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

Lichen Wang[1], Bin Sun[1], Joseph Robinson[1], Taotao Jing[1], Yun Fu[1,2]
[1]Department of Electrical and Computer Engineering, Northeastern University, Boston, MA
[2]Khoury College of Computer and Information Sciences, Northeastern University, Boston, MA
{wang.lich,sun.bi,robinson.jo,jing.t}@husky.neu.edu,yunfu@ece.neu.edu

ABSTRACT

Multi-modal human motion analysis is a critical and attractive research topic. Most existing multi-modal action datasets only provide visual modalities such as RGB, depth, or low quality skeleton data. In this paper, we introduce a new, large-scale dataset named EV-Action dataset. It consists RGB, depth, electromyography (EMG), and two skeleton modalities. Compared with others, our 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, and acceleration with higher accuracy, sample frequency, and more skeleton markers. (2) we include an EMG modality. While EMG is used as an effective indicator for biomechanics area, it has yet to be well explored in multimedia, computer vision, and machine learning areas. To the best of our knowledge, this is the first action dataset with EMG modality. In this paper, we introduce the details of EV-Action dataset. A simple yet effective framework for EMG-based action recognition is proposed. Moreover, we provide state-of-the-art baselines for each modality. The approaches achieve considerable improvements when EMG is involved, and it demonstrates the effectiveness of EMG modality in human action analysis tasks. We hope this dataset could make significant contributions to signal processing, multimedia, computer vision, machine learning, biomechanics, and other interdisciplinary fields.

KEYWORDS

Human motion analysis, Multi-modality, Kinect, Vicon system, Motion capturing, Skeleton, Biomechanics, Electromyography, EMG

1 INTRODUCTION

There are many use-cases of multi-modal systems for human motion understanding. For instance, in surveillance systems (*e.g.*, airports, offices, homes), forensic investigations (*e.g.*, event detection, behavior prediction), biomechanics (*e.g.*, gait analysis, joint mechanics, prosthetic designs, sports medicines), human-computer interaction (*e.g.*gaming, interactive physical therapy), and so on. It has attracted lots of attentions of researchers over the years in multimedia, computer vision, and signal processing area [23, 28, 47, 57]. The availability of datasets tends to directly impact the progress of research. From the start, action datasets only consisted RGB modality [47, 54, 57, 62]. Later on, as 3D sensors became more accessible, several datasets included the depth modality (*i.e.*, RGB-D) [29, 30, 35, 59, 70]. This paved a way for researchers to propose more effective approaches in terms of multi-modal methods [20, 23, 28, 50, 65]. After that, skeleton data was introdeced by some works [32, 55, 67].As a consequence, the performance of these systems have been further improved. However, most skeleton information of these datasets was directly obtained from Kinect sensors [71], resulting in low localization accuracy. Skeleton modal captured by more accurate devices was released [10], while RGB-D modals were not included.

In this paper, we introduce our EV-Action dataset, which includes all visual modalities mentioned above (*i.e.*, RGB, depth, and two skeleton modalities). An optical tracking-based Vicon system [39] is deployed to capture high-quality skeleton data. Compared with Kinect, Vicon achieves significantly higher sampling rate (*i.e.*, 100 vs. 30 fps), higher localization accuracy, and more skeleton markers (*i.e.*, 39 vs. 26). Thus, it could provide more comprehensive skeleton motion information in terms of location, trajectory, velocity, and acceleration. In addition to visual modalities, we also collected Electromyography (EMG) signals to measure the electrical activity of human skeletal muscles as a function of the intensity of force [33]. EMG is regularly used in medical and biomechanics fields. It has not yet been explored in the fields of multimedia or computer vision. In our dataset, all modalities are captured simultaneously with corresponding action labels frame by frame. To the best of our knowledge, this is the first action dataset that includes EMG signals. The goal of generating this dataset is to explore the correlation across different modalities and improve learning performances of action recognition systems. We believe that this dataset will contribute significantly to the research fields of signal processing, multimedia, computer vision, machine learning, biomechanics, and other interdisciplinary sub-fields. The contributions of this paper are listed as follows:

(1) We design and construct a data collection center with optical tracking based Vicon system and RGB-D modality Kinect-V2 systems. This allows us to build the proposed action data with all the visual modalities (*i.e.*, RGB, Depth, Skeleton-K, and Skeleton-V).
(2) EMG modality from skeletal muscles is extracted while actions were being performed. To the best of our knowledge, this is the first action dataset that includes EMG modal. EMG 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.

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

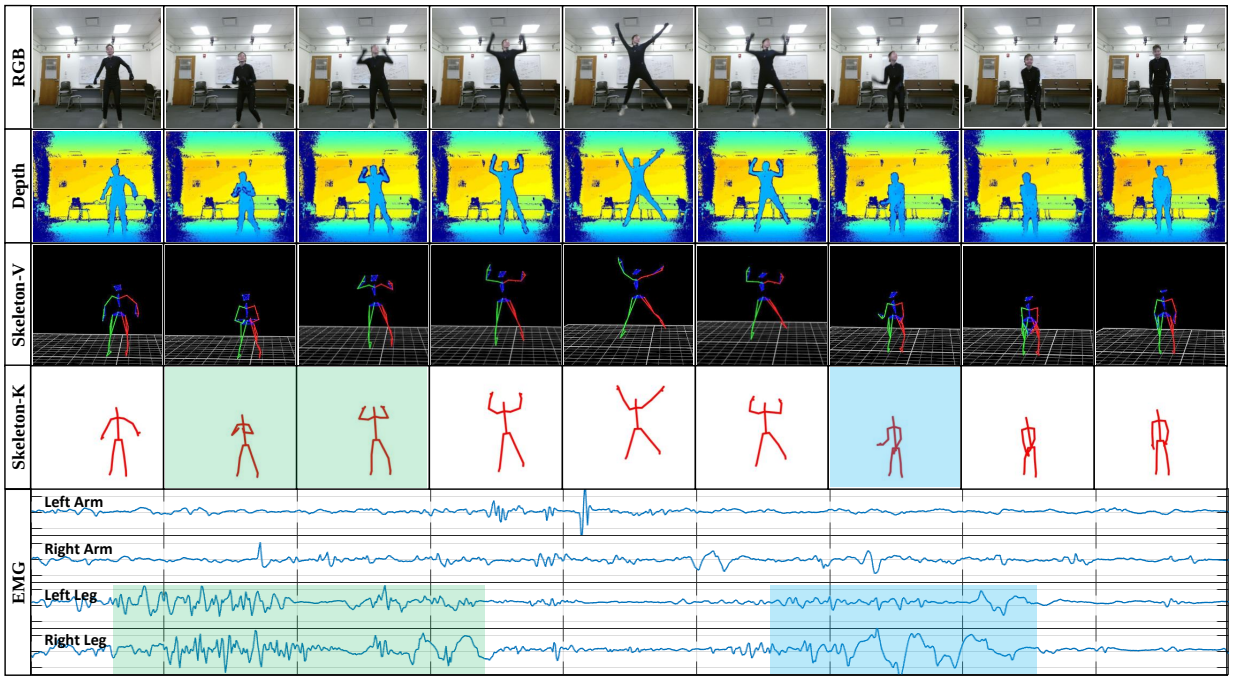


Figure 1: Visualization of sample frames of all the modalities in EV-Action dataset. The skeleton data were captured by both the KinectV2 (Skeleton-K) and Vicon system (Skeleton-V). Colored boxes show the correlations between visual modalities and EMG (*i.e., Take Off* and *Touch Down*). We can clearly 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 for each of the 20 action classes.

(4) We define experimental settings and provided the state-of-the-art benchmarks for each modality. EMG is merged with visual features to achieve considerable improvements which further demonstrates the complimentary of the EMG modal.

2 RELATED WORKS

2.1 RGB/D and Skeleton Datasets

Originally, small-scale datasets included several action classes (*e.g.*, Keck Gesture [24], Weizmann Activity [18], and UT-Interaction datasets [44]), each with tens of samples to learn action-based tasks [70]. Upon the arrival of deep learning, large-scale datasets were introduced with just the RGB modality (*e.g.*, UCF101 [52] and Kinetics [27]). Later on, as 3D and depth sensor became popular, RGB-D datasets were released (*e.g.*, MSR-Action3D [30], RGBD-HuDaAct [35], MSRC-12 [59] and MSR-Activity3D [29]). Due to the space and budgeting constraints, most RGB-D datasets were collected using low-cost Kinect sensors [49, 71]. In addition, Kinect sensors can extract skeleton data, as introduced in MAD [22], UCF-Kinect [15] and NTU-RGBD [48] datasets. However, since Kinect utilized depth information for estimating the skeletons, its accuracy and reliability were low. [26] proposed a large-scale dataset with RGB and skeleton modalities. Nonetheless, it was only designed for pose estimation without action labels or multi-modalities.

2.2 Multi-Modal Action Datasets

We consider the datasets containing more than RGB-D modality as being multi-modal datasets. Currently, only a few datasets provide additional action information. NTU-RGBD [48] and PKU-MMD [9] contained infrared frames captured by a Kinect. CMU-MMAC [10] utilized an optical tracking technique associated with multi-visual modalities to capture action sequences without any specific labels. UTD-MHAD [5] utilized a Kinect and a single wearable inertial sensor to capture RGB-D, skeleton, and inertial signals. However, the single of the inertial sensor was not captured in a consistent manner for different subjects. Therefore, this modality is severely limited and sporadic. Compared with these datasets, our dataset utilizes 39 markers to capture precise location, trajectory and acceleration information at a high frame rate (100 Hz). Thus, we propose the **most accurate and comprehensive** dataset of this kind.

2.3 EMG Signal

Electromyography (EMG) is an electrodiagnostic technique to evaluate the electrical activity produced by skeletal muscles [43]. Comprehensive surveys on the topic can be found in [11, 33, 43]. EMG signals provide easy access to physiological processes that cause muscles to generate force, produce movement, and perform actions. Typically, EMG is used in neural science, biomechanics, and signal processing fields. For Instance, EMG is used in hand gesture recognition [6, 7, 40], robot arm control [16], expression classification [19, 45] and human-computer interactions [3]. Since EMG activates before motion are visible [33] which could foresee potential information such as intention, force, and even mental activities–information that cannot be recognized with visual evidence alone. 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 connections

Table 1: 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	-	$84.1° \times 53.8°$	1	-	8-bit
Kinect-V2	Depth	512×424	30	-	$70.6° \times 60.0°$	1	0.5-4.5 m	16-bit
	Skeleton-K	-	30	26	-	-	-	-
Vicon-T40s	Skeleton-V	$2,336 \times 1,728$	100	39	$98.1° \times 50.1°$	8	12 m	10-bit
Delsys-Trigno	EMG	-	1,000	-	-	4	± 22 mV	16-bit

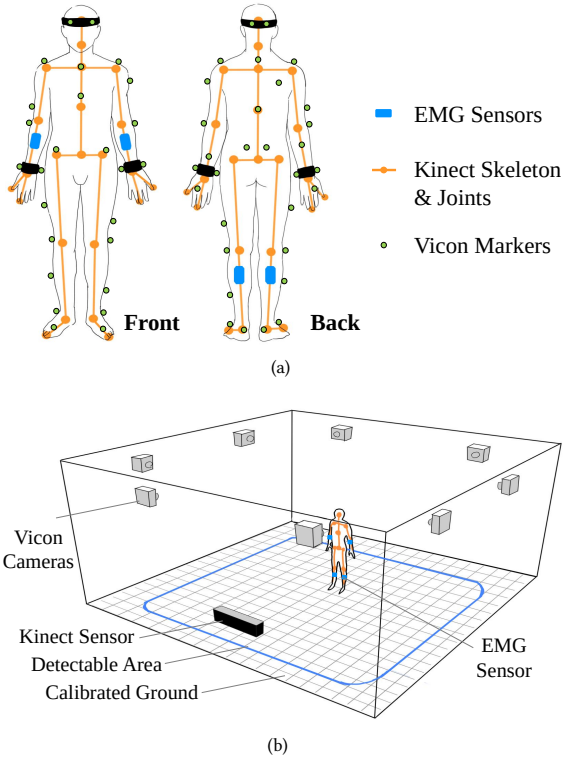


Figure 2: (a) Sensor placement schemes. Orange lines and spots indicate Kinect skeleton with 26 joints. Small gray points denote Vicon markers. And blue blocks indicate the EMG sensors. (b) Data collection center environment setup.

that can be explored by this association, and the potential applications that could be had, we generated EV-Action dataset with EMG as one of the critical modalities available for human action analysis.

3 EV-ACTION DATASET

In this section, we introduce the details of EV-Action dataset. Specifically, we discuss data formats, collection process, and data statistics.

3.1 Sensors and Setup

We review details and specifications for each of these, along with the setup of our data collection center, which consisted of a Kinect-V2 sensor [1], 4 wireless EMG sensors, and 8 Vicon-T40s cameras.

Kinect Sensors [1, 49, 71] capture RGB-D modalities of subjects, from which skeleton information can be extracted from the depth information. We used a second generation Kinect for the data collection (*i.e.*, Kinect-V2 [1, 38]). In summary, the inner workings of Kinect-V2 can be characterized by two-fold. First, a high resolution camera captures HD frames (1,920×1,080) at 30 fps with a wide field of view (70°×60°). Second, a depth sensor with a resolution of 512×424 is more robust and efficient for doing pose estimation with reference to 26 joints (Figure 2(a)). During data collection, the subjects perform actions in front of a Kinect-V2 sensor (Figure 2(b)).

Vicon System utilizes optical tracking-based technology to capture skeleton data with more accurate and comprehensive motion information [39]. We deploy 8 Vicon-T40s infrared cameras to capture the stickup marks on each subject (Figure 2(a)). The cameras sample data points as 10-bit gray-scale frames at 100 fps and with a resolution of 2336×1728. Then, the frames were calibrated and labeled to obtain skeleton information. We follow the standard full-body marker placement scheme [56] by placing 39 markers around human body, as shown in Figure 2(a). To this end, it can capture precise and comprehensive motion information, such as the second bounce in the *Fall Down* action class. Also, due to the high frame rate and localization precision, high quality trajectories and accelerations were obtainable in reference to ground coordinates. Figure 3 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.

EMG Sensors capture electromyography signals from human muscles. For the convenience and comfort of the human subjects, we use wireless EMG sensors manufactured by DELSYS. Each sensor captures 16-bit EMG signal at 1,000 Hz. This enables the sensors to record signals lower than 500 Hz, which, thus, covers the whole frequency spectrum of skeletal EMG (*i.e.*, 20-450 Hz). We attached 4 sensors to each subject: the middle of each forearm and the shank muscles (Figure 2(a)). The reasons for selecting these locations are explainable in three-fold. First, the most common actions utilize arms and legs. Second, based on prior research [11], the location of each muscle (mid-line of the muscle in the belly that is between the nearest innervation zone and the myotendinous junction) gives off a signal of highest amplitude, which makes the signal most responsive to the corresponding action performed. Third, 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 Vicon cameras placed around the parameter of a $4.6\text{m} \times 4.6\text{m}$ room. Thus, this room has a detectable area of $3\text{m} \times 3\text{m}$. From this, 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 inside the detectable area with the face of this Kinect sensor as the front. Also, 4 EMG sensors were connected to each subject (Figure 2(b)).

3.2 Dataset Description

Completeness, comprehensiveness, and diversity were highly considered when building our EV-Action dataset. To make the dataset

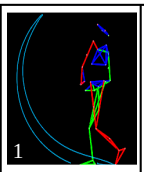
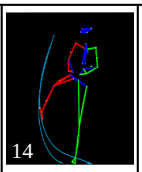
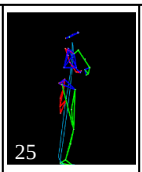
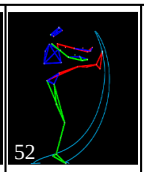
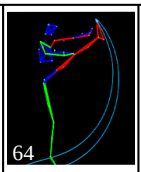
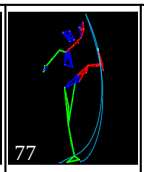
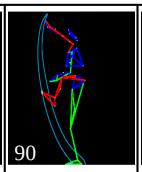
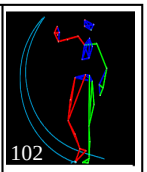
Vicon SK	 1	 14	 25	 39	 52	 64	 77	 90	 102
Angle	0°	45°	90°	135°	180°	225°	270°	315°	360°
Time(s)	0.00	0.13	0.25	0.40	0.53	0.65	0.78	0.91	1.04

Figure 3: Visualization of a subject performing a kicking action across view angles and time (i.e., across Vicon sensors in different locations and video frames, respectively). The blue curve highlights the trajectory of the marker placed on the toe of the subject. Clearly, the Vicon system precisely captures detailed motion information of the human body. Frame numbers are shown in left bottom which indicate the high sampling rate of Vicon system.

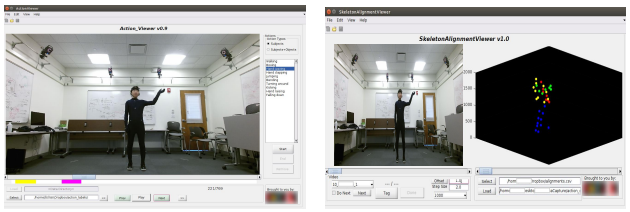
Table 2: Comparison between EV-Action dataset and other popular multi-modal datasets. EV-Action is one of the largest multi-modal datasets and significantly outperforms other datasets in modal diversity, subject numbers, and sample clips.

Datasets	Samples	Classes	Subjects	FrameRate(fps)	Sensors	Modalities
MSR-Action3D [30]	567	20	10	30	RGB-Cam	D+SK
RGBD-HUDA [35]	1189	13	30	30	KinectV1	RGB+D
CAD-60 [53]	60	12	4	30	KinectV1	RGB+D+SK
Action4^2 [8]	6844	14	24	30	KinectV1	RGB+D
CAD-120 [29]	120	20	4	30	KinectV1	RGB+D+SK
3D Action Pairs [37]	360	12	10	30	KinectV1	RGB+D+SK
Multiview 3D Event [66]	3815	8	8	30	KinectV1	RGB+D+SK
Online RGB+D Action [69]	336	7	24	30	KinectV1	RGB+D+SK
Northwestern-UCLA [60]	1475	10	10	30	KinectV1	RGB+D+SK
UWA3D Multiview [42]	900	30	10	30	KinectV1	RGB+D+SK
Office Activity [61]	1180	20	10	30	KinectV1	RGB+D
UTD-MHAD [5]	861	27	8	30+50	KinectV1+WIS	RGB+D+SK
UWA3D Multiview II [36]	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 3: 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

practical and generalizable enough for real-world use-cases, we included 20 common actions (*i.e.*, 10 were done by a single subject and the other 10 were done by that same subject interacting with different objects). A complete list of action classes is listed in Table 3. The dataset includes 70 subjects performing each of the action classes 5 times (*i.e.*, 100 action clips per subject). To introduce diversities, we asked the subjects to use a slightly different style in each sample. All-in-all, resulting in 7,000 action clips at multiple views. Table 2 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 action 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 time. Referencing figure 1, we notice that the EMG signal activates prior to the actions of *Take Off* and *Touch Down* becoming visible. We also notice that the durations of the EMG signals are typically longer than the visual modalities. These patterns are unrecognizable from any visual modal. It demonstrates



(a) (b)

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

that EMG does provide unique and useful information for more deep and sophisticated action analytical tasks.

3.3 Data Labeling

There are two steps for data labeling: (1) annotating the actions in RGB modality which is captured by the Kinect, and then (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 3. Actions are listed in the same order as performed associated with shortcut keys. These tricks allow the labeling to be done both quicker and with less human input. **Data Alignment:**

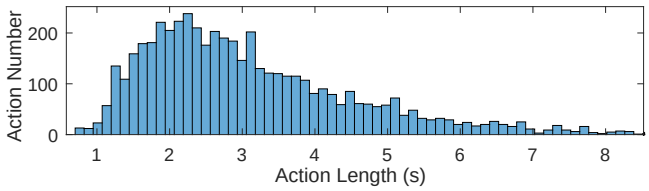


Figure 5: Histogram of the length distribution of all videos.

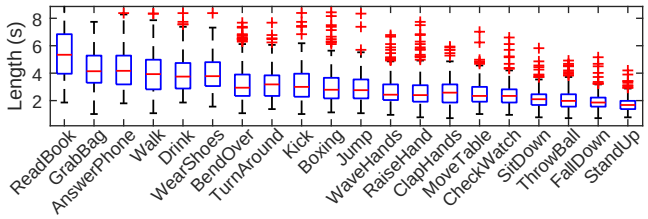


Figure 6: Video length distribution of each action.

We then align the clip with the clip captured by the Vicon system. This allow the action labels to be propagated across the different modalities. We develop another MATLAB tool that allow us to visually align single frame of the RGB and the skeleton from the Kinect and the Vicon systems, respectively (Figure 4(b)).

4 DATASET 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 on average per action category. 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 for action understanding tasks while increase the potential of dataset to generalize well, make it better suited for practical use cases.

Root Mean Squared (RMS) is an effective method to pre-process EMG data. We obtain the average of RMS value of each action and surprisingly notice that the shank muscles have significantly higher (2 times) EMG 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 their left legs even for *Check Watch* (might also for body balance need; in order to rise left arm to check watch, they should hold/balance their body by left leg to take over left arm).

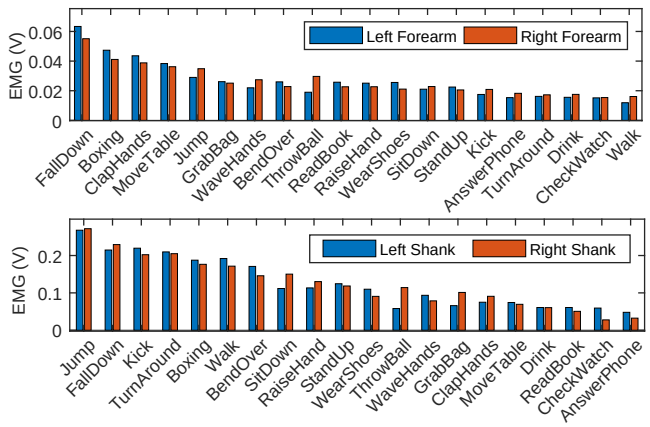


Figure 7: The average of Root Mean Square (RMS) value of the EMG recordings in different actions. We separate the value of upper body (left and right forearm) and lower body (left and right shank) for better discussion.

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 all the markers pasted on each subject, there were situations that some markers were obstructed to the views of every cameras, such as *Fall Down* and *Sit Down* due to a large area of a subject's body being in contact with the ground/chair. Once the occluded marker is again detected, Vicon could re-localized the respective 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 methods can also explore EV-Action dataset. This also leads to believe that more sophisticated algorithms are needed to achieve higher performance rating (*e.g.*, missing-modality or multi-view algorithms). Therefore, we split the actions into two action subsets which are *Single-Person* and *Person-Object* sets. The rest modalities (*i.e.*, RGB, Depth, Skeleton-K and EMG) are stable across all actions. The errors for these modalities are negligible.

5 EXPERIMENTS

Several state-of-the-art methods 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. To the best of our knowledge, this work is the first to model the non-visual EMG signal for action recognition. We achieved considerable performance improvements by employing a simple, yet effective fusion technique (*i.e.*, fused at the feature-level). This is not only the evidence of the complimentary information of the EMG data when fused with visual data, but also a great promise for further improvement by providing more sophisticated learning frameworks and fusion techniques. All modalities can be involved in the learning procedure. 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.

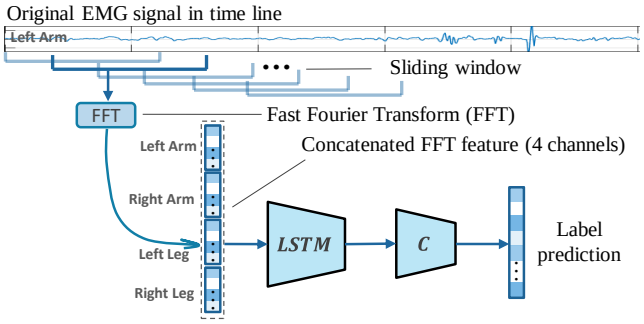


Figure 8: EMG based action recognition framework.

5.1 Experimental Settings

Benchmarks on EV-Action dataset followed conventional classification settings. The action clips from 56 subjects were used during training (*i.e.*, 5,600 clips), while the other 14 subjects were set aside for testing (*i.e.*, 1,400 clips). All experiments were evaluated in terms of classes separately. To this end, the class-specific evaluations allowed for a finer-grained analysis between different actions.

5.2 EMG Signal

Signal processing methods associated with hand-crafted features are usually deployed for EMG data analysis. In this section, we first introduce the classical EMG processing baselines; then, we design a novel deep-structure framework for EMG action recognition; the performances of all baselines are introduced and discussed lastly.

5.2.1 Conventional EMG Processing Methods. We first introduce the conventional procedures for EMG classification.

Noise is inherited in EMG. Butterworth filter [4] yields a flat frequency response which is effective to filter out EMG noise. The generalized equation of the frequency response of n-th order Butterworth filter 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. Root Mean Square (RMS) is a simple and effective feature extraction approach for EMG signal [33]. The expression of EMG 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. Since the obtained feature from sliding windows is large size, thus, dimension reduction is required. Principal Component Analysis (PCA) [25] and Linear Discriminative Analysis (LDA) [34] are utilized in our baselines.

Three classifiers are tested on including Support Vector Machine (SVM) [46], K-Nearest Neighbors (KNN) [14], and Random Forests (RF) [31]. The results are shown in Table 4, we find that random forest method on features after PCA dimension reduction has the best classification performance which is 35.12%. As noise of raw EMG signals occurred during collecting period, choosing the best way to extract features and reduce the dimension are the crucial works to improve classification performance.

Table 4: EMG classification accuracy based on different dimension reduction (Dim-Red) approaches and classifiers.

Classifier \ Dim-Red	(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**	-	-

5.2.2 Our approach for EMG Processing. We then introduce our deep modal-based EMG action recognition approach.

It is a novel, effective, and end-to-end trainable deep neural network structure for EMG action recognition. Both the noise elimination and the EMG recognition are done simultaneously. The framework of our approach is shown in Figure 8. Sliding windows are first employed to extract EMG signal from each channel. Differently, we utilize Fast Fourier Transform (FFT) [2] instead of RMS as the initial feature extraction 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 provides more comprehensive signal information in each sliding window. After obtaining FFT feature, 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) [21] networks. LSTM is an effective approach to analyze time series data which remembers values over arbitrary time intervals. Mathematically, the updating principles can be written as:

$$\begin{aligned} \mathrm{i}_t &= \sigma(U_i X_t + W_i h_{t-1}) \\ \mathrm{f}_t &= \sigma(U_f X_t + W_f h_{t-1}) \\ \mathrm{o}_t &= \sigma(U_o X_t + W_o h_{t-1}) \\ \tilde{c}_t &= \tanh(U_c X_t + W_c h_{t-1}) \\ \mathrm{c}_t &= \mathrm{f}_t \cdot \mathrm{c}_{t-1} + \mathrm{i}_t \cdot \tilde{c}_t \\ h_t &= \tanh(\mathrm{c}_t) \cdot o_t, \end{aligned} \tag{1}$$

where X_t is the t-th input FFT vector, i_t is the t-th input gate activation vector, h_t is the t-th hidden state, f_t is the t-th forget gate activation vector, o_t is the t-th output gate activation vector, c_t is the tth cell state vector, and $\tilde{c}_t$ is the t-th new memory content. LSTM outputs the representations of FFT features and a classifier is trained to predict the final label. The classifier $C(\cdot)$ is a fully connected network with softmax activation. We choose the last hidden layer of LSTM as the input of $C(\cdot)$. The loss function is as follow.

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

where Y is the instance label, X is the original data in time domain. $F(\cdot)$ is the feature extractor using sliding window and FFT. $G(\cdot)$ is LSTM network. $C(\cdot)$ is a full connected network. In the implementation, we use half-overlapping sliding window (*i.e.*, overlap ratio 50%) with the window size 200 to extract data blocks. If the video has several data points remain after extraction, we will ignore these data because the last few points could not considerably affect the performance. 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 can be represented by a 400-dimension vector. Since different clips may have different lengths when we configured our model to input 5-second long videos. We repeat the actions again to satisfy the required length for short videos. Besides, we cut off the long videos which does not considerably effect the performance. The LSTM structure has a hidden layer of 1024-dimension. The result (Table 4) denotes that the our approach significantly outperforms conventional methods which also indicates the effectiveness of EMG in action analytical tasks.

5.3 RGB & Depth

We first evaluate single-modal action recognition baselines using the RGB and depth modal captured by Kinect sensors.

Action Vector of Local Aggregated Descriptor (Action-VLAD) [17] is an effective video representation for action classification which aggregates local convolutional features across the entire spatio-temporal extent. It integrates two-stream networks with learnable spatio-temporal feature aggregation and is trainable in an end-to-end manner. We fine-tune the softmax layer on the top of the ImageNet [12] initialization for our evaluation.

Temporal Segment Networks (TSN) [63] combines a sparse temporal sampling approach with video-level supervision. In this way, the entire video was learned effectively. The Caffe framework was utilized for the implementation.

Long-term Recurrent Convolutional Networks (LRCN) [13] deploys a hierarchical visual feature extractor associated with a temporal dynamic recognition model. LRCN is dynamic for different lengths of videos and has the potential to recognize actions in complex videos. LRCN is capable of end-to-end training, we used L-2 regularization and implemented it based on Torch.

Weighted Depth Motion Maps (WDMM) [41] recognizes dynamic 3D human gestures and actions from depth videos. WDMM is based on linearly aggregation of spatio-temporal information. It utilizes a video summarization step for hierarchical representation learning, which results in increasing intra-class similarity and inter-class dissimilarities. We chose 80 as the number of visual words and 110 as the number of PCA components to obtain depth features.

Weighted Hierarchical Depth Motion Maps (WHDMM) [64] utilizes a weighted hierarchical depth motion map and a three-channel deep convolutional neural network together for action recognition. We only use the front view to train and test since we do not have multi-view depth images. The remaining parameters followed the original work.

5.4 Skeleton-Kinect

In this section, we introduce the baselines of skeleton based action recognition approaches.

Temporal Convolutional Networks (TCN) [51] learns an interpretable spatio-temporal representation for human action recognition. To train the model on the Kinect Skeleton modality, we modify the data as the same format of NTU-RGB-D[48].

Two Stream Recurrent Neural Network (TSRNN) [58] provides a two-stream RNN architecture to explore both temporal dynamics and spatial configurations for action classification. We process the data in the same way as TCN. We use bidirectional LSTM as the paper suggested, set the batch size to 256, learning rate to 0.02 and maximum iteration number to $2,000$.

Spatial Temporal Graph Convolution Network (STGCN) [68] 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 our data with 80 epochs, using SGD as the optimizer, and a learning rate of 0.01.

5.5 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** [51] 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** [58] needs the part of the body, *i.e.*, two arms, two legs and one trunk, as well as whole body for action recognition. We use the index groups of different body parts via the 39 joints of Vicon and keep other parameters be consistent. **STGCN** [68] needs a joint adjacency graph. Thus, we generate the connection graph for 39-joint while other parameters are consistent.

5.6 Skeleton-(Kinect/Vicon) & EMG

To prove the effectiveness and complementariness of EMG modal, we combine EMG as well as skeleton modalities together in low-level domain for action recognition. The results are compared with the performances with single Kinect or Vicon skeleton modality.

TCN-RMS We first calculate the RMS value in each sliding window to obtain 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 value directly with the skeleton data (considered as skeleton modality and auxiliary EMG modality) and input the combination data to TCN [51] for classification. The hyper parameters we used are the same as the parameters mentioned in section 5.4. 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 tested 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.

5.7 Results and Analysis

We used top-1 accuracy to evaluate each baseline (Table 5). For RGB modality, TSN outperformed other benchmarks with an accuracy of 73.6%. Action-VLAD was the next best, and then LRCN. 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 21.6% 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 79.0% accuracy obtained from the skeleton data alone, the added EMG

Table 5: 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[63]	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	73.6
	LRCN[13]	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	65.4
	VLAD[17]	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[41]	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[64]	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[51]	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	80.1
	TSRNN[58]	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	79.1
	STGCN[68]	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	79.6
SK-V	TCN[51]	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[58]	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[68]	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
S-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	81.7
S-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	67.4
	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	51.1	64.4

signal improved this by 4% (*i.e.*, TCN-RMS with 83.6%). After the analysis of the results, we conclude that the performance on 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, especially when a subject wears a black suit. 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 some actions are extremely high(*i.e.*, *Jump*) which shows that EMG signal is distinctive in some actions. However, similar actions(*i.e.*, *Wave Hand* and *Raise Hand*; *Stand Up* and *Sit down*), which can be easily classified by other modality, are always hard for EMG to distinguish. 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 and the models are not robust enough to handle these missing points scenarios. (2) The generation strategy of spatial graph for Vicon data is different from the Kinect skeleton model default setting. More robust and well-matching connection graphs could be used for this modality. To this end, there are still many open questions and challenges left for future exploration.

Accuracies of the benchmarks on skeleton data do not match up with the high scores reported on other datasets. Comparing our skeleton dataset with the NTU-RGBD [48] 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 than other datasets. The number of training samples for these benchmarks drastically affects performance ratings. (2) We only utilized 3D reference frame from the skeleton modality, while the approach of [48] 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. Clearly, the EMG data complements the visual modalities, as there is a significant improvement in results by fusing the modalities in just a simple manner. We believe more advanced feature extraction methods and multi-modality fusion strategies could further improve the learning performance.

6 CONCLUSION

We have introduced a new multi-modal human action recognition dataset, EV-Action dataset. It consisting of RGB, depth, skeleton, and EMG data. All modalities have been labeled and aligned across 7,000 samples of 70 different subjects. Our dataset has made two major improvements compared with other action-based video data collections. (1) we have utilized an optical tracking based Vicon system to capture more accurate and comprehensive skeletal data, and (2) we have included a non-visual EMG modality associated with other visual modalities. We also have provided several the state-of-the-art benchmarks for each modality. Moreover, we have proposed simple yet effective framework for EMG-based action recognition and achieved highest performance. Our results have demonstrated that the effective and complimentary information is extractable from EMG signals for human action analytical tasks. The proposed dataset can serve in widespread researches concerning human motion understanding. We hope EV-Action dataset could have a significant impact on multi-media, computer vision, biomechanics, and other interdisciplinary areas.

REFERENCES

[1] Clemens Amon, Ferdinand Fuhrmann, and Franz Graf. 2014. Evaluation of the spatial resolution accuracy of the face tracking system for kinect for windows v1 and v2. In *Proc. of Alps Adria Acoustics Association*. 16–17.

[2] E Oran Brigham and E Oran Brigham. 1988. *The fast Fourier transform and its applications*. Vol. 1. prentice Hall Englewood Cliffs, NJ.

[3] Nan Bu, Masaru Okamoto, and Toshio Tsuji. 2009. A hybrid motion classification approach for EMG-based human–robot interfaces using bayesian and neural networks. *IEEE Trans. Robotics* 25, 3 (2009), 502–511.

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

[5] Chen Chen, Roozbeh Jafari, and Nasser Kehtarnavaz. 2015. Utd-mhad: A multimodal dataset for human action recognition utilizing a depth camera and a wearable inertial sensor. In *Proc. IEEE ICIP*. 168–172.

[6] Xun Chen and Z Jane Wang. 2013. Pattern recognition of number gestures based on a wireless surface EMG system. *Biomedical Signal Processing and Control* 8, 2 (2013), 184–192.

[7] Xiang Chen, Xu Zhang, Zhang-Yan Zhao, Ji-Hai Yang, Vuokko Lantz, and Kong-Qiao Wang. 2007. Hand gesture recognition research based on surface EMG sensors and 2D-accelerometers. In *Proc. IEEE International Symposium on Wearable Computers*. 11–14.

[8] Zhongwei Cheng, Lei Qin, Yituo Ye, Qingming Huang, and Qi Tian. 2012. Human daily action analysis with multi-view and color-depth data. In *Proc. ECCV*. Springer, 52–61.

[9] Liu Chunhui, Hu Yueyu, Li Yanghao, Song Sijie, and Liu Jiaying. 2017. PKU-MMD: A Large Scale Benchmark for Continuous Multi-Modal Human Action Understanding. *arXiv preprint arXiv:1703.07475* (2017).

[10] Fernando De la Torre, Jessica Hodgins, Adam Bargteil, and others. 2008. Guide to the carnegie mellon university multimodal activity (CMU-MMAC) database. *Robotics Institute* (2008), 135.

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

[12] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei. 2009. ImageNet: A Large-Scale Hierarchical Image Database. In *Proc. IEEE CVPR*.

[13] Jeffrey Donahue, Lisa Anne Hendricks, Sergio Guadarrama, Marcus Rohrbach, Subhashini Venugopalan, Kate Saenko, and Trevor Darrell. 2015. Long-term recurrent convolutional networks for visual recognition and description. In *Proc. IEEE CVPR*. 2625–2634.

[14] Ute Dreher. 1973. DUDA HART PATTERN CLASSIFICATION AND SCENE ANALYSIS. (1973).

[15] Chris Ellis, Syed Zain Masood, Marshall F Tappen, Joseph J LaViola, and Rahul Sukthankar. 2013. Exploring the trade-off between accuracy and observational latency in action recognition. *IJCV* 101, 3 (2013), 420–436.

[16] Osamu Fukuda, Toshio Tsuji, Makoto Kaneko, and Akira Otsuka. 2003. A human-assisting manipulator teleoperated by EMG signals and arm motions. *IEEE Trans. Robotics and Automation* 19, 2 (2003), 210–222.

[17] Rohit Girdhar, Deva Ramanan, Abhinav Gupta, Josef Sivic, and Bryan Russell. 2017. ActionVLAD: Learning spatio-temporal aggregation for action classification. In *Proc. IEEE CVPR*, Vol. 2. 3.

[18] Lena Gorelick, Moshe Blank, Eli Shechtman, Michal Irani, and Ronen Basri. 2007. Actions as Space-Time Shapes. *IEEE Trans. PAMI* 29, 12 (2007).

[19] Jamin Halberstadt, Piotr Winkielman, Paula M Niedenthal, and Nathalie Dalle. 2009. Emotional conception: How embodied emotion concepts guide perception and facial action. *Psychological Science* 20, 10 (2009), 1254–1261.

[20] Alejandro Hernandez Ruiz, Lorenzo Porzi, Samuel Rota Bulò, and Francesc Moreno-Noguer. 2017. 3D CNNs on Distance Matrices for Human Action Recognition. In *Proc. ACM Multimedia*. 1087–1095.

[21] Sepp Hochreiter and Jürgen Schmidhuber. 1997. Long short-term memory. *Neural computation* 9, 8 (1997), 1735–1780.

[22] Dong Huang, Shitong Yao, Yi Wang, and Fernando De La Torre. 2014. Sequential max-margin event detectors. In *Proc. ECCV*. Springer, 410–424.

[23] Chengcheng Jia, Yu Kong, Zhengming Ding, and Yun Fu. 2016. Rgb-d action recognition. In *Human Activity Recognition and Prediction*. Springer, 87–106.

[24] Zhuolin Jiang, Zhe Lin, and Larry Davis. 2012. Recognizing human actions by learning and matching shape-motion prototype trees. *IEEE Trans. PAMI* 34, 3 (2012), 533–547.

[25] Ian Jolliffe. 2011. Principal component analysis. In *International Encyclopedia of Statistical Science*. 1094–1096.

[26] Hanbyul Joo, Hao Liu, Lei Tan, Lin Gui, Bart Nabbe, Iain Matthews, Takeo Kanade, Shohei Nobuhara, and Yaser Sheikh. 2015. Panoptic Studio: A Massively Multiview System for Social Motion Capture. In *Proc. IEEE ICCV*.

[27] Will Kay, Joao Carreira, Karen Simonyan, and others. 2017. The kinetics human action video dataset. *arXiv preprint arXiv:1705.06950* (2017).

[28] Yu Kong and Yun Fu. 2015. Bilinear heterogeneous information machine for RGB-D action recognition. In *Proc. IEEE CVPR*. 1054–1062.

[29] Hema Swetha Koppula, Rudhir Gupta, and Ashutosh Saxena. 2013. Learning human activities and object affordances from rgb-d videos. *The International Journal of Robotics Research* 32, 8 (2013), 951–970.

[30] Wanqing Li, Zhengyou Zhang, and Zicheng Liu. 2010. Action recognition based on a bag of 3d points. In *Proc. IEEE CVPR*. 9–14.

[31] Andy Liaw, Matthew Wiener, and others. 2002. Classification and regression by randomForest. *R news* 2, 3 (2002), 18–22.

[32] Diogo Carbonera Luvizon, Hedi Tabia, and David Picard. 2017. Learning features combination for human action recognition from skeleton sequences. *Pattern Recognition* 99 (2017), 13–20.

[33] Roberto Merletti and Philip A Parker. 2004. *Electromyography: physiology, engineering, and non-invasive applications*. Vol. 11. John Wiley & Sons.

[34] Sebastian Mika, Gunnar Ratsch, Jason Weston, Bernhard Scholkopf, and Klaus-Robert Mullers. 1999. Fisher discriminant analysis with kernels. In *IEEE Neural Networks for Signal Processing*. 41–48.

[35] Bingbing Ni, Gang Wang, and Pierre Moulin. 2013. Rgbd-hudaact: A color-depth video database for human daily activity recognition. In *Consumer Depth Cameras for Computer Vision*. Springer, 193–208.

[36] Eshed Ohn-Bar and Mohan Trivedi. 2013. Joint angles similarities and HOG2 for action recognition. In *Proc. IEEE CVPR Workshop*. 465–470.

[37] Omar Oreifej and Zicheng Liu. 2013. Hon4d: Histogram of oriented 4d normals for activity recognition from depth sequences. In *Proc. IEEE CVPR*. 716–723.

[38] Diana Pagliari and Livio Pinto. 2015. Calibration of Kinect for Xbox One and Comparison between the Two Generations of Microsoft Sensors. 15 (10 2015), 27569–27589.

[39] Alexandra Pfister, Alexandre M. West, Shaw Bronner, and Jack Adam Noah. 2014. Comparative abilities of Microsoft Kinect and Vicon 3D motion capture for gait analysis. *Journal of Medical Engineering & Technology* 38, 5 (2014), 274–280.

[40] Angkoon Phinyomark, S Hirunviriya, C Limsakul, and P Phukpattaranont. 2010. Evaluation of EMG feature extraction for hand movement recognition based on Euclidean distance and standard deviation. In *Proc. IEEE ECTI-CON*. 856–860.

[41] S. Kasaei Sergio Escalera R. Azad, M. Asadi. 2017. Dynamic 3D Hand Gesture Recognition by Learning Weighted Depth Motion Maps. https://github.com/rezazad68/Dynamic-3D-Action-Recognition-on-RGB-D-Videos.git. (2017).

[42] Hossein Rahmani, Arif Mahmood, Du Q Huynh, and Ajmal Mian. 2014. HOPC: Histogram of oriented principal components of 3D pointclouds for action recognition. In *Proc. ECCV*. Springer, 742–757.

[43] Gordon Robertson, Graham Caldwell, Joseph Hamill, Gary Kamen, and Saunders Whittlesey. 2013. *Research methods in biomechanics, 2E*. Human Kinetics.

[44] M. S. Ryoo and J. K. Aggarwal. 2009. Spatio-Temporal Relationship Match: Video Structure Comparison for Recognition of Complex Human Activities. In *Proc. IEEE ICCV*.

[45] Wataru Sato, Tomomi Fujimura, and Naoto Suzuki. 2008. Enhanced facial EMG activity in response to dynamic facial expressions. *International Journal of Psychophysiology* 70, 1 (2008), 70–74.

[46] Bernhard Scholkopf and Alexander J Smola. 2001. *Learning with kernels: support vector machines, regularization, optimization, and beyond*. MIT press.

[47] Paul Scovanner, Saad Ali, and Mubarak Shah. 2007. A 3-dimensional sift descriptor and its application to action recognition. In *Proc. ACM Multimedia*. 357–360.

[48] Amir Shahroudy, Jun Liu, Tian-Tsong Ng, and Gang Wang. 2016. NTU RGB+ D: A large scale dataset for 3D human activity analysis. *arXiv:1604.02808* (2016).

[49] Jan Smisek, Michal Jancosek, and Tomas Pajdla. 2013. 3D with Kinect. In *Consumer depth cameras for computer vision*. Springer, 3–25.

[50] Xinhang Song, Chengpeng Chen, and Shuqiang Jiang. 2017. RGB-D Scene Recognition with Object-to-Object Relation. In *Proc. ACM Multimedia*. 600–608.

[51] Tae Soo Kim and Austin Reiter. 2017. Interpretable 3D Human Action Analysis With Temporal Convolutional Networks. In *Proc. IEEE CVPR*. 20–28.

[52] Khurram Soomro, Amir Roshan Zamir, and Mubarak Shah. 2012. UCF101: A dataset of 101 human actions classes from videos in the wild. *arXiv preprint arXiv:1212.0402* (2012).

[53] Jaeyong Sung, Colin Ponce, Bart Selman, and Ashutosh Saxena. 2011. Human activity detection from RGBD images. In *Proc. AAAI Workshops*.

[54] Gul Varol, Ivan Laptev, and Cordelia Schmid. 2017. Long-term temporal convolutions for action recognition. *IEEE Trans. PAMI* (2017).

[55] Raviteja Vemulapalli, Felipe Arrate, and Rama Chellappa. 2014. Human action recognition by representing 3d skeletons as points in a lie group. In *Proc. IEEE CVPR*. 588–595.

[56] ViconSystem. 2010. The Standard Vicon full-body model (Plug-in Gait) Marker Placement Scheme. (2010).

[57] Heng Wang, Alexander Kläser, Cordelia Schmid, and Cheng-Lin Liu. 2011. Action recognition by dense trajectories. In *Proc. IEEE CVPR*. 3169–3176.

[58] Hongsong Wang and Liang Wang. 2017. Modeling temporal dynamics and spatial configurations of actions using two-stream recurrent neural networks. In *Proc. IEEE CVPR*.

[59] Jiang Wang, Zicheng Liu, Ying Wu, and Junsong Yuan. 2012. Mining actionlet ensemble for action recognition with depth cameras. In *Proc. IEEE CVPR*.

[60] Jiang Wang, Xiaohan Nie, Yin Xia, Ying Wu, and Song-Chun Zhu. 2014. Cross-view action modeling, learning and recognition. In *Proc. IEEE CVPR*. 2649–2656.

[61] Keze Wang, Xiaolong Wang, Liang Lin, Meng Wang, and Wangmeng Zuo. 2014. 3D human activity recognition with reconfigurable convolutional neural networks. In *Proc. ACM Multimedia*. ACM, 97–106.
[62] Lichen Wang, Zhengming Ding, and Yun Fu. 2018. Learning Transferable Subspace for Human Motion Segmentation. In *Proc. AAAI*.
[63] Limin Wang, Yuanjun Xiong, Zhe Wang, Yu Qiao, Dahua Lin, Xiaoou Tang, and Luc Van Gool. 2016. Temporal segment networks: Towards good practices for deep action recognition. In *Proc. ECCV*. Springer, 20–36.
[64] Pichao Wang, Wanqing Li, Zhimin Gao, Jing Zhang, Chang Tang, and Philip O Ogunbona. 2016. Action recognition from depth maps using deep convolutional neural networks. *IEEE Trans. Human-Machine Systems* 46, 4 (2016), 498–509.
[65] Pichao Wang, Zhaoyang Li, Yonghong Hou, and Wanqing Li. 2016. Action recognition based on joint trajectory maps using convolutional neural networks. In *Proc. ACM Multimedia*. 102–106.
[66] Ping Wei, Yibiao Zhao, Nanning Zheng, and Song-Chun Zhu. 2013. Modeling 4d human-object interactions for event and object recognition. In *Proc. IEEE ICCV*. 3272–3279.
[67] Lu Xia, Chia-Chih Chen, and JK Aggarwal. 2012. View invariant human action recognition using histograms of 3d joints. In *Proc. IEEE CVPR*. 20–27.
[68] Sijie Yan, Yuanjun Xiong, and Dahua Lin. 2018. Spatial Temporal Graph Convolutional Networks for Skeleton-Based Action Recognition. *arXiv preprint arXiv:1801.07455* (2018).
[69] Gang Yu, Zicheng Liu, and Junsong Yuan. 2014. Discriminative orderlet mining for real-time recognition of human-object interaction. In *Proc. ACCV*. Springer, 50–65.
[70] Jing Zhang, Wanqing Li, Philip O Ogunbona, Pichao Wang, and Chang Tang. 2016. RGB-D-based action recognition datasets: A survey. *Pattern Recognition* (2016), 86–105.
[71] Zhengyou Zhang. 2012. Microsoft kinect sensor and its effect. *IEEE Multimedia* 19, 2 (2012), 4–10.