Berkeley MHAD: A Comprehensive Multimodal Human Action Database

Ferda Ofli[1], Rizwan Chaudhry[2], Gregorij Kurillo[1], René Vidal[2] and Ruzena Bajcsy[1]
[1]Tele-immersion Lab, University of California, Berkeley
[2]Center for Imaging Sciences, Johns Hopkins University

Abstract

Over the years, a large number of methods have been proposed to analyze human pose and motion information from images, videos, and recently from depth data. Most methods, however, have been evaluated on datasets that were too specific to each application, limited to a particular modality, and more importantly, captured under unknown conditions. To address these issues, we introduce the Berkeley Multimodal Human Action Database (MHAD) consisting of temporally synchronized and geometrically calibrated data from an optical motion capture system, multi-baseline stereo cameras from multiple views, depth sensors, accelerometers and microphones. This controlled multimodal dataset provides researchers an inclusive testbed to develop and benchmark new algorithms across multiple modalities under known capture conditions in various research domains. To demonstrate possible use of MHAD for action recognition, we compare results using the popular Bag-of-Words algorithm adapted to each modality independently with the results of various combinations of modalities using the Multiple Kernel Learning. Our comparative results show that multimodal analysis of human motion yields better action recognition rates than unimodal analysis.

1. Introduction

Human pose estimation, motion tracking and action recognition have been studied extensively by the computer vision community over the last few decades. Mouslend *et al*. [17, 18] provide a functional taxonomy based review of the human motion capture and activity analysis. Similarly Aggarwal and Ryoo [3] published a more recent review based on an approach-specific taxonomy. Most of the studies in this domain have focused primarily on image and video data obtained from various collections on the Internet (*e.g*., YouTube, Flickr). Although such collections provide practically endless source of data, the images and videos usually have to be manually labeled, segmented or otherwise preprocessed to obtain the ground truth. Users of such datasets have very little information on the hardware and conditions used to capture particular images or videos, which can bias the recognition algorithms. Some of the issues pertaining to popular image datasets used by the computer vision community were previously discussed in [23]. More recently, Torralba and Efros analyzed several examples of popular object recognition databases and showed that uncontrolled data collection (*i.e*., "in the wild") can create biased results which limit the progress of algorithms developed and evaluated against such datasets [28]. These studies suggest that the experiments should be performed first on carefully controlled datasets in order to understand the limitations of a new algorithm and to compare it with existing algorithms.

A large body of literature in human motion analysis and action recognition has been dedicated to motion capture (mocap) data, where sparse movement information is extracted either from a small number of active or passive markers detected by an optical system or from direct motion measurements obtained by an inertial or mechanical motion system. Mocap data, which is frequently used for animation and video games, has a much lower spatial resolution but higher temporal resolution as compared to video data. The high temporal resolution is especially important when studying dynamics of human motion. The spatial segmentation and labeling of mocap data with respect to corresponding body segments is usually trivial; however, as in video, the automatic temporal segmentation of longer sequences into low-level primitives is still an open problem [4, 19]. In comparison to the various video collections, the availability of mocap data is significantly limited due to high costs of the acquisition systems and tedious preparations needed to collect such data.

Recently, there has been a revival of interest in 3D human motion tracking, pose estimation and action recognition, thanks to the advances in active depth sensing, especially with the launch of the Microsoft Kinect. This technological development has started a considerable trend in 3D data collection and analysis, causing a shift from mocap-video clip suite datasets to RGB+D datasets (*i.e*., including video/image and depth data). The low cost and simplicity of capturing with the Kinect are creating an opportunity for average users to create Flickr/YouTube type of datasets [1].

Similarly as in video data collections, such datasets are by and large uncontrolled and very little information is given on the conditions under which the data were collected.

There exist only a handful of datasets that include more than two modalities, as outlined in the next section, while there is a great need in the multimedia community to develop and evaluate multimodal algorithms, such as algorithms for action recognition, pose estimation, motion segmentation and dynamic 3D scene reconstruction. For this purpose, we created the Berkeley Multimodal Human Action Database (MHAD) consisting of temporally synchronized and geometrically calibrated data from an optical mocap system, multi-baseline stereo cameras from multiple views, depth sensors, accelerometers and microphones. We believe this database will provide a comprehensive testbed for new algorithms in the aforementioned research domains.

In the next section, we review several existing databases that have been used by the computer vision community in the past decade. We outline several disadvantages of these datasets which motivated us to create the proposed multimodal human action database. We describe our database in more details in Section 3. To demonstrate the advantage of multimodal data for action recognition, we report several experiments based on the Bag-of-Words algorithm in Section 4. Finally, we conclude the paper in Section 5.

2. Existing Databases

One of the first video-based human action datasets were KTH by Schuldt *et al.* [25] and Weizmann by Blank *et al.* [5]. Both datasets were captured in a relatively controlled setting with minimal occlusion and clutter in the scene. KTH and Weizmann datasets have both been used extensively by the computer vision community in the past several years, reaching the limits for any further improvements for future research on these particular datasets [33].

To provide the community with new challenges for human action recognition, several datasets with richer sets of activities and more complex environments, more faithfully representing the real-world scenarios, have been published. These include collections of video clips from various television shows, movies and user contributed videos, such as YouTube. These datasets include Hollywood2 by Marszalek *et al.* [16], UCF50 by Liu *et al.* (as an extension to UCF YouTube [15]), Olympic Sports by Niebles *et al.* [22] and HMDB51 by Kuehne *et al.* [11]. Despite the fact that the majority of the computer vision community focused on monocular video datasets, a number of research groups invested time and effort in creating multiview video datasets, such as the CMU Motion of Body (MoBo) Database by Gross and Shi [10], IXMAS by Weinland *et al.* [32], ViHASi: Virtual Human Action Silhouette Data by Ragheb *et al.* [24] and i3DPost Multi-view and 3D Human Action/Interaction Database by Gkalelis *et al.* [9]. These datasets not only increased the amount of information available for analysis, but also introduced baselines, either in 2D or in 3D, for quantitative evaluation of motion tracking and pose estimation algorithms.

Database	Modality					# Sub	# Act	# Seq
	V	M	D	A	O			
KTH [25]	1					25	6	2391
Weizmann [5]	1					9	10	90
Hollywood2 [16]	1					-	12	2517
UCF50 [15]	1						50	>5000
Olympic Sports [22]	1					-	16	800
HMDB51 [11]	1					-	51	>5151
CMU MoBo [10]	6					25	1	100
IXMAS [32]	5					11	13	1148
ViHASi [24]	<40					9	20	≈180
i3DPost [9]	8					8	12	104
CMU Mocap [6]	1	*				>100	109	2605
HDM05 [20]		*				5	>70	≈1500
HumanEva I [26]	7	*				4	6	56
HumanEva II [26]	4	*				2	1	2
TUM Kitchen [27]	4	*			*	4	4	17
CMU MMAC [7]	6	*		5	*	43	5	185
MSR-Action3D [14]			1			10	20	567
MSRDailyActivity3D [31]	1		1			10	16	960
MHAD (Our)	12	*	2	4	*	12	11	>647

Table 1. Comparison of MHAD to other datasets available in the computer vision literature. Sensor modalities: (V)ideo, (M)ocap, (D)epth, (A)udio, (O)thers.

In parallel to the efforts of producing video-based human action datasets, several mocap datasets were published, such as the CMU Motion Capture Dataset by the CMU Graphics Lab [6] and the HDM05 Mocap Database by Müller *et al.* [20]. The CMU Motion Capture Dataset is the first and most extensive publicly available mocap dataset, which contains various actions ranging from locomotion to sports and pantomime. The CMU dataset includes monocular video in addition to the mocap data for most of the sequences; however, the video camera is of low quality and not calibrated with respect to the mocap system, thus not allowing for coupling of the video and mocap data that would be required, for example, for the evaluation of video-based tracking algorithms. Furthermore, the dataset is not well structured as it does not provide a balanced distribution of the motion sequences across different action categories and subjects, which makes it less useful for classification tasks. The HDM05 Mocap Database, on the other hand, provides several hours of pure mocap data that consists of systematically recorded set of motion sequences, containing multiple realizations of action categories for several subjects. Nevertheless, this dataset lacks other supporting media, such as video, that can be used in conjunction with the mocap data, either for reference or to evaluate multimodal algorithms.

In addition to the video- and mocap-centric datasets, there are several datasets that combine video and mocap systems in a more systematic way by collecting synchronized and calibrated data. Such examples include the HumanEva I and II datasets by Sigal *et al.* [26]. The creation of HumanEva datasets were motivated mainly by the need for having ground truth that can be used for quantitative evaluation and comparison of both 2D and 3D pose estimation and tracking algorithms. Although the HumanEva datasets have been extensively used in establishing the state-of-the-art in human action recognition, their application areas remain limited to evaluation of 2D and 3D motion and pose estimation based on video and mocap data only.

There are a number of other multimodal datasets that enhance the standard mocap-video data with additional modalities, such as magnetic sensors or microphones. The TUM Kitchen Dataset by Tenorth *et al*. [27], which consists of activities in a kitchen setting (*i.e*., subjects setting a table in different ways), for example includes also RFID tag and magnetic sensor readings in addition to the multi-view video and mocap data. Similarly, the CMU Multimodal Activity (CMU-MMAC) Dataset by De La Torre *et al*. [7] contains multimodal measures captured from subjects performing tasks such as meal preparation and cooking. The set of modalities utilized in this dataset is rather comprehensive, consisting of video, audio, mocap, internal measurement units (*i.e*., accelerometers, gyroscopes and magnetometers) and wearable devices (*i.e*., BodyMedia and eWatch). These two datasets are the first examples of publicly available multimodal datasets with a rich selection of various modalities.

In the past, datasets based purely on depth data were used primarily for object and scene recognition, segmentation and rendering purposes by the computer vision and graphics communities. However, with the advent of the Microsoft Kinect, new 3D depth datasets have emerged for human motion tracking, pose estimation and action recognition, such as MSR-Action3D Dataset by Li *et al*. [14] and MSRDailyActivity3D Dataset by Wang *et al*. [31]. These datasets provide a rich depth representation of the scene at each time instant, allowing for both spatial and temporal analyses of human motion. However, they are captured from a single viewpoint, and hence, prone to occlusions in the scene, resulting in inaccurate pose estimations. The two datasets also do not include any accurate ground truth information for extracted pose.

Table 1 provides comparative summary of several existing datasets. By introducing a new multimodal dataset we are addressing several drawbacks of the aforementioned datasets, such as the lack of multimodal synchronized data, balanced distribution of actions and their repetitions, performance of same actions by multiple subjects. Our dataset therefore consists of multi-view video, depth and color data from multiple Kinect cameras, movement dynamics from wearable accelerometers and the accurate mocap data with the skeleton information. In addition, we also recorded and synchronized ambient sound during the action performance to reveal discriminative cues for human motion analysis. Our goal is to provide this database to the computer vision and other research communities to further advance the development and evaluation of various algorithms, in particular the algorithms that take advantage of the multimodal input. To the best of our knowledge, this is the only database to-date that systematically combines multiple depth cameras with multi-view video and mocap that are geometrically calibrated and temporally synchronized with other modalities such as accelerometry and sound.

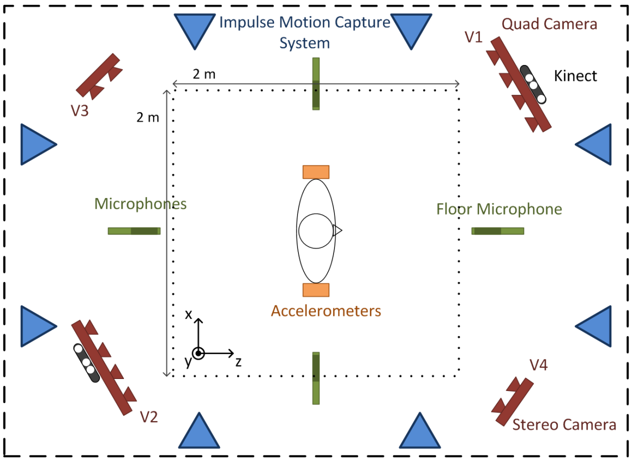


Figure 1. Diagram of the data acquisition system.

3. Multimodal Human Action Database

3.1. Database Acquisition

In this section we describe the components of the multimodal acquisition system used for the collection of our database. Each action was simultaneously captured by five different systems: optical mocap system, four multi-view stereo vision cameras, two Microsoft Kinect cameras, six wireless accelerometers and four microphones. Figure 1 shows the layout of the sensors used in our setup.

Mocap System: For the ground truth we used optical motion capture system Impulse (PhaseSpace Inc., San Leandro, CA, USA) which captured 3D position of active LED markers with the frequency of 480 Hz. The mocap system was calibrated using manufacturer software. The data acquisition server provided the time synchronization for other sensors (i.e., cameras, Kinect, accelerometers) through the NTP using Meinberg NTP client service[1]. To capture the position of different parts of the body via the mocap system, we used a custom-built tight-fitting suit with 43 LED markers. We post-processed the 3D marker trajectories to extract the skeleton of each subject using MotionBuilder software (Autodesk Inc., San Rafael, CA, USA).

Camera System: Multi-view video data was captured by 12 Dragonfly2 cameras (Point Grey Research Inc., Richmond, BC, Canada) with the image resolution of 640×480 pixels. The cameras were arranged into four clusters: two clusters for stereo and two clusters with four cameras for multi-view capture, deployed as shown in Figure 1. The focal length was set at approximately 4 mm to provide full

[1] http://www.meinberg.de/english/sw/ntp.htm

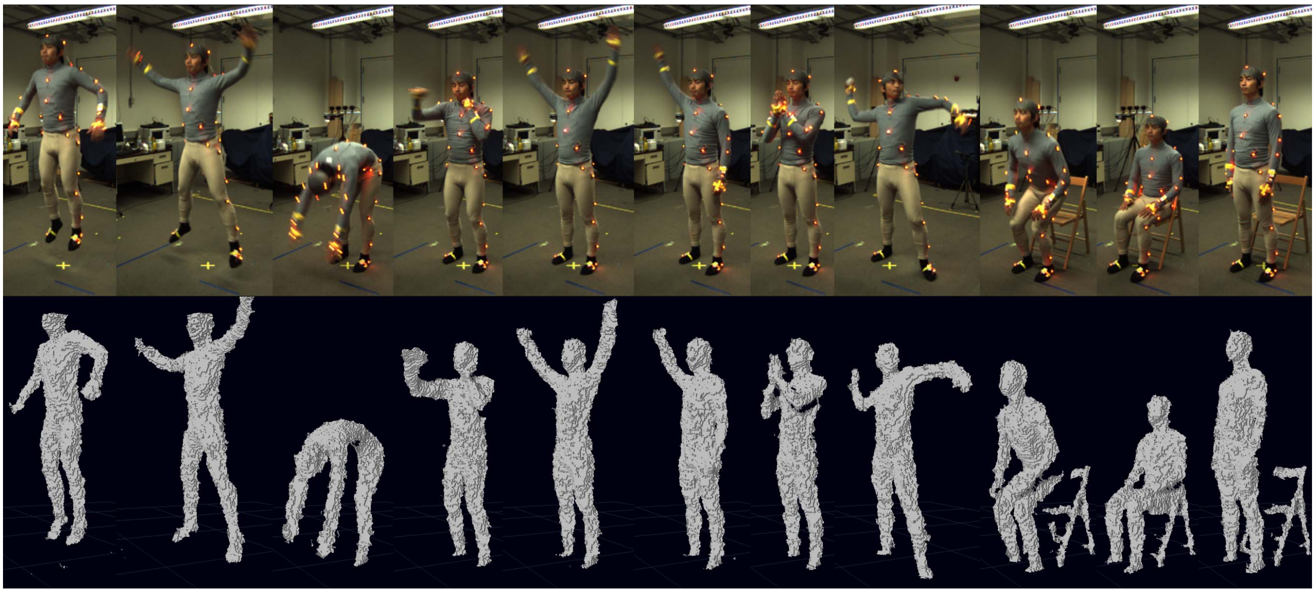


Figure 2. Snapshots from all the actions available in the database are displayed together with the corresponding point clouds obtained from the Kinect depth data. Actions (from left to right): *jumping*, *jumping jacks*, *bending*, *punching*, *waving two hands*, *waving one hand*, *clapping*, *throwing*, *sit down/stand up*, *sit down*, *stand up*.

view of the subject while performing various actions. The images were captured with a frame rate of about 22 Hz using hardware triggering for temporal synchronization across all views. Prior to data collection, the cameras were geometrically calibrated and aligned with the mocap system.

Kinect System: For the depth data acquisition we have positioned two Microsoft Kinect cameras approximately in opposite directions to prevent interference between the two active pattern projections. Each Kinect camera captured a color image with a resolution of 640×480 pixels and a 16-bit depth image, both with an acquisition rate of 30 Hz. Although the color and depth acquisition inside the Kinect are not precisely synchronized, the temporal difference is not noticeable in the output due to the relatively high frame rate. The Kinect cameras were temporally synchronized with the NTP server and geometrically calibrated with the mocap system. We used the OpenNI[2] driver for the acquisition of the images since the official Microsoft Kinect driver was not yet available at the time of the data collection.

Accelerometers: To capture dynamics of the movement, we have applied six three-axis wireless accelerometers (Shimmer Research, Dublin, Ireland). The accelerometers were strapped or inserted into the mocap suit to measure movement at the wrists, ankles and hips. The accelerometers captured the motion data with the frequency of about 30 Hz and delivered data via the Bluetooth protocol to the acquisition computer where time stamps were applied to each collected frame. Due to the delays in the wireless communication, there was a lag of about 100 ms between the acquisition and recording times, which was compensated for by the time-stamp adjustments.

Audio System: The importance of audio in video analysis was demonstrated in past research, e.g., the work by Abdullah and Noah [2] who improved human action detection performance by integrating audio with visual information. Therefore, we decided to also record audio during performance of each action using four microphones arranged around the capture space as shown in Figure 1. Three of the microphones were placed on tripods about 65 cm off the ground while the fourth microphone was taped to the floor to capture the sounds generated at the ground surface. The audio recording was performed with the frequency of 48 kHz through analog/digital converter/amplifier which was connected via an optical link to a digital recorder.

3.2. Database Description

Our database contains 11 actions performed by 7 male and 5 female subjects in the range 23-30 years of age except for one elderly subject. All the subjects performed 5 repetitions of each action, yielding about 660 action sequences which correspond to about 82 minutes of total recording time. In addition, we have recorded a T-pose for each subject which can be used for the skeleton extraction; and the background data (with and without the chair used in some of the activities). Figure 2 shows the snapshots from all the actions taken by the front-facing camera and the corresponding point clouds extracted from the Kinect data. The specified set of actions comprises of the following: (1) actions with movement in both upper and lower extremities, *e.g.*, *jumping in place*, *jumping jacks*, *throwing*, etc., (2) ac-

[2] http://www.openni.org/documentation

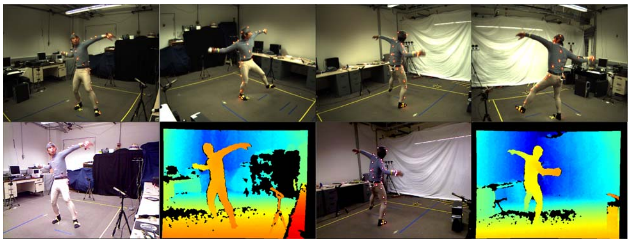


Figure 3. *Throwing* action is displayed from the reference camera in each camera cluster as well as from the two Kinect cameras.

tions with high dynamics in upper extremities, *e.g.*, *waving hands*, *clapping hands*, etc. and (3) actions with high dynamics in lower extremities, *e.g.*, *sit down*, *stand up*.

Prior to each recording, the subjects were given instructions on what action to perform; however no specific details were given on how the action should be executed (*i.e.*, performance style or speed). The subjects have thus incorporated different styles in performing some of the actions (*e.g.*, *punching*, *throwing*). Figure 3 shows a snapshot of the *throwing* action from the reference camera of each camera cluster and from the two Kinect cameras. The figure demonstrates the amount of information that can be obtained from multi-view and depth observations as compared to a single viewpoint. In Figure 4 we compare the outputs of the mocap, accelerometer and audio data for the *jumping* action. The trajectories from all three modalities show distinct patterns typical for this action. The database is available online at *http://tele-immersion.citris-uc.org/berkeley_mhad*.

4. Action Recognition Experiments

To demonstrate the use of various modalities included in our database, we perform action recognition experiments based on the popular Bag-of-Words approach to model the action sequence from a particular modality. Briefly, for each modality, we extract features from the entire sequence and quantize them using k-medoids. We choose k-medoids over the more common k-means, as it is computationally more efficient and in general does not show significant difference in classification performance. We then compute the distance of each feature to these codewords to assign a codeword label. Each training and test sequence is then represented as a histogram of codewords. We can therefore compute the (dis)similarity between two sequences by computing the χ^2 distance between the histograms representing the two sequences and use the k-Nearest Neighbors (k-NN) classifier to classify the action of each sequence. We also use kernel-SVM with the χ^2 kernel for classification. We then use Multiple Kernel Learning (MKL) [8] to compute the optimal linear combination of multimodal/multi-view similarities for the task of action recognition.[3] In all experiments, we use first 7 subjects for training and last 5 subjects for testing. In the subsequent sections, we first describe the feature extraction process from each modality, and then show results for action classification using that modality.

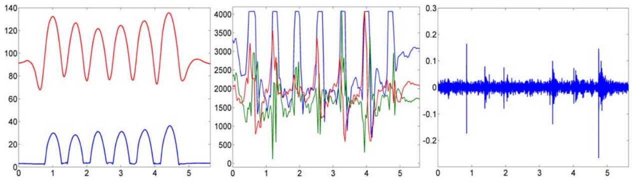


Figure 4. Comparison of mocap, accelerometer and audio data for *jumping* action. On the left, only the vertical trajectories of the left wrist (red) and the right foot (blue) are plotted. The plot in the middle shows acceleration measurements in all three axes for the right wrist whereas the plot on the right shows audio waveform from the ground level microphone. It is notable that the peak locations are well aligned in all three modalities.

	V-1	V-2	V-3	V-4	MKL
1-NN	91.61	91.97	93.43	84.67	
3-NN	90.88	90.51	92.43	84.31	
K-SVM	91.97	87.96	96.35	87.23	99.27

Table 2. Action classification results using video data from different camera views (V-1 through V-4).

4.1. Multi-view Video Data

State-of-the-art action recognition methods compute spatio-temporal interest points from videos and extract features from a spatio-temporal window around each interest point. Wang *et al.* [30] performed a comparative study of several spatio-temporal features across several databases and showed that the Space-Time-Interest-Points (STIP) detection method of Laptev *et al.* [12] and Histogram of Gradients (HOG) and Histogram of Flow (HOF) features proposed in [13] tend to perform the best. Therefore, we extract HOG/HOF features from each video as in [13]. Using the aforementioned training/test breakup, we quantize the training HOG/HOF features into 20 codewords and perform the classification experiments. Table 2 shows the classification results for each view for k-NN with different values of k as well as kernel-SVM. The overall classification rate is high because the video data has a static background with little noise and the subjects remain around the same spatial location in the scene throughout the duration of motion, in which case the spatio-temporal features perform very well. Furthermore, combination of all views via MKL yields almost perfect action classification at 99.27%.

4.2. Kinect Depth Data

Feature extraction from depth videos for action recognition is, for the most part, an open problem. In this study, we follow the approach of Ni *et al.* [21] that extends the spatio-temporal features of Laptev *et al.* [13] to the depth domain. We first divide the depth videos into $m = 8$ disjoint Depth-Layered Multi-Channel (DLMC) videos. This

[3] We used the multi-class MKL implementation of Vedaldi *et al.* [29]

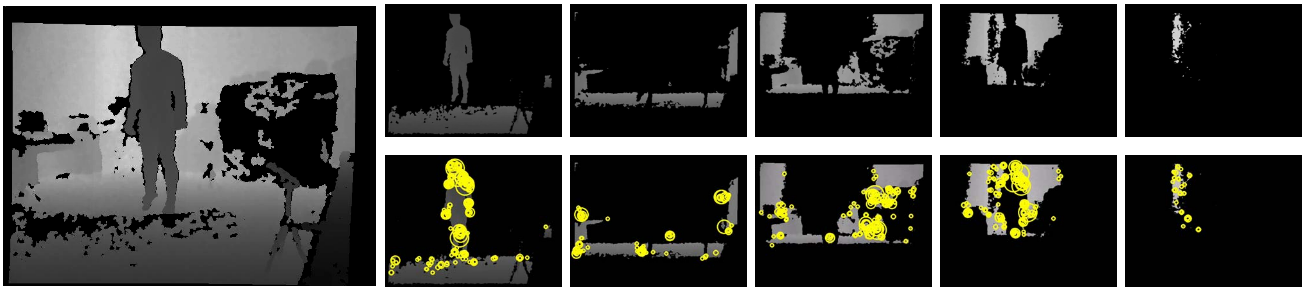


Figure 5. Left: Grayscale depth image for a person performing the jump action. Right - Top row: Depth-Layered Multi-Channel (DLMC) images for channels 3 to 7. Channels 1, 2 and 8 did not have significant depth data. Right - Bottom row: Spatio-temporal interest points (STIP) detected in each DLMC. The STIP show up mostly at depth discontinuities around the person, noisy areas as well as depth discontinuities in the background due to occlusion by the motion of the subject.

	C-3	C-4	C-5	C-6	C-7	MKL
1-NN	77.37	35.40	28.47	35.77	58.03	
3-NN	76.28	30.29	19.34	29.93	55.84	
K-SVM	70.07	24.45	24.09	27.01	40.15	91.24

Table 3. Action classification results using the Kinect depth-layered multi-channel data (for channels C-3 through C-7).

	MC	Acc	Aud
1-NN	74.82	79.20	32.12
3-NN	75.55	81.75	28.10
K-SVM	79.93	85.40	35.04

Table 4. Action classification results using mocap (MC), accelerometer (Acc) and audio (Aud) data.

is done by dividing the entire depth range into m equal depth layers and, for each depth layer, keeping the pixels that fall into the corresponding depth layer. We then extract the usual HOG/HOF features from each DLMC videos, and separately quantize the features from the training data extracted from each layer. Figure 5 shows a sample Kinect frame and the corresponding DLMC frames as well as the detected feature points for channels $m = 3 - 7$ (*i.e.*, C-3 through C-7). Channels $m = 1, 2, 8$ (*i.e.*, C-1, C-2 and C-8) did not have significant depth data and therefore did not have any features. After computing the codewords, we compute m histograms from each depth sequence, each corresponding to a particular range of depths. Table 3 shows the results of action recognition using the Kinect depth data. Note that C-3 gives the best results because the subjects almost always fall into the depth layer at $m = 3$. C-7 also performs slightly better than the rest of the channels due to the characteristic depth discontinuities in the background created by the occlusions from subject's motions.

4.3. Mocap Data

Based on the skeletal poses extracted from the mocap data, we compute the joint angles for a set of 21 joints in the skeleton hierarchy which allows us to represent an action sequence as a collection of 21-dimensional joint angle features. For the action recognition task, we partition each action sequence into $N_s = 60$ temporal windows and compute the variance of the joint angles in that temporal window as our local temporal feature descriptor. As a result, we represent each action sequence by a set of N_s feature descriptors of size 21. We then quantize the training features into 20 codewords and perform the corresponding classification experiment. Table 4 shows the classification results for the mocap data under the column labeled "MC."

4.4. Accelerometer Data

Each of the six accelerometers provides a time series of 3-dimensional acceleration values. As with the skeleton data, we extract $N_s = 30$ temporal windows from each accelerometer sequence and compute the variance of the acceleration in each direction in each temporal window. Concatenating variance values from all accelerometers, we get an 18-dimensional local temporal feature descriptor per temporal window. Hence, we represent an action sequence with a set of N_s feature descriptors of size 18. We then quantize the features into 20 codewords and perform classification. Table 4 shows under the column labeled "Acc" that unimodal classification results for the accelerometer data are promising despite that the extracted features are simple.

4.5. Audio Data

For audio analysis, we compute the 13-dimensional time-series of MFCC coefficients from each audio sequence and, following the procedure for other modalities, divide each sequence into $N_s = 60$ temporal segments. We then compute the mean MFCC coefficients for each temporal segment and quantize all features into 20 codewords for classification. Table 4 shows the classification results for the audio data under the column labeled "Aud." Not surprisingly, action recognition using audio data does not provide good results by itself. However, the results are sufficiently better than random guess (9%) and should be helpful for action recognition when combined with other modalities. Note that MFCC features are commonly used for speech modeling and may not be the best feature representation for action recognition from audio data.

4.6. Multimodal Results

Finally, we learn several classifiers by combining various modalities through Multiple Kernel Learning (MKL). Table 5 shows action recognition results for several different combinations of modalities. As expected, combining several modalities increases the recognition performance because, in general, the features extracted from one modality complement the drawbacks of the features extracted from other modalities. For instance, combining mocap (MC) and accelerometer (Acc) features alone, the recognition performance jumps to 97.45% from 79.93% for the mocap and 85.40% for the accelerometer data as shown before in Table 4. Furthermore, combining audio (Aud) with other modalities does not necessarily reduce the recognition rate as MKL learns the best linear combination of feature kernels. In fact, combining audio with mocap, kinect (Kin) and accelerometer data increases the recognition rate from 98.18% (MC+Acc+Kin) to 98.54% (MC+Acc+Kin+Aud). Figure 6 shows the weights computed by MKL for different modalities for each action in a one-vs-all classification framework. In particular, Figure 6(c) illustrates the importance of audio features in discriminating *clapping*, *throwing* and *stand up* actions from others. Finally, combining all 5 modalities, we attain 100% action recognition performance, thanks to contribution from all but Kinect-based features as shown in Figure 6(d).

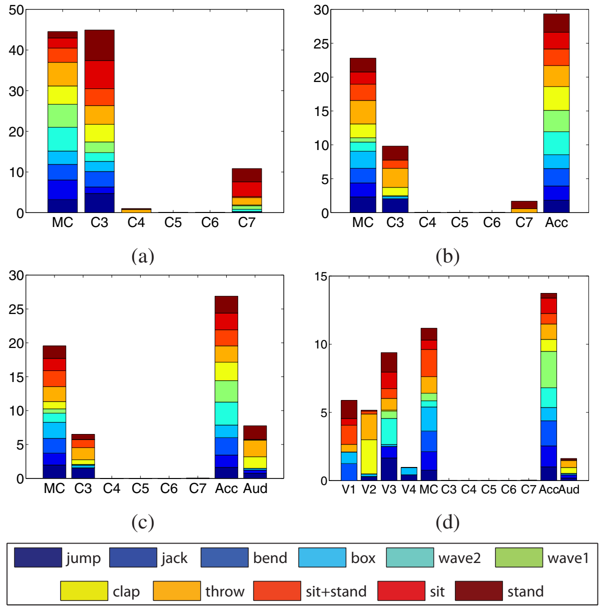
5. Discussion and Conclusion

We explain the high recognition accuracy by the nature of the controlled data collection, especially when complemented with multiple modalities. However, with carefully designed experimental dataset, as the one presented, one can get better insight into the workings of the recognition algorithms and feature selection, and explain better the failures. Therefore, we anticipate that the data will be useful to learn and tune new classifiers on modalities, such as video or accelerometry, for recognition on the data "in the wild," which will most likely include only one or two modalities. The data can be used as ground truth for robustness analyses where noise can be introduced into the data on different levels, e.g., reduce resolution of images, reduce temporal resolution of mocap data, add noise to mocap markers, etc.

We have presented a unique multimodal human action database which is, to the best of our knowledge, currently the largest database with more than 80 minutes of data captured from 12 subjects, performing 11 actions in 5 trials, by combining various modalities such as mocap, multi-view video, depth, acceleration and audio. All data captured by different sensors are geometrically calibrated and synchronized for temporal alignment. Our action recognition experiments based on the state-of-the-art Bag-of-Words algorithm together with the Multiple Kernel Learning demonstrated that the multimodal analysis of human motion performs superior than the unimodal analysis. We believe this multimodal dataset has broader applications, beyond action recognition, such as human motion tracking, pose estimation, motion segmentation and 3D reconstruction.

Modalities	Rec (%)
MC + Acc	97.45
Kin + Vid	99.27
MC + Kin	93.80
Acc + Kin	97.81
Vid + Aud	99.27
MC + Acc + Kin	98.18
MC + Acc + Aud	97.45
Vid + Kin + Aud	98.91
MC + Acc + Kin + Aud	98.54
MC + Acc + Kin + Vid	100.00
MC + Acc + Kin + Vid + Aud	100.00

Table 5. Integrating multiple modalities using MKL.



(a)

(b)

(c)

(d)

Figure 6. MKL weights for 1-vs-all classification for all actions. Left to right: MC+Kin, MC+Kin+Acc, MC+Kin+Acc+Aud, all modalities.

Acknowledgements We would like to thank Po Yan and Lin Zhang for their contribution in collection and post-processing of the database. This work was funded in part by grants NSF 0941362, NSF 0941463 and NSF 0941382.

References

[1] Kinect@Home. url: http://www.kinectathome.com. Center for Autonomous Systems, Royal Institute of Technology, Sweden. Accessed: April 11, 2012.

[2] L. Abdullah and S. Noah. Integrating audio visual data for human action detection. In *Computer Graphics, Imaging and Visualisation (CGIV), 2008 5th Int. Conf. on*, pages 242–246, Aug. 2008.

[3] J. Aggarwal and M. Ryoo. Human activity analysis: A review. *ACM Comput. Surv.*, 43:16:1–16:43, Apr. 2011.

[4] J. Barbic, J.-Y. Pan, C. Faloutsos, J. K. Hodgins, and N. Pollard. Segmenting motion capture data into distinct behaviors. In *Graphics Interface 2004*, pages 185–194, May 2004.

[5] M. Blank, L. Gorelick, E. Shechtman, M. Irani, and R. Basri. Actions as space-time shapes. In *Computer Vision (ICCV), 2005. IEEE Int. Conf. on*, volume 2, pages 1395–1402, Oct. 2005.

[6] CMU. CMU motion capture database. url: http://mocap.cs.cmu.edu/, 2003.

[7] F. De La Torre, J. Hodgins, J. Montano, S. Valcarcel, R. Forcada, and J. Macey. Guide to the Carnegie Mellon University Multimodal Activity (CMU-MMAC) database. Technical Report CMU-RI-TR-08-22, Robotics Intstitute, Carnegie Mellon University, Pittsburgh, PA, July 2009.

[8] P. Gehler and S. Nowozin. On feature combination for multiclass object classification. In *Computer Vision (ICCV), 2009. IEEE Int. Conf. on*, pages 221–228, Oct. 2009.

[9] N. Gkalelis, H. Kim, A. Hilton, N. Nikolaidis, and I. Pitas. The i3DPost multi-view and 3D human action/interaction database. In *2009 Conf. for Visual Media Production*, CVMP'09, pages 159–168, 2009.

[10] R. Gross and J. Shi. The CMU motion of body (MoBo) database. Technical Report CMU-RI-TR-01-18, Robotics Institute, Carnegie Mellon University, Pittsburgh, PA, June 2001.

[11] H. Kuehne, H. Jhuang, E. Garrote, T. Poggio, and T. Serre. HMDB: A large video database for human motion recognition. In *Computer Vision (ICCV), 2011. IEEE Int. Conf. on*, pages 2556–2563, Nov. 2011.

[12] I. Laptev and T. Lindeberg. Space-time interest points. In *Computer Vision (ICCV), 2003. IEEE Int. Conf. on*, pages 432–439, Oct. 2003.

[13] I. Laptev, M. Marszalek, C. Schmid, and B. Rozenfeld. Learning realistic human actions from movies. In *Computer Vision and Pattern Recognition (CVPR), 2008. IEEE Conf. on*, pages 1–8, June 2008.

[14] W. Li, Z. Zhang, and Z. Liu. Action recognition based on a bag of 3D points. In *Computer Vision and Pattern Recognition Workshops (CVPRW), 2010. IEEE Computer Society Conf. on*, pages 9–14, June 2010.

[15] J. Liu, J. Luo, and M. Shah. Recognizing realistic actions from videos "in the wild". In *Computer Vision and Pattern Recognition (CVPR), 2009. IEEE Conf. on*, pages 1996–2003, June 2009.

[16] M. Marszalek, I. Laptev, and C. Schmid. Actions in context. In *Computer Vision and Pattern Recognition (CVPR), 2009. IEEE Conf. on*, pages 2929–2936, June 2009.

[17] T. B. Moeslund and E. Granum. A survey of computer vision-based human motion capture. *Computer Vision and Image Understanding*, 81:231–268, 2001.

[18] T. B. Moeslund, A. Hilton, and V. Krüger. A survey of advances in vision-based human motion capture and analysis. *Computer Vision and Image Understanding*, 104:90–126, 2006.

[19] M. Müller, A. Baak, and H.-P. Seidel. Efficient and robust annotation of motion capture data. In *ACM SIGGRAPH/Eurographics Symposium on Computer Animation (SCA)*, pages 17–26, Aug. 2009.

[20] M. Müller, T. Röder, M. Clausen, B. Eberhardt, B. Krüger, and A. Weber. Documentation mocap database HDM05. Technical Report CG-2007-2, Universität Bonn, June 2007.

[21] B. Ni, G. Wang, and P. Moulin. RGBD-HuDaAct: A color-depth video database for human daily activity recognition. In *Computer Vision Workshops (ICCVW), 2011. IEEE Int. Conf. on*, pages 1147–1153, Nov. 2011.

[22] J. C. Niebles, C.-W. Chen, and L. Fei-Fei. Modeling temporal structure of decomposable motion segments for activity classification. In *11th European Conf. on Computer Vision: Part II*, ECCV'10, pages 392–405, 2010.

[23] J. Ponce, T. L. Berg, M. Everingham, D. A. Forsyth, M. Hebert, S. Lazebnik, M. Marszałek, C. Schmid, B. C. Russell, A. Torralba, C. K. I. Williams, J. Zhang, and A. Zisserman. Dataset issues in object recognition. In J. Ponce, M. Hebert, C. Schmid, and A. Zisserman, editors, *Toward Category-Level Object Recognition*, volume 4170 of *LNCS*, pages 29–48. Springer, 2006.

[24] H. Ragheb, S. Velastin, P. Remagnino, and T. Ellis. ViHASi: Virtual human action silhouette data for the performance evaluation of silhouette-based action recognition methods. In *Distributed Smart Cameras (ICDSC), 2008. 2nd ACM/IEEE Int. Conf. on*, pages 1–10, Sep. 2008.

[25] C. Schuldt, I. Laptev, and B. Caputo. Recognizing human actions: a local SVM approach. In *Pattern Recognition (ICPR), 2004. 17th Int. Conf. on*, volume 3, pages 32–36, Aug. 2004.

[26] L. Sigal, A. Balan, and M. Black. HumanEva: Synchronized video and motion capture dataset and baseline algorithm for evaluation of articulated human motion. *Int. Journal of Comput. Vis.*, 87:4–27, 2010.

[27] M. Tenorth, J. Bandouch, and M. Beetz. The TUM Kitchen Data Set of everyday manipulation activities for motion tracking and action recognition. In *Computer Vision Workshops (ICCVW), 2009. IEEE Int. Conf. on*, pages 1089–1096, Oct. 2009.

[28] A. Torralba and A. Efros. Unbiased look at dataset bias. In *Computer Vision and Pattern Recognition (CVPR), 2011. IEEE Conf. on*, pages 1521–1528, June 2011.

[29] A. Vedaldi, V. Gulshan, M. Varma, and A. Zisserman. Multiple kernels for object detection. In *Computer Vision (ICCV), 2009. IEEE Int. Conf. on*, pages 606–613, Oct. 2009.

[30] H. Wang, M. M. Ullah, A. Klaser, I. Laptev, and C. Schmid. Evaluation of local spatio-temporal features for action recognition. In *British Machine Vision Conf.*, pages 124.1–124.11. BMVA Press, 2009.

[31] J. Wang, Z. Liu, Y. Wu, and J. Yuan. Mining actionlet ensemble for action recognition with depth cameras. In *Computer Vision and Pattern Recognition (CVPR), 2012. IEEE Conf. on*, pages 2929–2936, June 2012.

[32] D. Weinland, E. Boyer, and R. Ronfard. Action recognition from arbitrary views using 3D exemplars. In *Computer Vision (ICCV), 2007. IEEE Int. Conf. on*, pages 1–7, Oct. 2007.

[33] D. Weinland, R. Ronfard, and E. Boyer. A survey of vision-based methods for action representation, segmentation and recognition. *Computer Vision and Image Understanding*, 115(2):224–241, 2011.