_k is the RMS value and x_i is the EMG signal of the i-th frame in the k-th time window period. N is the size of sliding window. We obtain the RMS values from each channel to obtain RMS features. Linear Discriminative Analysis (LDA) [26] and Principal Component Analysis (PCA) [21] are utilized for dimension reduction. Three classifiers are tested including SVM [35], K-Nearest Neighbors (KNN) [13], and Random Forests (RF) [25]. The results (Table IV) indicates that RF after PCA has the best classification performance which is 35.12%.

We then introduce our deep modal-based EMG action recognition approach (Figure 8). It is an effective and efficient framework in an end-to-end scenario, while both the noise elimination and the recognition are done simultaneously. Sliding windows are first employed to extract EMG signal from each channel. Differently, we utilize Fast Fourier Transform (FFT) [3] instead of RMS as the initial approach. This strategy has two advantages. Firstly, FFT decomposes the time series EMG data in frequency domain which automatically separates EMG with high/low noise. Thus, denoising procedures can be omitted. Secondly, FFT preserves more comprehensive information. Then, we utilize the amplitude of each frequency as feature vector, and concatenate the four channels (*i.e.*, left/right forearm/shank) together and input them into a Long Short-Term Memory (LSTM) [19] networks. LSTM outputs the feature representations. A classifier, $C(\cdot)$, is utilized to obtain the final label.

TABLE IV: EMG classification accuracy based on different dimension reduction (Dim-Red) approaches and classifiers.

Methods	Dimension Reduction		
	(None)	LDA	PCA
Random Forest	33.72	16.81	35.12
KNN	22.16	13.55	26.18
SVM	23.74	16.12	25.65
FFT-LSTM (Ours)	**44.13**	-	-

The loss function is shown as below.

$$L = \|Y - C(G(F(X))\|_F^2, \tag{1}$$

where Y is the instance label, X is the original data in time domain. $F(\cdot)$ is the FFT feature extractor. $G(\cdot)$ is an LSTM network. In the implementation, we deploy half-overlapping sliding window associated with the window size 200 to extract data blocks. By deploying FFT to the extract block, we obtain a 100-dimension feature vector in frequency domain. Since there are four EMG channels, each data block is represented by a 400-dimension vector. The LSTM structure has a hidden layer of 1024-dimension. The result (Table IV) denotes that the our approach significantly outperforms conventional methods which also indicates the effectiveness of EMG in action analytical tasks.

C. RGB & Depth

We evaluated single-view action recognition baselines on EV-Action. Details are introduced below:
Action Vector of Local Aggregated Descriptor (Action-VLAD) [16] is an effective video descriptor. A trainable aggregation framework is designed to capture spatio-temporal features. We fine-tune the last layer before SoftMax pre-trained by ImageNet [11] our evaluation. **Temporal Segment Networks (TSN)** [52] sparsely samples the videos to capture the temporal information in supervised scenario. In this way, the entire video was learned effectively. **Long-term Recurrent Convolutional Networks (LRCN)** [12] deploys a hierarchical visual representation learning associated with a temporal dynamic recognition module. LRCN is capable of end-to-end training. **Weighted Depth Motion Maps (WDMM)** [2] recognizes actions from depth videos. It utilizes a video summarization step for hierarchical representation learning. WDMM effectively increases inter-class dissimilarities and intra-class similarities. 110 is set as the number of PCA components and 80 is set as the visual words for extracting depth feature. **Weighted Hierarchical Depth Motion Maps (WHDMM)** [53] utilizes a convolutional neural network to extract three-channel features. A hierarchical depth motion extractor is further deployed for action recognition. We only use the front view to train and test. The remaining parameters followed the original work.

D. Skeleton-Kinect

We introduce the action recognition baselines based on skeleton modal in this section.
Temporal Convolutional Networks (TCN) [39] learns an interpretable spatio-temporal representation. To train the model on the Kinect Skeleton modality, we modify the data as the same format of NTU-RGB-D [37]. **Two Stream Recurrent Neural Network (TSRNN)** [44] provides an RNN framework with two-stream structure to explore spatial configurations and temporal dynamics for action classification. We process the data in the same way as TCN. The batch size is set to 256, the maximum iteration number is set to 2,000, and the learning rate is set to 0.02. **Spatial Temporal Graph Convolution Network (STGCN)** [56] learns both the spatial and temporal patterns from data simultaneously. It overcomes the limited expressive power and difficulties of generalization. The data is processed in the same way as TCN. We train the model with 80 epochs, using SGD as the optimizer.

E. Skeleton-Vicon

Vicon system captures skeleton data with higher localization quality. In our evaluation, we deploy the same baselines as (Skeleton-K) while modify the data format to satisfy the requirements of each baselines. For **TCN** [39] approach, we change the spatial connection graph from 25 joints to 39 joints. Vicon data contains higher frame rate, and we also increase frames for other models. The remaining parameters are kept the same. The dimension of the feature is increased to 273. **TSRNN** [44] needs the part of the body (*i.e.*, one trunk, two legs, and two arms, as well as whole body). We use the index groups of different body parts via the 39 joints of Vicon and keep other parameters be consistent. **STGCN** [56] needs a joint adjacency graph. Thus, we generate the connection graph for 39-joint while other parameters are consistent.

F. Skeleton-(Kinect/Vicon) & EMG

To prove the effectiveness and complementariness of EMG modal, we combine EMG with skeleton modalities together in low-level domain.
TCN-RMS. We first obtain the EMG features. The time window has the same size as the sampling time between two frames of the Skeleton-K. We then concatenate the RMS directly with the skeleton data and input the combination data to TCN [39] for classification. The hyper parameters we used are the same as the parameters aforementioned. And we find out that such combination features have improved the performance on the action recognition task. **TCN-FFT.** FFT-based feature merging strategy is also evaluated since EMG is a temporal signal which can be explored in frequency space. Similar as TCN-RMS, we set a time window to extract the frequency distribution feature of the signal in each channel, and forward to classifiers.

G. Results and Analysis

We used top-1 accuracy to evaluate each baseline (Table V). The left column shows the modals utilized for evaluation. *SK-K* and *SK-V* indicate the single-view skeleton data Skeleton-K and Skeleton-V respectively. *SK-K-E* and *SK-V-E* denote the multi-modal setting where EMG modal is combined with Skeleton-K and Skeleton-V respectively. For RGB modality, TSN outperformed other benchmarks with

TABLE V: Action recognition accuracy scores (%) for all benchmarks.

		Single-Person										Person-Object										
		Walk	*Box*	*Wave Hand*	*Clap Hands*	*Jump*	*Bend*	*Turn Around*	*Kick*	*Raise Hand*	*Fall Down*	*Ans. Phone*	*Check Watch*	*Stand Up*	*Sit Down*	*Grab Bag*	*Throw Ball*	*Drink Water*	*Tie Shoes*	*Read Book*	*Move Table*	ACC
RGB	TSN [52]	**56.1**	**94.1**	**25.3**	**83.9**	**88.5**	**94.3**	**68.3**	**95.6**	95.1	**86.2**	**69.5**	**37.6**	**87.0**	**54.3**	**86.9**	**75.7**	**56.8**	**84.8**	**96.7**	**59.1**	**74.7**
	LRCN [12]	44.2	84.0	19.8	69.4	71.6	78.0	57.9	82.1	90.0	71.3	55.6	28.5	72.1	43.4	72.0	62.5	46.8	70.2	85.4	44.2	62.4
	VLAD [16]	47.5	91.8	21.6	75.9	78.3	85.3	63.3	89.7	**98.4**	77.9	60.7	31.1	78.8	47.5	78.7	68.3	50.8	76.7	93.4	48.3	68.2
Dep	WDMM [2]	44.3	76.3	11.4	31.4	36.5	**43.7**	**17.2**	47.4	**72.7**	**36.2**	**27.9**	**12.3**	45.1	**16.8**	**27.2**	**48.2**	**23.4**	**28.4**	42.1	13.5	35.1
	WHDMM [53]	**78.5**	**84.5**	**62.7**	**64.7**	**66.1**	12.3	17.2	**72.3**	67.9	20.1	12.5	11.7	**61.1**	10.1	16.7	22.5	17.0	11.2	**71.5**	**23.5**	**40.2**
SK-K	TCN[39]	**91.2**	82.0	**71.4**	**86.0**	**92.2**	**91.7**	**87.6**	**93.0**	**89.2**	**92.6**	**57.5**	**76.0**	**92.9**	**87.8**	**66.8**	**70.5**	**95.0**	**76.1**	**76.1**	76.4	**82.6**
	TSRNN [44]	90.0	**85.0**	70.6	81.0	91.0	90.5	86.6	91.8	86.6	91.4	56.7	75.1	91.7	86.8	66.0	69.7	93.8	75.1	65.1	**85.4**	81.5
	STGCN [56]	90.6	83.5	71.0	83.5	91.6	91.1	87.1	92.4	88.7	92.0	57.1	75.6	92.3	87.3	66.4	70.1	94.4	75.6	75.6	75.9	82.1
SK-V	TCN [39]	82.1	**77.2**	**67.2**	**87.2**	**83.8**	83.3	80.1	84.4	**81.4**	84.0	36.0	50.9	**64.3**	60.3	43.4	46.4	66.0	50.9	50.9	51.1	64.1
	TSRNN [44]	**83.0**	77.2	67.1	77.4	82.1	**84.4**	**80.5**	**84.9**	79.9	**84.1**	**38.4**	**64.1**	58.3	**64.0**	**46.3**	**49.4**	**70.1**	**54.1**	**64.1**	**64.3**	**67.5**
	STGCN[56]	57.7	53.2	45.2	53.2	58.4	58.0	55.5	58.9	56.5	59.6	36.4	48.2	58.7	55.6	42.3	44.6	60.1	45.2	25.2	54.3	50.7
EMG	LSTM-FFT	72.3	51.6	35.1	54.8	90.6	40.0	30.3	36.6	11.9	72.8	51.2	56.5	16.1	41.6	17.3	48.4	45.7	31.4	46.2	33.0	44.1
SK-K-E	TCN-RMS	91.1	83.0	**73.4**	**88.0**	93.2	**94.7**	**87.8**	91.0	**91.4**	**95.6**	60.5	**79.8**	**91.9**	88.8	**70.8**	**72.5**	94.0	74.1	**78.1**	74.4	83.6
	TCN-FFT	**92.0**	**83.7**	72.1	85.7	**94.0**	93.5	87.3	**94.8**	91.0	94.4	**60.6**	78.5	91.3	**89.6**	70.1	71.9	**94.8**	**79.5**	77.6	**77.9**	**84.0**
SK-V-E	TCN-RMS	**86.7**	**80.7**	**70.3**	**87.9**	**87.1**	**84.5**	**83.6**	**85.1**	**82.1**	83.6	**63.5**	**51.6**	64.4	**60.3**	**45.4**	46.0	**65.8**	50.5	51.2	51.1	**69.1**
	TCN-FFT	82.2	77.5	67.3	87.3	83.8	83.4	80.5	84.7	81.7	**84.5**	37.0	51.4	**64.5**	60.0	43.5	**47.4**	64.0	**53.9**	**52.9**	**51.1**	66.8

an accuracy of 74.8%. For each action, we noticed that some actions were easy to recognize and received over 90% accuracy (*i.e.*, *Box*, *Raise Hand*, and *Move Table*). However, *Wave Hand* only got 19.8% accuracy. This is because this kind of action has low visual distinctiveness especially when the subjects wear black suit in data collection procedure. For depth modality, WDMM obtained 35.1% average accuracy, while WHDMM greatly outdid that with 40.2%. Compared with the 82.6% accuracy obtained from the skeleton data alone, the added EMG signal improved this by 1.4% (*i.e.*, TCN-FFT with 84.0%). Thus, we conclude that the actions with slightly movements (*i.e.*, *Check Watch*, *Answer Phone*) have been improved with EMG features. Since the EMG signal can react to the action without obvious motion, while the visual feature is indistinctive. As a consequence, the result is reasonable. The reason why EMG with FFT does not have significant improvement may be that the FFT features are complicated and significantly different compared with skeleton motion structure. If these two modalities are simply and directly concatenated together, the extra dynamic information of FFT features could not be fully utilized. The EMG with LSTM-FFT also proves EMG is useful. The accuracies of several actions are extremely high (*i.e.*, *Jump*). However, similar actions(*i.e.*, *Wave Hand* and *Raise Hand*; *Stand Up* and *Sit down*), which can be easily classified by other modality, are sometimes hard for EMG. Therefore, a good fusion technique may help us get better results. We conjecture there are two reasons for the relatively low baseline performances of Vicon data. (1) Some missing points make the data more challenging. (2) The generation strategy of spatial graph for Vicon data is different from the Kinect skeleton model default setting. Comparing our skeleton dataset with NTU-RGBD [37] dataset, which included the same skeleton baselines as ours, we consider the differences in scores are justifiable in two-fold. (1) Our dataset contains less but more challenging action clips. (2) We only utilized 3D reference frame from the skeleton modality, while [37] fully utilized more motion information such as orientation. Regardless, the evaluation of our dataset can be further boosted with the added non-visual modality, the EMG signal. We believe more advanced feature extraction methods and multi-modality fusion strategies could further improve the learning performance. To this end, a lot of open questions and challenges are left for future exploration.

VI. Conclusion

We have introduced a new multi-modal human action dataset in this paper which is called EV-Action dataset. The proposed dataset consists of RGB, depth, skeleton, and EMG data. All modalities have been labeled and aligned across 7,000 samples collected from 70 human subjects. In general, EV-Action has two major advantages over the other action-based video collections. (1) we have utilized an optical tracking based Vicon system to capture more accurate and comprehensive skeletal data; (2) we have introduced a non-visual EMG modality associated with other visual modalities. We also have provided several state-of-the-art benchmarks for each modality to prove the effectiveness. Moreover, we have designed effective and simple approach for EMG-based action recognition task and achieved highest performance. Further, the experiments have demonstrated that the effective and complimentary information is extractable from EMG for human action analytical tasks. Overall, EV-Action can serve in widespread research and applications concerning human motion understanding. We hope EV-Action can have a significant impact on motion understanding, computer vision, biomechanics, and other interdisciplinary areas.

References

[1] C. Amon, F. Fuhrmann, and F. Graf. Evaluation of the spatial resolution accuracy of the face tracking system for kinect for windows v1 and v2. In *Proc. AAAA*, pages 16–17, 2014.

[2] R. Azad, M. Asadi-Aghbolaghi, S. Kasaei, and S. Escalera. Dynamic 3D hand gesture recognition by learning weighted depth motion maps. *IEEE TCSVT*, 29(6):1729–1740, 2018.

[3] E. O. Brigham and E. O. Brigham. *The fast Fourier transform and its applications*, volume 1. prentice Hall Englewood Cliffs, NJ, 1988.

[4] S. Butterworth. On the theory of filter amplifiers. *Wireless Engineer*, 7(6):536–541, 1930.

[5] C. Chen, R. Jafari, and N. Kehtarnavaz. NTD-MHAD: A multimodal dataset for human action recognition utilizing a depth camera and a wearable inertial sensor. In *Proc. ICIP*, pages 168–172, 2015.

[6] X. Chen, X. Zhang, Z.-Y. Zhao, J.-H. Yang, V. Lantz, and K.-Q. Wang. Hand gesture recognition research based on surface EMG sensors and 2D-accelerometers. In *Proc. IEEE ISWC*, pages 11–14, 2007.

[7] Z. Cheng, L. Qin, Y. Ye, Q. Huang, and Q. Tian. Human daily action analysis with multi-view and color-depth data. In *Proc. ECCV*, pages 52–61. Springer, 2012.

[8] L. Chunhui, H. Yueyu, L. Yanghao, S. Sijie, and L. Jiaying. PKU-MMD: A large scale benchmark for continuous multi-modal human action understanding. *arXiv:1703.07475*, 2017.

[9] F. De la Torre, J. Hodgins, A. Bargteil, and others. Guide to the carnegie mellon university multimodal activity (CMU-MMAC) database. *Robotics Institute*, page 135, 2008.

[10] C. J. De Luca. The use of surface electromyography in biomechanics. *Journal of applied biomechanics*, 13(2):135–163, 1997.

[11] J. Deng, W. Dong, R. Socher, L. Li, K. Li, and L. Feifei. ImageNet: A large-scale hierarchical image database. In *Proc. IEEE CVPR*, pages 248–255, 2009.

[12] J. Donahue, L. Anne Hendricks, S. Guadarrama, M. Rohrbach, S. Venugopalan, K. Saenko, and T. Darrell. Long-term recurrent convolutional networks for visual recognition and description. In *Proc. IEEE CVPR*, pages 2625–2634, 2015.

[13] U. Dreher. Duda hart pattern classification and scene analysis. 1973.

[14] C. Ellis, S. Z. Masood, M. F. Tappen, J. J. LaViola, and R. Sukthankar. Exploring the trade-off between accuracy and observational latency in action recognition. *IJCV*, 101(3):420–436, 2013.

[15] O. Fukuda, T. Tsuji, M. Kaneko, and A. Otsuka. A human-assisting manipulator teleoperated by EMG signals and arm motions. *IEEE Trans. on Robotics and Automation*, 19(2):210–222, 2003.

[16] R. Girdhar, D. Ramanan, A. Gupta, J. Sivic, and B. Russell. Action-VLAD: Learning spatio-temporal aggregation for action classification. In *Proc. IEEE CVPR*, volume 2, page 3, 2017.

[17] L. Gorelick, M. Blank, E. Shechtman, M. Irani, and R. Basri. Actions as space-time shapes. *IEEE TPAMI*, 29(12), 2007.

[18] A. Hernandez Ruiz, L. Porzi, S. Rota Bulò, and F. Moreno-Noguer. 3D CNNs on distance matrices for human action recognition. In *Proc. ACM MM*, pages 1087–1095, 2017.

[19] S. Hochreiter and J. Schmidhuber. Long short-term memory. *Neural computation*, 9(8):1735–1780, 1997.

[20] D. Huang, S. Yao, Y. Wang, and F. De La Torre. Sequential max-margin event detectors. In *Proc. ECCV*, pages 410–424, 2014.

[21] I. Jolliffe. Principal component analysis. In *IESS*, pages 1094–1096. 2011.

[22] W. Kay, J. Carreira, K. Simonyan, and others. The kinetics human action video dataset. *arXiv:1705.06950*, 2017.

[23] H. S. Koppula, R. Gupta, and A. Saxena. Learning human activities and object affordances from RGB-D videos. *IJRR*, 32(8):951–970, 2013.

[24] W. Li, Z. Zhang, and Z. Liu. Action recognition based on a bag of 3D points. In *Proc. IEEE CVPR*, pages 9–14, 2010.

[25] A. Liaw, M. Wiener, et al. Classification and regression by random-forest. *R news*, 2(3):18–22, 2002.

[26] S. Mika, G. Ratsch, J. Weston, B. Scholkopf, and K.-R. Mullers. Fisher discriminant analysis with kernels. In *IEEE NNSP*, pages 41–48, 1999.

[27] B. Ni, G. Wang, and P. Moulin. RGBD-HuDaAct: A color-depth video database for human daily activity recognition. In *Consumer Depth Cameras for Computer Vision*, pages 193–208. Springer, 2013.

[28] E. Ohn-Bar and M. Trivedi. Joint angles similarities and HOG2 for action recognition. In *Proc. IEEE CVPRW*, pages 465–470, 2013.

[29] O. Oreifej and Z. Liu. HON4D: Histogram of oriented 4D normals for activity recognition from depth sequences. In *Proc. IEEE CVPR*, pages 716–723, 2013.

[30] D. Pagliari and L. Pinto. Calibration of kinect for Xbox One and comparison between the two generations of microsoft sensors. 15:27569–27589, 10 2015.

[31] A. Pfister, A. M. West, S. Bronner, and J. A. Noah. Comparative abilities of Microsoft Kinect and Vicon 3D motion capture for gait analysis. *JMET*, 38(5):274–280, 2014.

[32] H. Rahmani, A. Mahmood, D. Q. Huynh, and A. Mian. HOPC: Histogram of oriented principal components of 3D pointclouds for action recognition. In *Proc. ECCV*, pages 742–757, 2014.

[33] G. Robertson, G. Caldwell, J. Hamill, G. Kamen, and S. Whittlesey. *Research methods in biomechanics*. Human Kinetics, 2013.

[34] W. Sato, T. Fujimura, and N. Suzuki. Enhanced facial EMG activity in response to dynamic facial expressions. *International Journal of Psychophysiology*, 70(1):70–74, 2008.

[35] B. Scholkopf and A. J. Smola. *Learning with kernels: support vector machines, regularization, optimization, and beyond*. MIT press, 2001.

[36] P. Scovanner, S. Ali, and M. Shah. A 3-dimensional SIFT descriptor and its application to action recognition. In *Proc. ACM MM*, pages 357–360, 2007.

[37] A. Shahroudy, J. Liu, T.-T. Ng, and G. Wang. NTU RGB+D: A large scale dataset for 3D human activity analysis. In *Proc. IEEE CVPR*, pages 1010–1019, 2016.

[38] J. Smisek, M. Jancosek, and T. Pajdla. 3d with kinect. In *Consumer depth cameras for computer vision*, pages 3–25. Springer, 2013.

[39] T. Soo Kim and A. Reiter. Interpretable 3D human action analysis with temporal convolutional networks. In *Proc. IEEE CVPR*, pages 20–28, 2017.

[40] K. Soomro, A. R. Zamir, and M. Shah. UCF101: A dataset of 101 human actions classes from videos in the wild. *arXiv:1212.0402*, 2012.

[41] J. Sung, C. Ponce, B. Selman, and A. Saxena. Human activity detection from RGBD images. In *Proc. AAAI*, 2011.

[42] R. Vemulapalli, F. Arrate, and R. Chellappa. Human action recognition by representing 3D skeletons as points in a lie group. In *Proc. IEEE CVPR*, pages 588–595, 2014.

[43] ViconSystem. The standard vicon full-body model (plug-in gait) marker placement scheme, 2010.

[44] H. Wang and L. Wang. Modeling temporal dynamics and spatial configurations of actions using two-stream recurrent neural networks. In *Proc. IEEE CVPR*, 2017.

[45] J. Wang, X. Nie, Y. Xia, Y. Wu, and S.-C. Zhu. Cross-view action modeling, learning and recognition. In *Proc. IEEE CVPR*, pages 2649–2656, 2014.

[46] K. Wang, X. Wang, L. Lin, M. Wang, and W. Zuo. 3D human activity recognition with reconfigurable convolutional neural networks. In *Proc. ACM MM*, pages 97–106, 2014.

[47] L. Wang, Z. Ding, and Y. Fu. Adaptive graph guided embedding for multi-label annotation. In *Proc. IJCAI*, pages 2798–2804, 2018.

[48] L. Wang, Z. Ding, and Y. Fu. Learning transferable subspace for human motion segmentation. In *Proc. AAAI*, 2018.

[49] L. Wang, Z. Ding, and Y. Fu. Low-rank transfer human motion segmentation. *IEEE TIP*, 28(2):1023–1034, 2019.

[50] L. Wang, Z. Ding, Z. Tao, Y. Liu, and Y. Fu. Generative multi-view human action recognition. In *Proc. IEEE ICCV*, pages 6212–6221, 2019.

[51] L. Wang, Y. Liu, C. Qin, G. Sun, and Y. Fu. Dual relation semi-supervised multi-label learning. In *Proc. AAAI*, 2020.

[52] L. Wang, Y. Xiong, Z. Wang, Y. Qiao, D. Lin, X. Tang, and L. Van Gool. Temporal segment networks: Towards good practices for deep action recognition. In *Proc. ECCV*, pages 20–36, 2016.

[53] P. Wang, W. Li, Z. Gao, J. Zhang, C. Tang, and P. O. Ogunbona. Action recognition from depth maps using deep convolutional neural networks. *IEEE THMS*, 46(4):498–509, 2016.

[54] P. Wei, Y. Zhao, N. Zheng, and S.-C. Zhu. Modeling 4D human-object interactions for event and object recognition. In *Proc. IEEE ICCV*, pages 3272–3279, 2013.

[55] L. Xia, C.-C. Chen, and J. Aggarwal. View invariant human action recognition using histograms of 3D joints. In *Proc. IEEE CVPR*, pages 20–27, 2012.

[56] S. Yan, Y. Xiong, and D. Lin. Spatial temporal graph convolutional networks for skeleton-based action recognition. *arXiv preprint arXiv:1801.07455*, 2018.

[57] G. Yu, Z. Liu, and J. Yuan. Discriminative orderlet mining for real-time recognition of human-object interaction. In *Proc. ACCV*, pages 50–65, 2014.

[58] J. Zhang, W. Li, P. O. Ogunbona, P. Wang, and C. Tang. RGB-D-based action recognition datasets: A survey. *Pattern Recognition*, pages 86–105, 2016.