2019

Comparison of Activity Recognition Using 2D and 3D Skeletal Joint Data

Fiona Marshall
Ulster University

Shuai Zhang
Ulster University

Bryan Scotney
Ulster University

Recommended Citation

Marshall, F., Zhang, S., & Scotney, B. (2019). Comparison of activity recognition using 2D and 3D skeletal joint data. *IMVIP 2019: Irish Machine Vision & Image Processing*, Technological University Dublin, Dublin, Ireland, August 28-30. doi:10.21427/rxw5-q154

Comparison of Activity Recognition using 2D and 3D Skeletal Joint Data

Fiona Marshall, Shuai Zhang and Bryan Scotney

School of Computing, Ulster University

Abstract

With the recent development of cheap and accurate depth sensors, activity recognition research has largely focused on the use of features created from 3D, rather than 2D, skeletal joint location. Nevertheless, conventional 2D RGB cameras remain an attractive data collection tool due to their low cost and ease of use. This study investigates the benefits of using 2D skeletal joints for activity recognition using visualisation and an exemplar classifier. Results show that 2D models can be as informative as 3D models, demonstrating the informativeness of joints extracted from RGB video.

Keywords: Activity Recognition, 2D Skeletal Joints, 3D Skeletal Joints, RGB

1 Introduction

Automatic recognition of human activity has many applications in fields as diverse as surveillance, human-machine interaction and virtual reality, and over the past two decades has been the focus of much research [Han, et al., 2017]. Whilst conventional cameras have frequently been used for data capture [Aggarwal and Xia, 2014], the rapid development of depth sensors over the past decade has resulted in most recent activity recognition approaches using 3D joint positions to describe posture and movement [Han, et al., 2017]. Yet, as humans can recognise activities from 2D video data, then the same should be true for computers. In many situations it is more practical to use a conventional camera to collect RGB video data than to collect 3D data with a specialist sensor. A conventional camera is cost effective; does not require specialist equipment; can provide a wide field of view and be used both indoors and outdoors.

Despite a general assumption that 3D data is more informative than 2D for activity recognition, and that it is more robust to variation in illumination, scale and rotation [Han, et al., 2017], we are not aware of any research that compares the discriminative ability of 3D and 2D skeletal joints, and of only limited research comparing the different data modalities. To guide future activity recognition research, this paper evaluates the discriminative abilities of features created from 2D and 3D skeletal joint positions extracted from depth and RGB data.

2 Related Work

Awareness of the discriminative ability of a small number of moving points extends from Johansson's experiments in visual motion perception [1975]. Intuitively, activity recognition is based upon the ability to represent the human pose, [Han, et al., 2017] but until recently the automatic extraction of skeletal joint positions was deemed challenging and not a viable option [Poppe, 2010]. In 2012 Microsoft launched the SDK toolkit for Kinect[1], providing cost effective 3D skeletal joint location. Since its introduction, many activity recognition approaches have been based on 3D joints. The automatic extraction of joints from 2D images remained elusive until the arrival of OpenPose [Wei, et al., 2016], an open source joint extraction tool for RGB data. This time lag has resulted in most activity recognition models being based on 3D rather than 2D joint location.

2D and 3D joint features can be constructed in a variety of ways. Joints can be described by their raw

[1] Microsoft Kinect, https://dev.windows.com/en-us/kinect (2012).

position [Pham, et al., 2018]; position relative to other joints [Yang and Tian, 2014]; orientation (absolute [Ofli, et al., 2014] or relative [Boubou and Suzuki, 2014]); or as a combination of the above; spatially or temporally, or both. Joints can be represented by concatenating data from each joint [Yang and Tian, 2014]; statistically combining data using histograms [Wang, et al., 2013]; or by summarising into a bag of features [Xu, et al., 2012]. Furthermore, joints can be represented individually [Zanfir, et al., 2013], or combined into body parts [Wang, et al., 2013]. Joint locations may be represented in a Euclidean space or on a manifold [Tanfous, et al., 2018].

Features can be handcrafted [Zanfir, et al., 2013], or selected using dictionary learning [Wang, et al., 2014], unsupervised learning [Yang and Tian, 2014] or deep learning. Classification approaches include discriminative, statistical and deep learning. Commonly used discriminative approaches are Support Vector Machines [Tanfous, et al., 2018] and Nearest Neighbour [Ofli, et al., 2014]. Statistical methods that consider the sequence of poses include Hidden Markov Models [Hai and Kha, 2016]. Deep Learning methods that directly encode temporal information include Convolutional Neural Networks (CNN) [Pham, et al., 2018] and Long Short-Term Memory (LSTM) [Baradel, et al., 2017].

Only a limited number of approaches are based on skeletal joints from 2D video, using either manually annotated joints or extracted using the method described by [Wang, et al., 2012]. Handcrafted features have been created and classified with SVMs: Xu, et al., [2012] created a codebook of activities with local motion information, whilst Wang, et al., [2013] used contrast mining to create a bag of words. CNNs have been employed with joint locations to classify activity from 2D video. Cheron, et al., [2015] used joint locations to select image patches from which pixel and optical flow values were input into multiple CNNs. Conversely, Cao, et al., [2016] used joint location to select relevant features after the filter layer of a CNN. Li, et al., [2018] input pixel, optical flow, joint location and motion values into separate CNNs.

Han, et al., [2017] reflected that whilst no single approach gives the best results over different datasets, approaches that integrate joint features from temporal and spatial sources perform better than those from a single source. Li, et al., [2018] demonstrated that features created from RGB images and skeletal joint location are complementary. Whilst many methods seek to improve accuracy by increasing features and using more complex networks, real-time classification is vital for many applications and computationally inexpensive approaches are important [Han, et al., 2017]. Pham, et al., [2018] created an efficient activity recognition model by reducing each activity sequence of 3D skeletal joints to a RGB image which was classified by a residual network.

Liu, et al., [2019] compared the classification ability of models based on different modalities (rgb, depth and skeletal joints). The RGB and depth models used a CNN with image and motion streams; whilst the skeletal joint model used a LSTM. It was observed that whilst joint location data achieved the highest accuracy for cross-set up, RGB data achieved the highest cross-subject accuracy suggesting that whilst 3D joints are more invariant to view, image-based methods benefit from appearance and texture. The highest accuracy was achieved by fusing the results of the 3 modalities. Joint location extracted from RGB data was not considered. Wang, et al., [2013] demonstrated that 2D joints provide similar discriminative information to 3D joints by using the same model for activity classification of a 3D dataset [Li and Way, 2010] and 2D dataset [Rodriguez, et al., 2008].

The activity recognition approaches reviewed demonstrate the benefits of using both spatial and temporal data, as well as the discriminative ability of both image and skeletal joint data. However, we have not found any strong evidence to demonstrate that 3D joint location is necessarily more informative than for activity recognition.

3 Method of Comparison

We investigate the value of depth data for activity recognition by comparing the discriminative ability of 2D and 3D skeletal joint location using T-distributed stochastic neighbour embedding (t-SNE) visualisation [Maaten, and Hinton, 2008] and an exemplar model of activity recognition. An intuitive approach to feature selection and classification was used to show the effect of data from different sources. As it has been demonstrated that both spatial and temporal information are important for activity recognition [Han, et al., 2017] joint position, acceleration and motion features were concatenated for each frame. T-SNE, an algorithm for the visualisation of high-dimensional datasets, was used to visualise the informativeness of features.

Data used was from public datasets collected from Kinect style sensors that provide depth, skeletal joint location and (where available) video data. Raw data from three sources were compared: 3D skeletal joint positions obtained from Kinect style depth sensors; 2D skeletal joint positions obtained from Kinect style depth sensors with the depth information discarded; and 2D skeletal joint position obtained using OpenPose from RGB video.

3.1 Joint Extraction

2D joint positions (x and y co-ordinates) are obtained from RGB video using OpenPose. OpenPose returns 18 skeletal points: nose, neck, left and right shoulders, elbows, hands, hips, knees, feet, eyes and ears (Figure 1). Kinect SDK returns 20 3D skeletal body points (as x, y and z co-ordinates): head, shoulder centre, spine, hip centre, left and right shoulders, elbows, wrists, hands, hips, knees, ankles and feet. 2D joint positions are obtained from Kinect data by discarding the z co-ordinate of each joint.

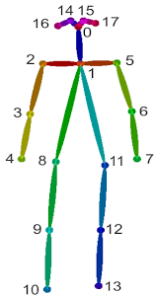

Figure 1: OpenPose model

3.2 Feature creation

14 skeletal joints were used: head, neck, right and left: shoulder, elbow, wrist, hip, knee and ankle. The joint positions were normalised for position with reference to the neck joint and scaled by dividing by the torso length d (the Euclidean distance between the neck joint and hip centre). For a single 2D frame, the location of the 14 joint locations is denoted $L = \{l_0, \dots, l_{13}\}$, as show in Figure 1, in the original image coordinate space $l_i = \{x_i, y_i\}$. Position is normalised around the neck joint (x_1, y_1). The spatial scale in each frame is also normalised using the torso length, d, resulting in the scale-normalised skeletal pose, $P = \{p_0, \dots, p_{13}\}$:

$$d = \sqrt{\left(\frac{1}{2}(x_8 - x_{11}) - x_1\right)^2 + \left(\frac{1}{2}(y_8 - y_{11}) - y_1\right)^2} \quad (1)$$

$$p_i = \left(\frac{x_i - x_1}{d}, \frac{y_i - y_1}{d}\right) \quad (2)$$

The neck co-ordinate (the origin) is discarded and each scale-normalised co-ordinate is smoothed along the time dimension using a 5x1 Gaussian filter over 5 frames (σ=1). Temporal information is encoded into each frame by calculating the acceleration and velocity around the current time frame (t_0) [Zanfir, et al., 2013]:

$$\delta P(t_0) \approx \frac{P(t_1) - P(t_{-1})}{2} \quad (3)$$

$$\delta^2 P(t_0) \approx \frac{P(t_2) + P(t_{-2}) - 2P(t_0)}{4} \quad (4)$$

A total of 78 features are concatenated as a linear vector for each frame (i.e. 13joints×3features×2D). The same method is used in 3 dimensions for 3D data, giving a total of 117 (13×3×3) features for each frame.

3.3 Classification

Each dataset was split into training and test sets according to the procedure used by other authors, as described in Section 4. A non-parametric multi-class classifier, kNN, is used to compare the discriminative power of the features. kNN is used due to its suitability to activity recognition tasks, with k=3 neighbours chosen as the average of the square roots of the number of classes in each dataset. Each frame in a sequence is classified individually, with the most frequently occurring frame label in a sequence (mode) selected as the sequence label.

4 Datasets

To compare the performance of approaches using 2D and 3D features, data collected by a 3D depth sensor is used. Four datasets have been used: MSR Action [Li and Way, 2010], KARD [Gaglio and Morana, 2015], MSR Daily Activity 3D [Wang, et al., 2012] and NTURGB-D [Shahroudy, et al., 2016]. Of these 4 datasets, 3 also provide the RGB video, enabling joint extraction using OpenPose (Table 1).

	No. of sequences	No. of Classes	No. of Subjects	Sensor	Modality	RGB Pixel resolution	Skeletal frames	RGB frames
MSR Action [Li and Way, 2010]	567	20	10	Kinect type	S,D	-	28,914	-
KARD [Gaglio and Morana, 2015]	540	18	10	Kinect v1	RGB,S,D	640x480	100,708	61,466
MSR Daily Action 3D [Wang, et al., 2012]	320	16	10	Kinect v1	RGB,S,D	640x480	160,663	45,021
NTU-RGB+D (small) [Shahroudy, et al., 2016]	192	8	12	Kinect v2	RGB,S,D	1920×1080	8,219	7,456

Table 1: Activity Recognition Datasets S=skeleton, D=depth

MSR Action [Li and Way, 2010] contains 3D skeletal joint location obtained from a Kinect type sensor. 10 subjects carry out 20 activities 3 times. RGB data is not available and there are some missing sequences. The activities are divided into 3 sets of similar activities for classification purposes. Action Set 1: *horizontal arm wave, hammer, forward punch, high throw, hand clap, bend, tennis serve, pickup & throw*. Action Set 2: *high arm wave, hand catch, draw x, draw tick, draw circle, two hand wave, forward kick, side-boxing*. Action Set 3: *high throw, forward kick, side kick, jogging, tennis swing, tennis serve, golf swing, pickup & throw*. Each dataset is split into training and test sets with half the subjects used for training. 10-fold cross validation was carried out for each action set and overall average accuracy is recorded for the 3 action sets.

KARD [Gaglio and Morana, 2015] contains RGB video and 3D skeletal joint location obtained from a Kinect v.1 device. 10 subjects carry out 18 activities 3 times. There are no missing sequences. Activities are divided into 3 datasets: Activity Set1: *horizontal arm wave, two hand wave, bend, phone call, stand up, forward kick, draw x, walk*; Activity Set2: *high arm wave, side kick, catch cap, draw tick, hand clap, forward kick, bend, sit down*; Activity Set3: *draw tick, drink, sit down, phone call, take umbrella, toss paper, high throw, horizontal arm wave*. 10-fold cross-validation was used each of the 3 action sets with the 3 testing methods as described in the original paper. To enable comparison with other approaches, accuracy is recorded for the average of all the tests and activity sets [Pham, et al., 2018].

MSR Daily Activity 3D [Wang, et al., 2012] was captured by using a Kinect v.1 device and contains 3D skeletal joint location and RGB video. 10 subjects carry out 16 activities: *drink, eat, read book, call cellphone, write on a paper, use laptop, use vacuum cleaner, cheer up, sit still, toss paper, play game, lie down on sofa, walk, play guitar*. 10 subjects perform each activity twice, *standing* and *sitting*. In line with other papers [Tao and Vidal, 2016], the training set comprised subjects 1,3,5,7,9 and the test set comprised subjects 2,4,6,8,10.

NTURGB-D [Shahroudy, et al., 2016] is a large dataset collected using a Kinect v.2 sensor with 56,880 action sequences of 40 subjects and 60 activities. Camera height, distance from the subject and camera angles are varied throughout. A small subset of NTURGB (192 sequences) was selected to contain similar activities and camera set up to the other datasets. The activities were: *drink, pick up, throw, sit down, stand up, clapping, wave and kick*. 12 people carried out each activity twice. Due to the large number of frames in which the subject did not

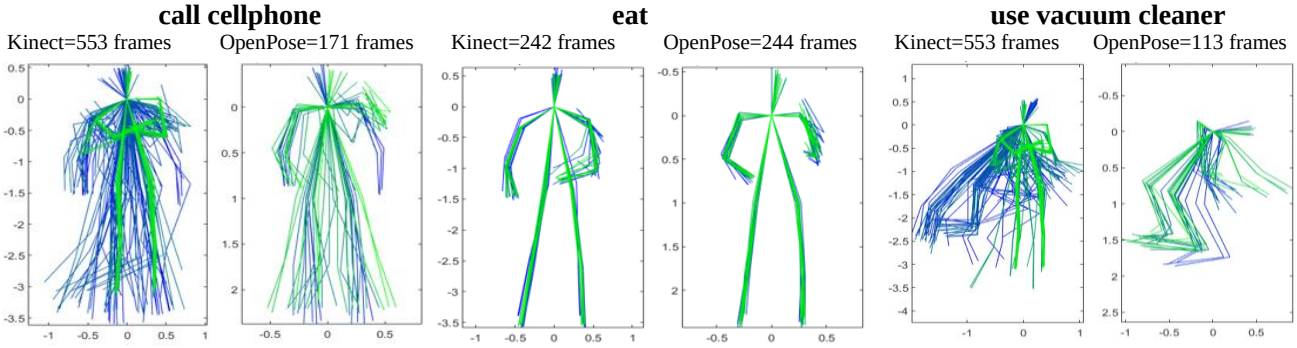

Figure 2: MSR Daily Activity. Comparison of Kinect 2D and OpenPose Skeletal Joints. Subject 1, Standing. Additional frames and higher collection rate for *call cellphone* and *use vacuum cleaner* can be seen. Movement is plotted for every 5th frame only, blue at start of activity, green at end.

move, each action sample was reduced by 20% to contain only the most informative consecutive frames using the AME method described in [Yang and Tian, 2014]. Action sequences are much shorter than other datasets, resulting in fewer frames. 10-fold cross subject testing was used, with 8 subjects for training and 4 for testing.

The KARD and MSR Daily datasets have significantly fewer frames of RGB than skeletal data for some sequences. Although collection rates are not given, a visualisation (Figure 2) suggests that RGB data is collected at a lower rate and over a shorter period. KARD and MSR Daily have a low pixel resolution, potentially reducing the accuracy of OpenPose joint location. NTURGB-D has a pixel resolution comparable to conventional cameras.

5 Results

Three sets of features were created from different sources: 2D joints extracted from RGB videos using OpenPose; 2D Kinect joints where depth values have been discarded; and 3D Kinect joints. We compare the informativeness of features created from the different modalities by visualisation and classification accuracies from the 4 datasets.

Features from the three modalities are visualised using t-SNE (Figure 3). T-SNE provides 2D visualisation of high dimensional data by grouping similar objects in hyper-dimensional space together and dissimilar objects further apart. Each point represents features from a single frame. Clustering of features can be seen for all 3 modalities; the 'ball' scatter seen in the Kinect plots for MSR Daily and KARD is due to uninformative frames at the start and end of activities where the subject is not moving; the shorter RGB sequences do not contain as many of these frames (Figure 2). Similar cluster patterns are seen in the 2D and 3D Kinect data for each dataset, indicating that there is little discriminative information gained by retaining depth values. The three NTURGB-D plots all show similar cluster patterns, indicating that features created from the different modalities are similar. Conversely, the

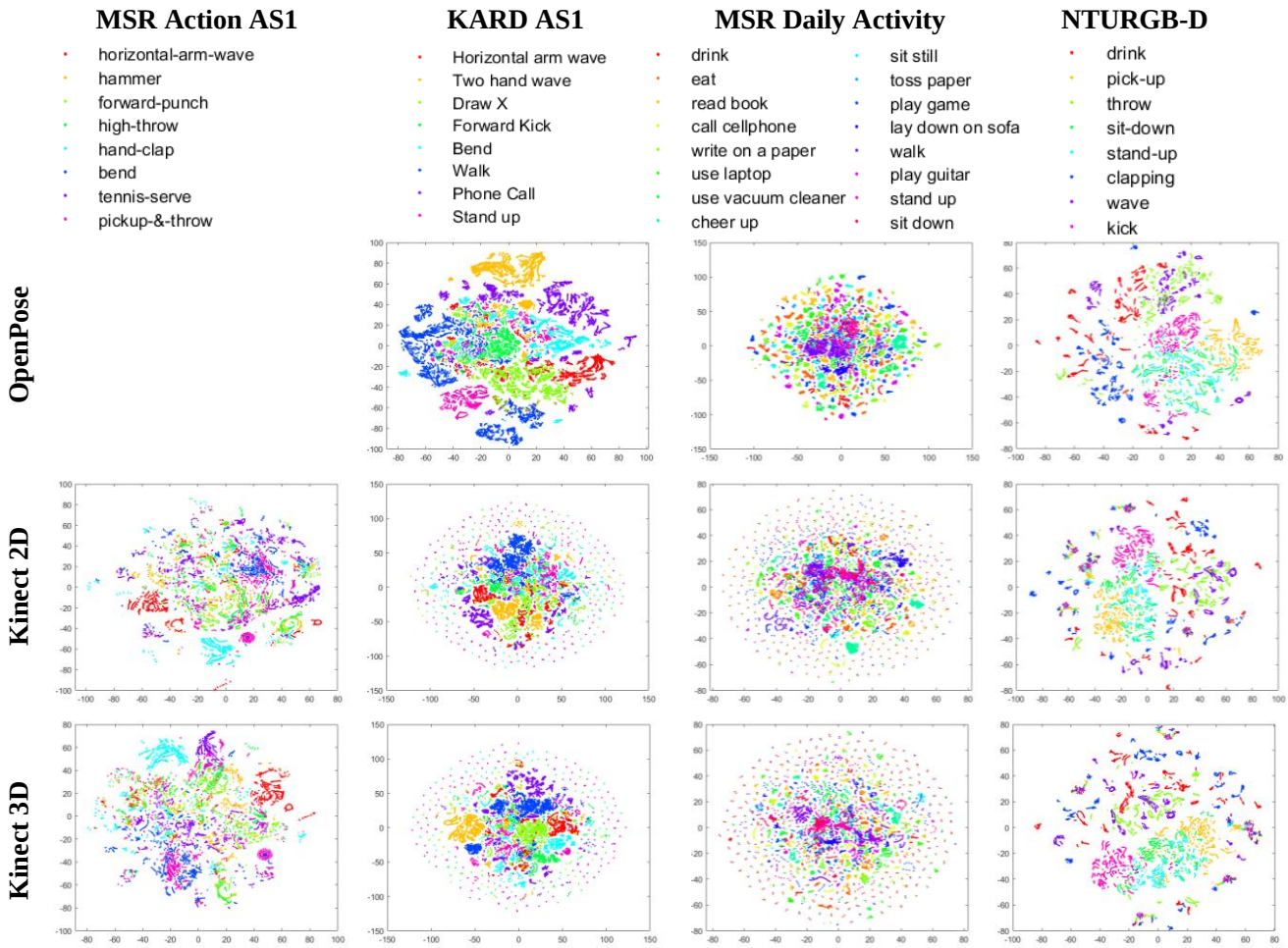

Figure 3: t-SNE plots showing similar patterns for each dataset demonstrate that data collected from the different sources provides similar information for classifying activity. Less 'ball' scatter in the KARD and MSR Daily Activity OpenPose plots is as a result of fewer frames (best viewed in colour)

OpenPose plots for KARD and MSR Daily, whose features were extracted from low-resolution images, look very different from the features created directly from Kinect joint position.

Results from our exemplar kNN classification model are compared for the 3 modalities (shaded region, Table 2). Accuracy is recorded for the entire activity sequence, and individual frames (in brackets). ANOVA[2] analysis of the 10 sets of cross-validation accuracies (MSR Action, KARD and NTURGB-D) show no significant difference between Kinect 2D and Kinect 3D results ($p=0.065$)[3]. Furthermore, no significant difference was found ($p=0.18$) between OpenPose and Kinect 3D cross-validation accuracies from KARD and NTURGB-D (small). ANOVA analysis was not carried out for the MSR Daily results due to cross-validation not being used. MSR Daily OpenPose data for performed poorly in comparison to the depth sensor data. This may be due to less RGB than joint frames (45,021 compared to 160,663) and low RGB pixel resolution. In contrast, with higher pixel resolution and a similar number of frames, the NTURGB-D OpenPose features outperform the 2D depth features, and ANOVA analysis showed no significant difference between cross-validation accuracies ($p=0.86$).

6 Discussion

Results suggest that joint extraction from a standard resolution video can provide features as informative as those from a depth sensor, demonstrating that 2D joint position can be as informative as 3D. Interestingly, the highest individual frame accuracies are from OpenPose features; potentially, in the case of MSR Daily and KARD, due to uninformative frames causing random classification; but also indicating that classifying a sequence by the mode of its individual frames is not ideal. When our results are compared with other published results that use 3D skeletal joint location (Table 2); all 3 data modalities of our exemplar model outperform those approaches which do not use temporal information. Approaches with similar accuracies to our model also encode temporal information: Mining Actionlet Ensemble [Wang, et al., 2012] employ Fourier Transformation Pyramids; Moving Poselets [Tao and Vidal, 2016] compute a parts based global feature; whilst the Moving Pose [Zanfir, et al., 2013]

	MSR Action (Average of 10 fold cv for 3 actions sets)	**KARD** (Average of 10 fold cv for 3 tests and 3 activity sets)	**MSR Daily Activity** (1 test / training set)	**NTURGB-D small** (10 fold cv)
Exemplar Model: OpenPose 2D	-	98.51 (90.6)	49.06 (41.2)	77.50 (66.7)
Exemplar Model: Kinect 2D	89.96 (61.8)	99.18 (77.85)	**81.25** $^{\alpha}$ (38.0)	76.88 (64.2)
Exemplar Model: Kinect 3D	**90.91** $^{\alpha}$ (63.9)	**99.47** $^{\alpha}$ (81.7)	79.38 (38.1)	**78.44** $^{\alpha}$ (63.9)
Eigenjoints [Yang and Tian, 2014]	83.3	-	-	-
Resnet 44 [Pham, et al., 2018]	**99.9** $^{\beta}$	**99.98** $^{\beta}$	-	-
Bag of 3D points [Li and Way, 2010]	74.7	-	-	-
HAR 3D posture [Gaglio and Morana, 2015]	-	90.83	-	-
Moving Poselets [Tao and Vidal, 2016]	93.6	-	74.5	-
Mining Actionlet Ensemble [Wang, et al., 2012]	88.2	-	**85.75** $^{\beta}$	-
The Moving Pose [Zanfir, et al., 2013]	91.7	-	73.8	-

Table 2: Comparison of Results: Accuracy for sequence classification (frame accuracy in brackets). $^{\alpha}$Highest accuracy from exemplar model, $^{\beta}$ Overall highest accuracy. Shaded region shows results from exemplar model.

[2] ANOVA - two factor with replication

[3] p-threshold of 0.05 [Fisher, R. A.,1925]

encodes velocity and acceleration. Resnet [Pham, et al., 2018], which uses LSTMs to retain sequence, outperforms our model. These results demonstrate the importance of temporal information whilst also showing that potentially depth may not be significant for activity recognition.

7 Conclusion and Future Work

Our work has shown that both 2D and 3D skeletal joint location can provide informative features for activity recognition, and that joints extracted from RGB images can be as useful as those extracted from 3D depth data.

Video cameras are an attractive data collection tool for activity recognition due to their low cost, large surveillance distance, ability to collect data outdoors and the availability of training data. Furthermore, as pose alone cannot describe human-object interaction [Han, et al., 2017], it is useful to fuse image data with joint position. This is straightforward when both sets of data are available from the same frame.

Our exemplar classification model has shown promising results for the use of skeletal joints extracted from 2D data. Future work will explore whether different feature selection and classification techniques show the same trend by comparing recognition accuracies of the different modalities of skeletal joint data with other classification techniques, including deep learning and statistical methods. It will also compare results from more complex datasets, including those with occlusion, and where subjects are not directly facing the camera.

8 References

[Aggarwal and Xia, 2014] Aggarwal, J.K., & Ryoo, M.S. (2011). *Human activity analysis: A review*. ACM Comput. Surv., 43, 16:1-16:43.

[Baradel, et al., 2017] Baradel, F., Wolf, C., & Mille, J. (2017). *Pose-conditioned Spatio-Temporal Attention for Human Action Recognition*. CoRR, abs/1703.10106.

[Boubou and Suzuki, 2014] Boubou, S., & Suzuki, E. (2014). *Classifying actions based on histogram of oriented velocity vectors*. Journal of Intelligent Information Systems, *44*(1), 49–65.

[Cao, et al., 2016] Cao, C., Zhang, Y., Zhang, C., & Lu, H. (2016). *Action Recognition with Joints-Pooled 3D Deep Convolutional Descriptors*. IJCAI.[Fisher, R, 1925]

[Fisher, R. A.,1925] Fisher, Ronald A., (1925). *Statistical Methods for Research Workers*. Edinburgh, UK: Oliver and Boyd. p. 43.

[Cheron, et al., 2015] Chéron, G., Laptev, I., & Schmid, C. (2015). *P-CNN: Pose-Based CNN Features for Action Recognition*. 2015 IEEE International Conference on Computer Vision (ICCV), 3218-3226.

[Gaglio and Morana, 2015] Gaglio, S., Re, G.L., & Morana, M. (2015). *Human Activity Recognition Process Using 3-D Posture Data*. IEEE Transactions on Human-Machine Systems, 45, 586-597.

[Hai and Kha, 2016] Hai, P.T., & Kha, H.H. (2016). *An efficient star skeleton extraction for human action recognition using hidden Markov models*. 2016 IEEE Sixth International Conference on Communications and Electronics (ICCE), 351-356.

[Han, et al., 2017] Han, F., Reily, B., Hoff, W.A., & Zhang, H. (2017). *Space-Time Representation of People Based on 3D Skeletal Data: A Review*. Computer Vision and Image Understanding, 158, 85-105.

[Johansson, 1975] Johansson, G. (1975). Visual Motion Perception. *Scientific American*, *232*(6), 76–89.

[Li and Way, 2010] Li, W., Zhang, Z., & Liu, Z. (2010). *Action recognition based on a bag of 3D points*. 2010 IEEE Computer Society Conference on Computer Vision and Pattern Recognition - Workshops, 9-14.

[Li, et al., 2018] Li, C., Tong, R., & Tang, M. (2018). *Modelling Human Body Pose for Action Recognition Using*

Deep Neural Networks. Arabian Journal for Science and Engineering, 43(12), 7777–7788.

[Liu, et al., 2019] Liu, J., Shahroudy, A., Perez, M.L., Wang, G., Duan, L., & Chichung, A.K. (2019). *NTU RGB+D 120: A Large-Scale Benchmark for 3D Human Activity Understanding.* IEEE transactions on pattern analysis and machine intelligence.

[Maaten, and Hinton, 2008] Maaten, L.V., & Hinton, G.E. (2008). *Visualizing Data using t-SNE*.

[Ofli, et al., 2014] Ofli, F., Chaudhry, R., Kurillo, G., Vidal, R., & Bajcsy, R. (2014). *Sequence of the most informative joints (SMIJ): A new representation for human skeletal action recognition*. Journal of Visual Communication and Image Representation, 25(1), 24–38.

[Pham, et al., 2018] Pham, H., Khoudour, L., Crouzil, A., Zegers, P., & Velastin, S.A. (2018). *Exploiting deep residual networks for human action recognition from skeletal data*. Computer Vision and Image Understanding, 170, 51-66.

[Poppe, 2010] Poppe, R. (2010). *A survey on vision-based human action recognition*. Image Vision Comput., 28, 976-990.

[Rodriguez, et al., 2008] Rodriguez, M.D., Ahmed, J., & Shah, M. (2008). *Action MACH a spatio-temporal Maximum Average Correlation Height filter for action recognition*. 2008 IEEE Conference on Computer Vision and Pattern Recognition, 1-8.

[Shahroudy, et al., 2016] Shahroudy, A., Liu, J., Ng, T., & Wang, G. (2016). *NTU RGB+D: A Large Scale Dataset for 3D Human Activity Analysis*. 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 1010-1019.

[Tanfous, et al., 2018] Tanfous, A.B., Drira, H., & Amor, B.B. (2018). *Coding Kendall's Shape Trajectories for 3D Action Recognition*. 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2840-2849.

[Tao and Vidal, 2016] Tao, L., & Vidal, R. (2015). *Moving Poselets: A Discriminative and Interpretable Skeletal Motion Representation for Action Recognition*. 2015 IEEE International Conference on Computer Vision Workshop (ICCVW), 303-311.

[Wang, et al., 2012] Wang, J., Liu, Z., Wu, Y., & Yuan, J. (2012). *Mining actionlet ensemble for action recognition with depth cameras*. 2012 IEEE Conference on Computer Vision and Pattern Recognition, 1290-1297.

[Wang, et al., 2013] Wang, C., Wang, Y., & Yuille, A. L. (2013). *An approach to pose-based action recognition*. Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition, 915–922.

[Wang, et al., 2014] Wang, J., Liu, Z., Wu, Y., & Yuan, J. (2014). *Learning Actionlet Ensemble for 3D Human Action Recognition.* IEEE Transactions on Pattern Analysis and Machine Intelligence, 36, 914-927.

[Wei, et al., 2016] Wei, S., Ramakrishna, V., Kanade, T., & Sheikh, Y. (2016). *Convolutional Pose Machines*. 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 4724-4732.

[Xu, et al., 2012] Xu, R., Agarwal, P., Kumar, S., Krovi, V.N., & Corso, J.J. (2012). *Combining Skeletal Pose with Local Motion for Human Activity Recognition*. AMDO.

[Yang and Tian, 2014] Yang, X., & Tian, Y. (2014). *Effective 3D action recognition using EigenJoints*. J. Visual Communication and Image Representation, 25, 2-11.

[Zanfir, et al., 2013] Zanfir, M., Leordeanu, M., & Sminchisescu, C. (2013). *The Moving Pose: An Efficient 3D Kinematics Descriptor for Low-Latency Action Recognition and Detection*. 2013 IEEE International Conference on Computer Vision, 2752-2759.