Labeling Subtle Conversational Interactions Within the CONVERSE Dataset

Michael Edwards, Jingjing Deng and Xianghua Xie
Department of Computer Science, Swansea University, Swansea, SA2 8PP, United Kingdom
http://csvision.swan.ac.uk

***Abstract*—The field of Human Action Recognition has expanded greatly in previous years, exploring actions and interactions between individuals via the use of appearance and depth based pose information. There are numerous datasets that display action classes composed of behaviors that are well defined by their key poses, such as 'kicking' and 'punching'. The CONVERSE dataset presents conversational interaction classes that show little explicit relation to the poses and gestures they exhibit. Such a complex and subtle set of interactions is a novel challenge to the Human Action Recognition community, and one that will push the cutting edge of the field in both machine learning and the understanding of human actions. CONVERSE contains recordings of two person interactions from 7 conversational scenarios, represented as sequences of human skeletal poses captured by the Kinect depth sensor. In this study we discuss a method providing ground truth labelling for the set, and the complexity that comes with defining such annotation. The CONVERSE dataset it made available online.**

I. Introduction

Recent methods in the detection and recognition of human actions have shown a range of domain applications in assisted living, surveillance, and computer generated animations. This has been facilitated by the rapid development of the Human Action Recognition (HAR) community in recent years, utilizing differing modalities and class complexities to recognize behavioral performances in a given scene. Both appearance and depth representations of observations are now regularly explored by the community, and depth based information has become increasingly more accessible due to the introduction of consumer level depth sensors and Motion Capture (MoCap) setups. Such depth information has resulted in kinematic tracking of the human pose without the need to estimate it from the 2D image. Despite this growth, there is still an imbalance in the representative datasets for each modality. Appearance based sets have started to mature to describe more complex interactions between individuals and day-to-day activities, however the depth based datasets still mostly concern themselves with describing interactions between individuals that are comprised of low level actions such as 'kick' and 'push'.

Appearance based sets, such as RGB videos, have been subjected to significant research in HAR; developing from capturing acted performances, to exhibiting more complex interactions between subjects and objects. Appearance modality HAR makes up a great number of publicly available datasets, displaying varying scenario problems from a wide range of domain applications. Many sets exhibit low level single action classes, such as 'walking' and 'jumping' [1], [2], [3]. Several present issues of occlusions and multi-view camera setups [4], [5], and some include a significant number of classes [6]. In recent years there has been significant effort to understand higher level interactions and daily activities [7], [8], [9]. Despite this, there are still often classes that are distinguishable in the image domain based on the poses of the subjects, such as 'pass object' or 'hug'. The plethora of appearance based sets in HAR have aided in the development and evaluation of countless methods in the detection and recognition of actions, including Space-Time Interest Point (STIP) [10], Scale Invariant Feature Transform (SIFT) [11] and temporal Harris corners [12]. Recently there has been work on using the image domain information within the deep learning Convolutional Neural Network (CNN) architecture to assist in HAR problems [13]. This study however, looks at the use of depth based pose information for the classification of observed actions.

The use of depth based information for HAR has seen an increase in interest since the release of the commercially available depth sensor, allowing a relatively cheap and effective method of obtaining human pose information from a scene by means of skeletal detection and tracking. The datasets available for depth based pose information, despite rapidly expanding, is still in its infancy. The majority of depth sets only display low level singular action observations; classes that are readily defined by their key poses, such as 'punch' and 'kick' [14], [15]. Recently there have been a few depth based interaction sets; however these still often reflect very primitive actions, where there is an active perpetrator of the action of interest [16], [17]. A few datasets explore the recognition of higher level activities, however these can often be very pose specific behaviors, such as 'setting a table' or 'using a vacuum cleaner' [18], [19].

The primitive classes provided by many of the current depth-based HAR datasets can often be condensed into a sequence of key poses and gestures. In reality however, it can often be that quite subtle interactions are composed of numerous small gestures and interactions over a long period of time. The CONVERSE dataset therefore introduces a problem that utilizes pose based information to represent subtle and complex interaction classes that are not readily definable by the poses they contain [20]. The conversational classes shown are common interactions in daily life; however they do not exhibit an explicit relationship to the pose of the individual at any given frame. This makes recognition from pose a difficult

task to solve, often requiring information from multiple frames to identify the interaction. Such a dataset has challenges in acquisition and definition of ground truth labels, and this paper will identify and present the methods used in the CONVERSE dataset. The dataset is publicly available, and can be obtained from [21]*.

In this paper we outline the CONVERSE dataset and its annotation, including the necessity for such a dataset. We present the use of depth based information for the Human Action Recognition problem, and the challenges that come with such a task. In Section II we provide details on the CONVERSE dataset. In Section III we discuss the method used to obtain ground truth labels for CONVERSE, including the use of appearance and audio information to label pose information, and the alignment of multi-modal recordings.

II. Proposed Dataset

The CONVERSE dataset is a collection of observations of two person interactions captured by the Kinect. The data represents the 3D movement of the joints of the full body human skeleton, as tracked by the built in skeletal tracker of the Kinect. The set includes seven conversational interactions scenarios, listed in Table I; ranging from telling a joke, to debating an topic. Baseline method performances on the CONVERSE set are outlined in [22]. More subject recordings have been obtained since the release of the original CONVERSE dataset, looking at expanding the significance of the results obtained from analysis performed on the dataset. Further recordings have expanded the number of observed interactions from 8 to 37, with a mixture of standing and seated positions for the participants. For this paper we discuss the original dataset currently available to the public.

A. Apparatus Setup

In the CONVERSE set, seven conversational action categories are captured using a two-Kinect setup to capture 3D pose during an interaction between two subjects. CONVERSE is recorded within an indoor lab space, (Figure 1). The interior space has a complex background and is not controlled, with natural lighting variation. Two Kinect sensors are positioned at opposite ends of the space, an approximate two meters away from a rough area the subject will stand. This position allowed the subjects to move freely within the observable area and still enabled the Kinect to capture the full body skeletal pose. Each subject was captured by their respective Kinect at the standard 30fps record rate of the sensor. To reduce the chance of subject occlusion, the Kinects were placed just to the front right of the subject. This allows an 'over Subject A's shoulder' view of Subject B. During the recordings the two subjects were free to move within the observable space. In order to obtain appearance information two PAL cameras (B cameras in Figure 1) were positioned to obtain a full body view of the subjects. A third camera (M in Figure 1) was used to monitor the experiment; these recordings allowed for sequence annotation, removed experimenter interference with recordings, and was used to ensure participant safety during the experiment. The appearance recordings are only used for annotation and conduction of the experiment. The appearance data is not provided with the dataset published in CONVERSE, as the set aims to provide a complex depth-based HAR problem.

*The current iteration of the CONVERSE set is available at http://csvision.swan.ac.uk/converse

TABLE I
Description of each of the tasks given to the participants to perform. The rightmost column describes whether the participants were told about the task and asked to prepare before attending.

Task Name	Description	Prepared in advance
Describing Work	Each participant describes their current work or project to partner. The partner then repeats the description back, to confirm they had understood.	Yes
Story Telling	Participant were asked to think of an interesting story they could tell their partner.	Yes
Problem Solving	Participants were given the problem "Do candles burn in space and if so what shape and direction?", and asked to think of the solution of together.	No
Debate	Participants prepared arguments for a given point of view, pro or con, on the topic "Should University education be free?", and then debated this between them.	Yes
Discussion	Participants were asked to jointly discuss issues surrounding the statement "Social Networks have made the world a better place", and come to agreement whether they believe the statement is true or not.	No
Subjective Question	Participants responded to the subjective question "If you could be any animal, what animal and why?"	No
Telling jokes	Participants were asked to take it in turn telling three separate jokes.	Yes

B. Action Descriptions

To obtain natural action performances the participants were asked to complete seven conversational scenarios, Table I. The subjects had no time limit on their tasks, with performances such as 'Telling Jokes' naturally taking less time to complete than those of 'Debate' and 'Describe Work'. Certain scenarios required participants to prepare material in advance, these are also identified in Table I.

Before each sequence, the subjects were asked if they were happy to continue and reminded of the scenario to be carried out. The observers left the room and the interaction could begin once both participants felt ready to proceed. When the interaction had come to its conclusion the participants signaled to the observers and the process was repeated until all seven scenarios had played out. There was no predefined script or suggestion as to how the scenario should be executed beyond those described above. This allowed CONVERSE to capture a more realistic and natural set of interactions between two individuals than current pose-based sets of emphasized action classes. Example appearance information for recordings can be seen in Figure 2.

Overall 16 participants took part in the CONVERSE dataset interactions. Participants were arranged into pairs and 8 interactions were captured in the dataset. Subjects came from a variety of backgrounds within the university.

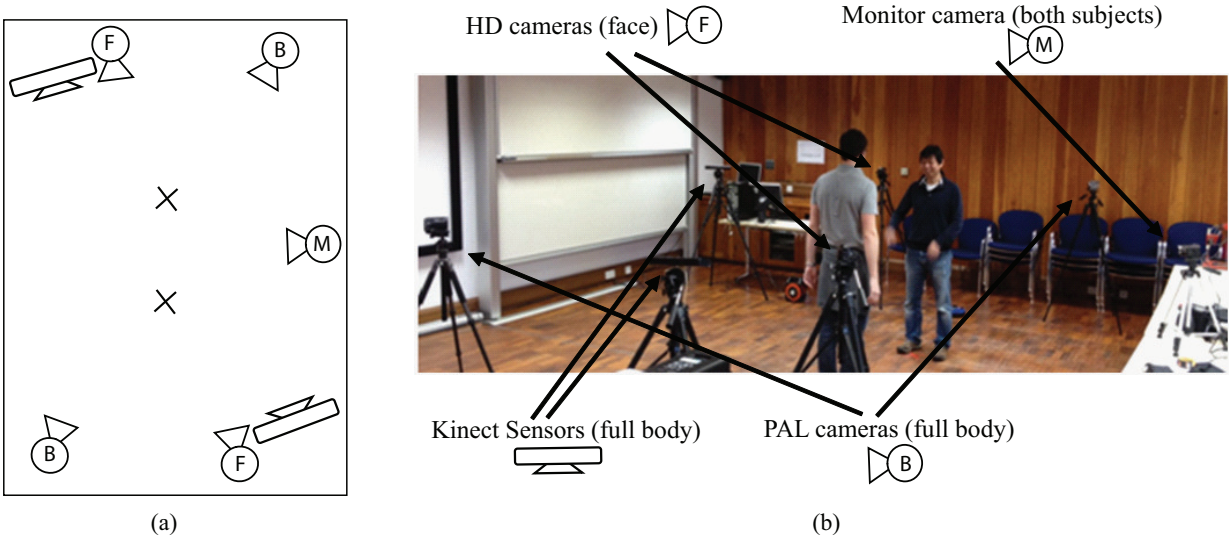

Fig. 1. CONVERSE data capture environment. a) Plan view of the capture setup b) Photo of subjects in situ.

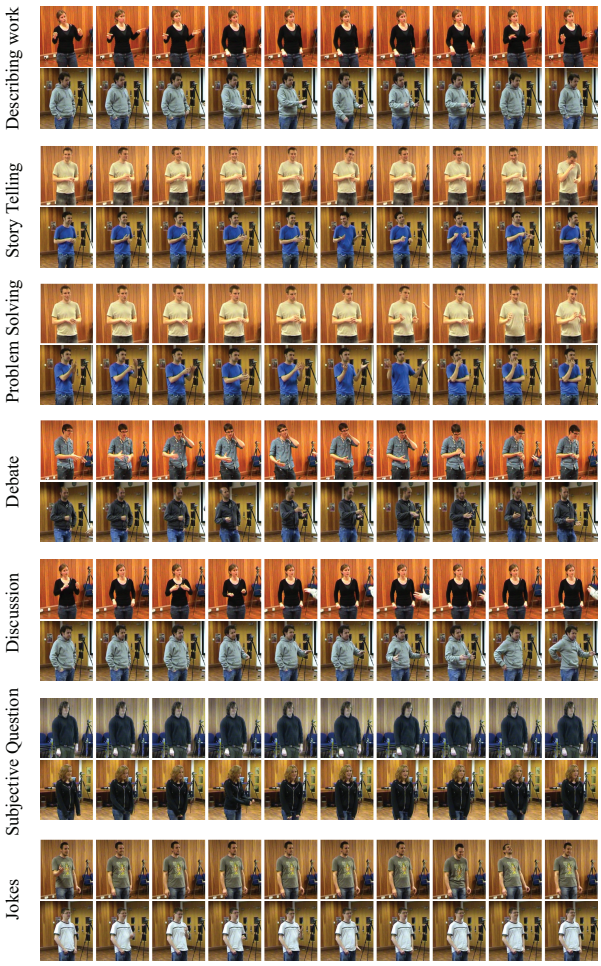

Fig. 2. Example recordings of CONVERSE classes. Appearance data was only used to annotate the depth based skeletal information.

C. Data Provided

CONVERSE provides human skeleton joints as tracked by the Kinect SDK. This skeleton presents the 3D coordinates for 20 joints, and the associated tracking confidence for each joint in a given frame. CONVERSE purposefully omits the appearance based information of the RGB recordings, and the pixel-wise depth information of the depth map recordings. Due to the private nature of many of the recordings, the audio has also been removed from CONVERSE sequences. This has several benefits to the dataset; firstly by allowing such unconstrained interaction it allows the dataset to represent more natural and unscripted observations. Secondly, by stripping the RGB and audio data, CONVERSE removes cues provided by audio features; this further adds to the complexity of understanding subtle conversational interactions through pose. See Figure 3 for skeletal information examples provided within CONVERSE, and the appearance information used to annotate the labelling.

III. Dataset Annotation

One of the major objectives of the CONVERSE dataset is to study the relationship between facial expression, bodily pose, and conversational scenarios from a computer vision perspective. Below we present some challenges in the labelling of subtle interactions over a long timeframe; including the use of sequence alignment and labelling. A custom MATLAB interface was written to explore and annotate the skeletal data from the Kinect and is provided online with the data [21]. For video annotation and labelling we utlised the iMovie software provided by Apple.

A. Labelling Challenges

Ground-truth annotation for such a subtle collection of tasks such as described by CONVERSE is a complex task. Both facial, bodily cameras and Kinect sensors were operated

Fig. 3. Skeletal data within CONVERSE and the associated appearance information. Appearance information was subsequently removed from the dataset.

independently. All were recording continuously across all scenarios, started once and ended once for each pair of participants. This means that the raw data contains not only events of interest, but also transitions between events where subjects are refreshed on the current class of interest. This continuous recording makes it essential to crop the actual events into discrete sequences for each class, allowing the classification of individual conversational scenarios. It is difficult for a human to recognize the interaction classes without visual and audio cues, and as such the facial and body RGB cameras are used to assist in locating appropriate start and end points for the pose based recordings. Another challenge to the dataset annotation is that the frame rates of video cameras and Kinect sensor differ, where the RGB cameras capture at 25fps, and the Kinect sensor records the skeleton kinematics at 30fps. More severely, the frame rates of individual devices are not a constant across time, which results in different length data sequences in terms of the number of frames.

B. Sequence Alignment

The monitor camera that captured two participants in a single view (M in 1) was used to locate the start and end points of events. These start and end points were then used as references to align other data; including facial recordings, full body videos and the Kinect kinematic skeleton sequences. To precisely locate the corresponding frames in different data sequences, frame-by-frame comparison was adopted for finding the optimal matchings. Individual conversational scenarios were manually annotated using the following standard procedures:

1) Locate the start and end frames of target event in the output of monitor camera, based on both audio and filming logs.
2) Find the corresponding RGB frames in output from the full body cameras, using the RGB image from monitor camera as reference.
3) Find the corresponding skeleton frames in output from the Kinect sensors, using the RGB images from body cameras as reference.
4) Find the corresponding RGB frames in both facial camera outputs using the RGB images from body cameras as references.
5) Use the full body recording of participant 1 as reference to re-sample all other sequences to constant frame rates (25FPS and 30FPS for RGB camera and Kinect sensor respectively) via frame-wise interpolation.
6) Remove the audio, and annotate all synchronized sequences with target event label.

C. Frame Labelling

Although several appearance based sets make use of frame-wise annotation of observations, these are commonly in the human action detection field and often focus on surveillance applications [7]. As the observation of a continuous conversational interaction can be highly complex, it was decided that sequence-wise labeling was more appropriate for the CONVERSE set. Obtaining a ground truth for frame-wise labelling of such recordings could be highly contentious, and as such may impact greatly on the performance of models that are evaluated on a frame-wise basis. The semantics behind an given individual frame is a complex result of frames that precede and succeed it, and therefore manual framewise labelling of long term interaction classes would provide variable interpretations. With such a large quantity of frames, it would also be a non-trivial task to obtain reliable manual frame-wise labelling of all sequences in the CONVERSE set, especially given the subjective nature of when a task has begun or been completed. By labelling the sequence as a whole it is possible to quickly obtain a labelling for an entire task, knowing that the class exists within the observation. This sequence-wise labelling also lowers the impact of a small variation in frame

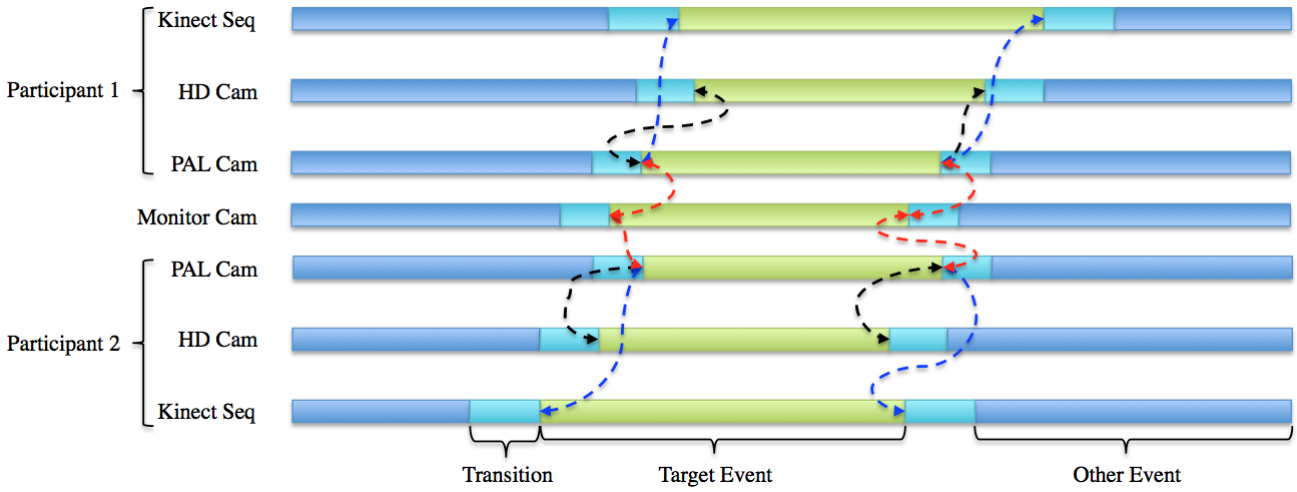

Fig. 4. Illustration of data annotation. Appearance information from the monitor camera is used to identify start and end points in the appearance recordings for each participant. This is in turn used to identify start and end points in the face and skeletal recordings.

labelling at the start and ends of an observation. The labels are not tightly bound to the interaction observed, instead the observed task is annotated as existing within the sequence.

Sequence-wise annotation was adopted for CONVERSE, where both facial appearance, bodily appearance and Kinematic sequences of individual conversational scenarios were aligned by finding the start and end time (See Figure 4). To assure the quality of annotation, the synchronized sequences and their associated labellings were then verified by a third person. As the study mainly investigates the connections of bodily motion and facial expression with conversational scenarios, the appearance and audio information was only used for annotation purpose and thus discarded. This results in the kinematic skeletal position data as represented within CONVERSE.

IV. Conclusion

CONVERSE provides a highly complex and subtle set of interactions for use in human action recognition purely from 3D pose information. The dataset utilizes appearance and audio data information in the ground truth labelling of classes that show strong correlation with appearance and audio cues. One of the main challenges proposed by CONVERSE asks if it is still possible to recognize such classes once such descriptive appearance and audio information has been stripped from the observation. This method allows the ground truth labelling to utilize a higher level of knowledge than is permitted to a system utilizing the data, posing a difficult problem within HAR. The dataset also provides a natural and unscripted execution of the interactions, and as such requires careful annotation of the recordings to identify suitable sequence labelling for use in subsequent evaluation. Due to the complex nature of the interactions and sheer volume of frames obtained, sequence based labelling was employed to annotate sequences with the scenario class observed. Although frame-wise labelling is beneficial to detection and localization problems, it is non-trivial to define suitable annotations of frames which display such complex human behaviors.

References

[1] C. Schuldt, I. Laptev, and B. Caputo, "Recognizing human actions : A local SVM approach," in *Pat. Rec.*, 2004, pp. 3–7.

[2] M. Blank, L. Gorelick, E. Shechtman, M. Irani, and R. Basri, "Actions as space-time shapes," in *Proc. Int. Conf. on Comp. Vis.*, 2005, pp. 1395–1402.

[3] ——, "Actions as space-time shapes," *IEEE Trans. Pat. Ana. & Mach. Int.*, vol. 29, no. 12, pp. 2247–53, 2007.

[4] A. Nghiem and F. Bremond, "ETISEO, performance evaluation for video surveillance systems," in *Advanced Video and Signal-Based Surveillance*, 2007.

[5] H. Ragheb and S. Velastin, "ViHASi: Virtual Human Action Silhouette data for the performance evaluation of silhouette-based action recognition methods," in *Int. Conf. on Distributed Smart Cameras*, 2008, pp. 1–10.

[6] J. Liu, "Recognizing realistic actions from videos in the wild," in *Proc. IEEE Conf. on Comp. Vis. and Pat. Rec.*, 2009, pp. 1996–2003.

[7] K. Huang, D. Tao, Y. Yuan, X. Li, and T. Tan, "View-independent behavior analysis." in *Proc. IEEE Int. Conf. Syst., Man and Cybern.*, vol. 39, no. 4, 2009, pp. 1028–35.

[8] R. Messing, C. Pal, and H. Kautz, "Activity recognition using the velocity histories of tracked keypoints," in *Proc. Int. Conf. on Comp. Vis.*, 2009.

[9] Y. Kong, Y. Jia, and Y. Fu, "Learning human interaction by interactive phrases," in *Proc. Euro. Conf. on Comp. Vis.*, vol. 7572, 2012, pp. 300–313.

[10] I. Laptev and M. Marszalek, "Learning realistic human actions from movies," in *Proc. IEEE Conf. on Comp. Vis. and Pat. Rec.*, 2008.

[11] P. Scovanner, S. Ali, and M. Shah, "A 3-dimensional SIFT descriptor and its application to action recognition," in *Proc. Int. Conf. on Multimedia*. New York, NY, USA: ACM, 2007, pp. 357–360.

[12] I. Laptev, "On space-time interest points," *Int. J. Comp. Vis.*, vol. 64, no. 2-3, pp. 107–123, 2005.

[13] J. Imran and P. Kumar, "Human action recognition using RGB-D sensor and deep convolutional neural networks," in *2016 Int. Conf. Advances in Computing, Communications and Informatics*, Sept 2016, pp. 144–148.

[14] W. Li, Z. Zhang, and Z. Liu, "Action recognition based on a bag of 3D points," in *IEEE Int. Workshop on CVPR for Human Communicative Behavior Analysis*, 2010.

[15] F. Ofli, R. Chaudhry, G. Kurillo, R. Vidal, and R. Bajcsy, "Berkeley MHAD: A comprehensive Multimodal Human Action Database," *Workshop on Applications of Computer Vision*, pp. 53–60, 2013.

[16] K. Yun, J. Honorio, D. Chattopadhyay, T. L. Berg, and D. Samaras, "Two-person interaction detection using body-pose features and multiple instance learning," in *Proc. IEEE Conf. on Comp. Vis. and Pat. Rec. Workshops*, 2012.

[17] T. Hu, X. Zhu, W. Guo, and K. Su, "Efficient interaction recognition through positive action representation," *Mathematical Problems in Engineering*, vol. 2013, pp. 1–11, 2013.

[18] J. Wang, Z. Liu, Y. Wu, and J. Yuan, "Mining actionlet ensemble for action recognition with depth cameras," in *Proc. IEEE Conf. on Comp. Vis. and Pat. Rec.*, 2012, pp. 1290–1297.

[19] M. Tenorth, J. Bandouch, and M. Beetz, "The TUM Kitchen Data Set of everyday manipulation activities for motion tracking and action recognition," *Proc. Int. Conf. on Comp. Vis. Workshops*, pp. 1089–1096, 2009.

[20] M. Edwards, J. Deng, and X. Xie, "From pose to activity: Surveying datasets and introducing CONVERSE," *Comp. Vis. Image Underst.*, vol. 144, no. C, pp. 73–105, Mar. 2016.

[21] Swansea University Computer Vision and Medical Image Analysis Group, "CONVERSE dataset," date accessed: 29/07/2015. [Online]. Available: http://csvision.swan.ac.uk/converse

[22] J. Deng, X. Xie, and B. Daubney, "A bag of words approach to subject specific 3D human pose interaction classification with random decision forests," *Graphical Models*, 2014.