Temporal Distance Matrices for Squat Classification

Ryoji Ogata
Waseda University
ryouji40040wsg@toki.waseda.jp

Edgar Simo-Serra
Waseda University
ess@waseda.jp

Satoshi Iizuka
Tsukuba University
iizuka@cs.tsukuba.ac.jp

Hiroshi Ishikawa
Waseda University
hfs@waseda.jp

Abstract

When working out, it is necessary to perform the same action many times for it to have effect. If the action, such as squats or bench pressing, is performed with poor form, it can lead to serious injuries in the long term. With the prevention of such harm in mind, we present an action dataset of videos where different types of poor form are annotated for a diversity of subjects and backgrounds, and propose a model for the form-classification task based on temporal distance matrices, both in the case of squats. We first run a 3D pose detector, then normalize the pose and compute the distance matrix, in which each element represents the normalized distance between two joints. This representation is invariant under global translation and rotation, as well as robust to individual differences, allowing for better generalization to real world data. Our classification model consists of a CNN with 1D convolutions. Results show that our method significantly outperforms existing approaches for the task.

1. Introduction

In recent years, working out or fitness has become popular in order to improve health and pursue a certain physique. Working out improves basal metabolism, can prevent metabolic syndrome and lower stress. It is not only for young people either for the elderly it has merits such as improvement of posture and the rehabilitation of restricted movement. One might go so far as to say that muscle development is essential for human beings to live healthily.

However, there is a danger in working out. Poor form of motion does not only utilize muscles incorrectly and thus decrease workout efficiency but also significantly increases the possibility of injury. Many beginners work out with poor form, leading to an assortment of problems. In recent years, being taught by personal trainers has become widespread, but it incurs an economical burden, and still does not guarantee that the correct form will be taught.

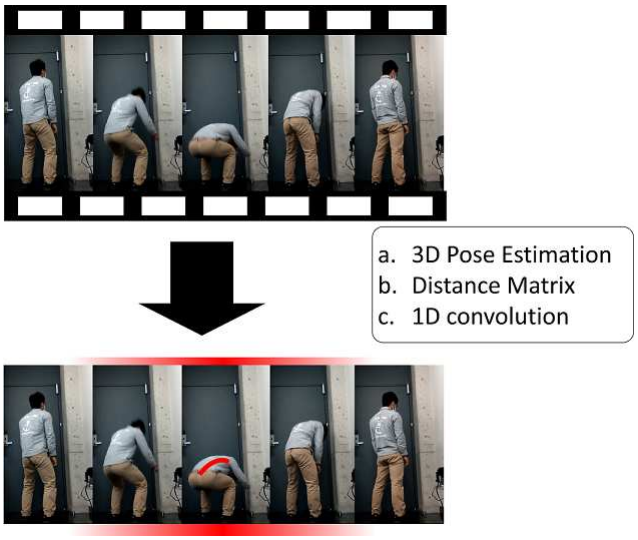
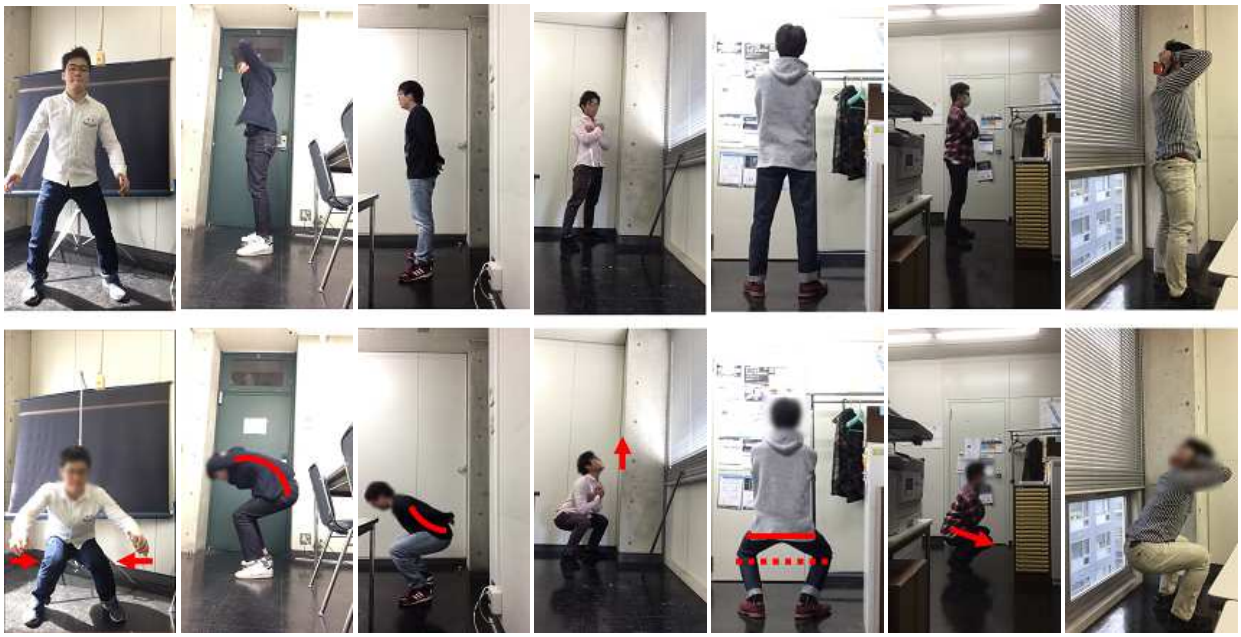


Figure 1: Given a video of a user performing squats, we estimate the 3D pose of each frame, and convert the 3D data to a temporal distance matrix representation. We then perform classification to detect mistakes in the squat form. Our data representation is independent of the image, allowing for strong generalization across scenes.

In this work, we attempt to automatically detect and correct poor form when working out, utilizing videos. In particular, we target squats and classify the different types of poor form beginners. Issues we face in the task include the following. First of all, there are no associated datasets for the task. For this reason, we have



Inward Knees Round Back Warped Back Upwards Head Shallowness Frontal Knee Good Squat

Figure 2: We have collected a multi-class dataset consisting of common squat mistakes and correct squats. The dataset contains a variety of backgrounds and individuals performing the squats.

created our own squat video dataset[1] with annotations on form, to serve as a testbed for different approaches (Fig. 2). Furthermore, the variety in lighting, background, clothing, individual, etc. makes the detection in a general setting difficult, and necessitates an impossibly large datasets, if a simple learning approach were to be taken. Therefore, it is important to develop algorithms that generalize well on inputs different from the training data. For this purpose, we developed a technique based on temporal distance matrices.

Our method is based on first extracting the 3D human pose from an input video as shown in Fig. 1. This gives a representation that is independent of the pixel information and thus independent of the background, illumination, etc. However, it is still sensitive to differences between individuals and between reference frames. Therefore, we perform a normalization of the different limb lengths, which gives a subject-independent representation, and compute the distance matrix of the 3D pose, i.e., the distance between all the different joints. This gives us a representation that is largely independent of the scene information, the individual bone length, and the reference frame. This representation is amenable to processing with Convolutional Neural Networks (CNN) with 1D convolutions. We propose a model based on ResNet that obtains high performance for the task of working out classification.

[1]Dataset will be made publicly available.

In this paper, we focus on six common kinds of poor squat form: inward knees, round back, warped back, upwards head, shallowness, and frontal knee, in addition to good squats. We collect several sources of data, including detailed data from a single individual, more data from multiple individuals with varying scenes, and videos from YouTube, and do an ablation analysis of the different components of our approach. We also provide a comparison with existing approaches for video classification and show a significant improvement for the squat classification task.

In summary, our contributions are the following:

1. A dataset for classification of good and various bad form of squats.
2. A method to assess the workout form from video by a feature extraction approach based on temporal distance matrices, which is robust to differences in scene, subject, and global translation and rotation.
3. An experimental validation of our method, in which it outperforms existing video classification approaches.

2. Related work

Our work is closely related to action recognition, action assessment, and pose estimation.

Action recognition is the major area of study related to our purpose, where the objective is to classify

what the person in a video is doing. Most of current literature relies on learning that requires a large training dataset of videos, which include UCF-101[30], kinects[14], ActivityNet[8], YouTube-8M[1], HMBD[9], and MACH[26]. These are vast collections of data aiming to contain as general actions and environments as possible. Recognizing video input requires exploiting three-dimensional features, representing sequential as well as spatial information. For instance, the two-stream method[29, 7] utilizes both the spatial features from single frames and the sequential information from optical flow, putting them into a CNN together to recognize actions. C3D[31] feeds video frames into a CNN directly, which turned out to be more effective. This has been improved by fine-tuning the result of learning with two-stream[3] and by incorporating global features by adding a non-local module[32]. Although LSTM [27] can relate temporal features, it becomes harder to train as the sequence gets longer. In our method, we treat temporal features by using residual networks [10]. For action recognition, using pose estimation is also effective, as it can remove background features and focus on human movement [5, 12, 21]. In recent years, multiview camera datasets [11, 15] have enabled 3D pose estimation. Using these datasets, 3D pose estimation can be done after 2D pose estimation, leading to improved action recognition [19, 18, 25, 17]. However, as these methods directly input coordinates to CNN, they have dependence on the orientation and the location. This can be remedied by computing the Euclidean distance matrix, which is independent of orientation and location [20]. Our method adds temporal features on this method using the distance matrix to improve accuracy.

Thus, our task here is closely related to action assessment, which estimates how well the action is performed. In the sport-related vision, there have been proposed methods to automatically score a dive or a skate performance mimicking human expert scorers[24, 23]. Besides scoring, the method in [23] also provides a feedback as to where the action can be improved, which is useful for the athletes. Out of sports, [6] gives relative scores on skills such as drawing and the use of chopsticks by comparing videos in the first-person view. This view dependence somewhat limits the applicability of the method under different conditions. For action assessment under more general conditions, there are methods[4, 22, 2] that utilize three-dimensional pose information acquired by kinect. The need for kinect, however limits the applicability in another way. In this paper, we use a single ordinary camera for pose estimation, giving more general applicability.

3. Dataset

We present four datasets of videos for the classification of good and differently bad squat forms: the Single Individual, the Multiple Individual, the Background Change, and the YouTube dataset. Table 1 summarizes the difference between the four.

3.1. Video

Each video shows one person performing squats and lasts for 10 seconds (300 frames), which translates to approximately three to five squats. The dataset contains a variety of individuals, clothing, and backgrounds. There are our original videos (i.e., videos we took), and videos downloaded from YouTube. In Table 1, the Source column indicates where the videos in each dataset come from (our original, YouTube, or both). The camera is static in all the videos.

3.2. Background

In the case of our original videos, the background of the videos are classified as "simple" or "complex". In Table 1, the Background column indicates if this classification is uniform or not in the test and training subset of the data. In the Single Individual and the Multiple Individual datasets, the data is divided into test and training subsets randomly, so each subset include videos with both "simple" and "complex" backgrounds. This is indicated as "mixed". In the Background Change dataset, on the other hand, all the videos with "simple" backgrounds are used as the training data and all with "complex" backgrounds are used as the test data, to test robustness to different kinds of background between training and test time. The videos in YouTube dataset have various background.

3.3. Single Individual Dataset

The single individual dataset serves as a baseline. It consists of our original videos, all showing the same individual performing squats. Details are shown in Table 2. Each video shows one of the seven (one good and six bad) classes of squat forms. Thus, the ground truth label is one-hot, i.e., each video has one of the seven labels as its true classification. Although it is conceivable to use a score instead of the one-hot label, we decided scoring squats is too difficult. The total of 2001 videos have been taken, and then randomly assigned to the test, training, and validation subsets, which contain 612, 1160, and 229 videos, respectively.

3.4. Multiple Individual Dataset

The Multiple Individual Dataset consists of 599 videos, including our original videos, showing seven

Table 1: Four squatting datasets details

	Source	#individuals	Background	Labels	#videos
Single Individual	original	1	mixed	One-hot	2001
Multiple Individual	both	14	mixed	Multi-label	599
Background Change	original	1	separated	One-hot	2001
YouTube	YouTube	7	-	Multi-label	23

Table 2: Single Individual Dataset. Number of videos with each label, divided into test, training, and validation subsets.

	Test	Train	Val	Total
Inward Knees	48	133	49	230
Round Back	95	161	24	280
Warped back	95	188	29	312
Upwards Head	86	154	32	272
Shallowness	105	180	34	319
Frontal Knee	110	159	26	295
Good Squat	73	185	35	293
Total	612	1160	229	2001

individuals, as well as those downloaded from YouTube. Table 3 shows the number of videos for each label and individual (A,...,G) or YouTube (YT). In this dataset, each video shows squats that may have multiple problems, as sometimes happens with a beginner, and thus may be multiply labeled. This is why the numbers for each individual's column do not add up to the bottom number, which is the actual number of videos showing that individual. A video is labeled "Good Squat" if and only if there is no problem. This dataset is meant to be tested with eight-fold cross validation, testing with the videos of one column (individual or YouTube) after training with the videos of other seven. In training, four videos are randomly chosen from the training data for use as validation set.

3.5. Background Change Dataset

This dataset is for testing the robustness to background change. The videos are same as those in the Single Individual dataset. Only the division into test, training, and validation subsets is different, as explained above in §3.2.

3.6. YouTube Dataset

The YouTube Dataset consists of videos downloaded from YouTube showing persons performing squats. This set is the same as those from YouTube in the Multiple Individual Dataset. Some examples are shown in Fig. 4. Videos showing squatting in YouTube seldom lasts for 10 seconds, so we could find only 23 videos.

Figure 3: Examples from the Background Change Dataset. Left two: test data with "complex" backgrounds (test, validation). Right two: training data with "simple" background (training).

Figure 4: YouTube (YT) Dataset examples. Various persons are performing squats in front of various background.

4. Proposed Approach

Our approach is based on exploiting the pose information. The main advantage of using pose information as a feature, rather than direct video information, is improved generalization. This is important as it is not simple to obtain and prepare data for sports activities such as squats. Not only is expert knowledge necessary, but a large amount of subjects must be recruited and different environment must be used for the video capture. Instead of using the video information, we use 3D pose information, which we condense into rotation and translation invariant representation using distance matrices, which are then amenable to further processing using 1D Convolutional Neural Networks. An overview of our approach is shown in Fig. 5.

Table 3: Multiple Individual Dataset details.

	A	B	C	D	E	F	G	YT	Total
Inward Knees	0	0	0	1	49	0	5	0	56
Round Back	17	0	0	10	109	0	40	1	177
Warped Back	22	0	3	6	39	0	6	11	87
Upwards Head	0	0	0	0	61	0	0	0	61
Shallowness	19	63	3	8	79	0	18	0	180
Frontal Knee	16	73	58	15	82	8	46	3	301
Good Squat	0	0	9	0	3	0	21	10	33
#videos	71	75	64	26	258	8	74	23	599

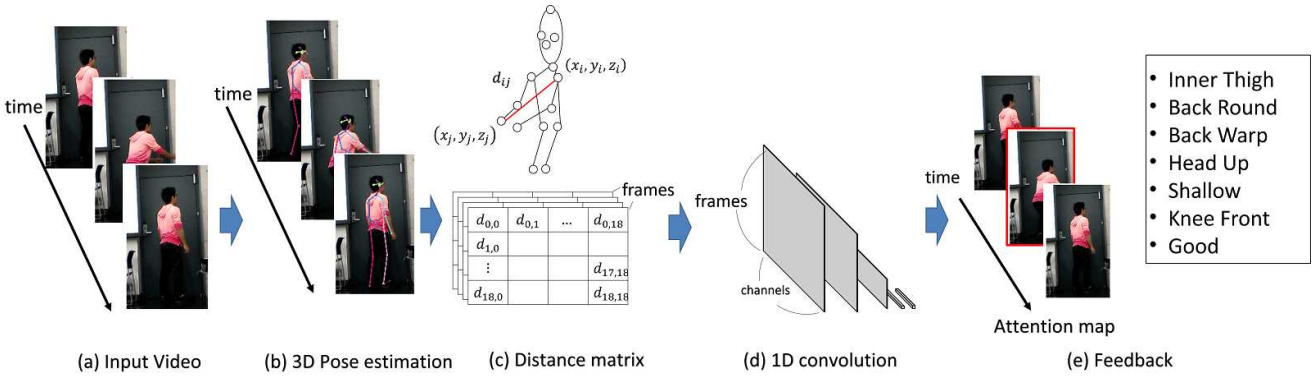


Figure 5: Overview of our proposed approach. a) Input video. b) We estimate the 3D human pose in each frame of the video. c) For each frame, we calculate the Euclidean distances between all pairs of the keypoints to obtain a symmetric distance matrix, take the upper-right triangle, and flatten it to a vector. d) The vectors for each frame are concatenated together temporally, and processed with a CNN based on 1D convolutions. e) The output classifies the action being done in the video.

Table 4: Background Change Dataset details.

	Test	Train	Val	Total
Inward Knees	22	203	5	230
Round Back	47	224	9	280
Warped Back	45	256	11	312
Upwards Head	32	232	8	272
Shallowness	36	266	17	319
Frontal Knee	40	246	9	295
Good Squat	27	258	8	293
Total	249	1685	67	2001

4.1. 3D Pose Estimation

For 3D pose estimation, we use the pose estimator described in [13]. This approach estimates the 3D pose of a single person from a monocular camera image. The output consists of 19 keypoints with their corresponding (x, y, z) 3D coordinates, and are normalized by the SMPL [16] model parameter. By using the SMPL model parameter, we can adjust according to individual body shape for more precise estimation, as well as detecting hidden keypoints.

4.2. Normalization

Even using the pose information, subject-specific information such as limb-length, which varies depending on individuals, still remain. When training with few individuals, this can lead to a bias and poor generalization of the results. To remedy this, we evaluate different types of normalization procedures, as shown in Fig. 6. Normalization is done by converting limbs to unit length.

4.3. Distance Matrices

A normalized pose is still not fully invariant, as it depends on the global reference frame. Although this variability can be partially absorbed by the 3D pose estimation network, we convert the 3D pose to a distance matrix, which is invariant to global translation,

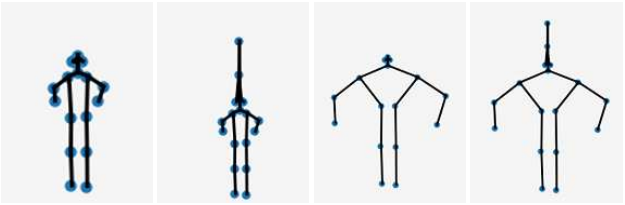


Figure 6: Examples of different types of 3D pose normalization. We normalize different limbs to unit length. This removes the dependency on individual characteristics of bone lengths. We test with normalization of different subsets of limbs.

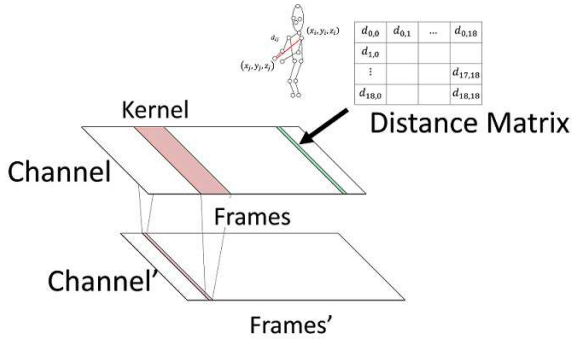


Figure 7: Illustrative example of processing temporal distance matrices with 1D convolutions. Each column represents a pose formed by the flattened upper-triangle of a distance matrix. Each output column is computed from the input column and neighboring columns which are determined by the width of the kernel.

rotation, and has a unique representation for each pose.

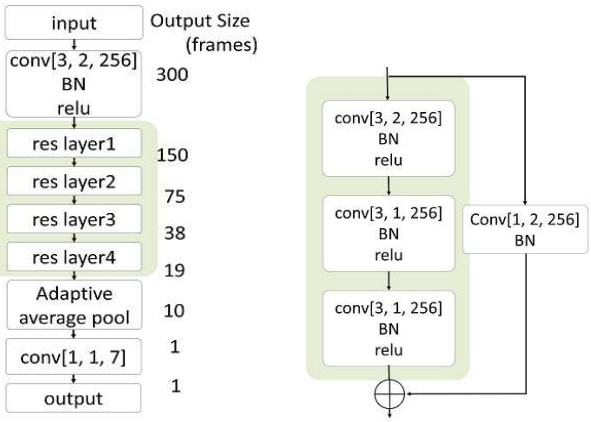
We compute each element $d_{i,j}$ in the distance matrix by computing the Euclidean distance between the i-th and j-th joints:

$$d_{i,j} = \sqrt{(x_i - x_j)^2 + (y_i - y_j)^2 + (z_i - z_j)^2} \qquad (1)$$

Flattening it to a vector representation, it contains, for a $N \times N$ distance matrix, $\frac{(N-1)(N-2)}{2}$ unique elements. In the case of our 3D pose with 19 joints, each can be represented by a 171-dimensional vector.

4.4. Classification Model

Once we compute the distance matrix for the pose of each frame, we can concatenate all the matrices for all the frames in a video to obtain a matrix, where each column represents a pose and the y-axis represents time. To perform classification with this input, we rely on one-dimensional convolutions. The distance-matrix features represent the different channels in the data.



a. proposal network structure b. res layer structure

Figure 8: Our network model. Left: the network as a whole. Right: structure of the res layer. The output size is 1 because of the adaptive average pooling before the last convolutional layer. For the convolution layers, we indicate in brackets the three main parameters: [kernel size, stride, channel].

An overview of this is shown in Fig. 7. Our model is based on ResNet [10] with a recursive structure. Fig. 8 shows our network.

4.5. Data Augmentation and Training

We train our model with a fixed number of frames. In order to improve the generalization, we sample a variable number of frames and use linear interpolation to convert it to the fixed number of frames, resulting in either a slow-down or speed-up of the video. This helps improve performance given that different individuals move at different speeds.

We also sample the training videos so that the first frame is in a canonical pose, i.e., the subject is standing, at the beginning of a squat. We do this by computing the knee angle between ankle and hip. If it is 150 degrees or more, we consider the subject to be in the canonical pose, and sample the fixed number of frames after the canonical pose. (See Fig. 9.)

We train our model using the AdaDelta algorithm [33] and employ a cross-entropy loss.

5. Experiments

We perform ablation experiments and quantitative comparison of our proposed approach with existing methods. Unless specified otherwise, we use the AdaDelta [33] optimizer with a batch size of 16 videos. We train for 8,000 iterations.

Table 5: Comparison with existing approaches on the Background Change dataset. We compare a 3D ResNet50 [10], Non-Local Networks [32], a 1D Resnet50 using Distance Matrices (DM), and our proposed approach. All approaches are using 128 frame inputs.

Input Network	Video Resnet50	Video Nonlocal	DM 1D-Resnet50	DM Ours
Accuracy	64.66	69.48	74.30	75.00

Table 6: Comparison of different hyper parameters on the Single Individual dataset. We compare a linear SVM approach with various CNN configurations. Frames indicates the length of the video used during training. Smoothing indicates whether or not a median filter is used for smoothing. Standing indicates whether or not the first frame shows the individual in a standing position.

Approach	SVM	CNN	CNN	CNN
Frames	200	200	180-220	180-220
Smoothing	No	No	No	3 frames
Standing	No	No	No	Yes
Accuracy	53.84	74.22	79.49	81.05

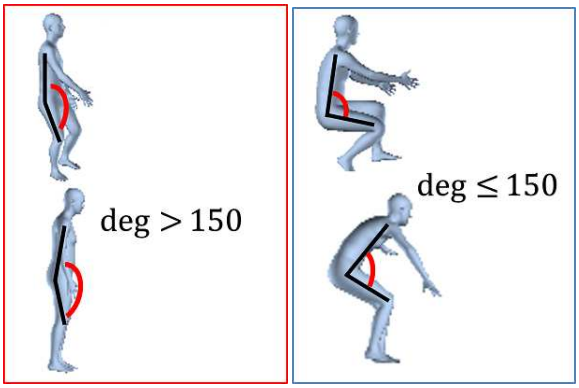


Figure 9: In training, start with standing position(left). Right shoulder, right hip, right knee degrees should be more than 150.

Table 7: Comparison of different types of normalization on the Multiple Individual and YouTube datasets. For all comparisons we use a CNN training with 180-220 frames with 3 frame smoothing and videos starting from a standing position.

Normalization	None	Head	Body	Full
Accuracy	87.84	87.16	88.93	88.86
Accuracy (YT)	73.91	67.08	78.26	77.02

5.1. Quantitative Comparison

We compare our method against existing approaches for video classification on the Background Change dataset.The results are shown in Table 5. We compare against a 3D variant of ResNet50 [10], Non-Local Neural Networks [32], and a 1D variant of ResNet50 trained using distance matrices. The 3D variant of ResNet50 is that of [31], while we replaced all the 2D convolutions with 1D convolutions for the 1D variant of ResNet50. For comparison purposes, we use the same training conditions as [32] for all approaches, such as training with Stochastic Gradient Descent instead of AdaDelta. Both video-based approaches fail to generalize to a diversity of backgrounds and individuals. On the other hand, using a distance matrix feature representation is invariant to the scene background, and shows higher generalization performance. Our proposed model also outperforms the 2D variant of ResNet50.

5.2. Training Hyperparameters

We show the result of ablation of different training parameters in Table 6. In particular, we compare randomizing the input frames, temporal smoothing, and forcing the videos to start with a standing position (Figure 9). We also compare with a linear SVM baseline. For all evaluations we use the distance matrices as input. We can see that randomizing the number of video frames gives a significant improvement. Further adding temporal smoothing and making the videos start with a standing position gives a further increase in performance.

5.3. Evaluation of the Normalization

We compare the effect of the different types of normalization on the Multiple Individual and the YouTube datasets in Table 7. For all comparisons, we use the best performing model, i.e., a CNN trained with randomized frames with temporal smoothing and videos starting from a standing position. We see that the YouTube dataset is significantly tougher than the other datasets, and the body normalization gives a significant improvement in that case, given the larger variety of individuals.

5.4. Visualization

We use Grad-cam [28] to perform a temporal and positional analysis of the proposed approach. The result is shown in Fig. 10. The frames showing first squat command high attention, while standing position between squats seems to attract much less attention.

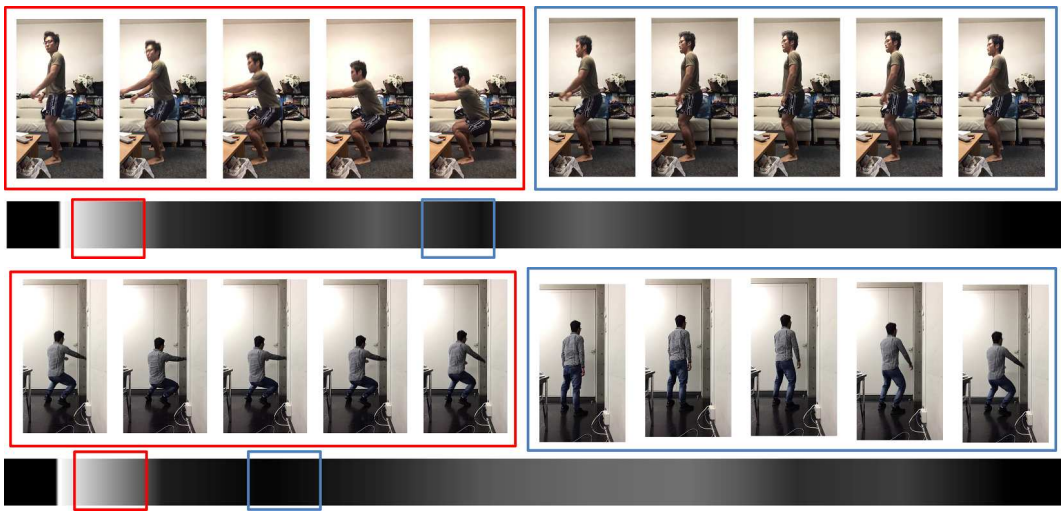


Figure 10: Attention map visualization by Grad-cam[28]. Lighter part in the band below the frames indicate higher attention.

5.5. Limitations

Failure cases in our approach are shown in Fig. 11. In this approach, failure in pose estimation can lead to failure of classification. The distinction between warped and round back seems especially hard because there is no keypoint in the middle of the back. We think this can be remedied by changing the placement of the keypoints. Also, inward bending knees can be missed because it requires the instantaneous motion of the knees at the crucial moment to be correctly detected. Other cases of failure tend to be when the person is too far way or when only part of the body is visible. To show the importance of the accuracy of the 3D pose estimation, we add Gaussian noise to the result of pose estimation in each frame. We train using the Single Individual dataset, and add the noise randomly to the training, validation, and test splits. The pose estimator we use, HMR [13], normalizes each joint coordinate to $[-1, 1]$, and we show the results in Fig. 12. We see how accuracy gets worse as noise becomes larger. It is most sensitive for the 'Inward Knees' and the 'Frontal Knee' classes. This is likely due to the importance few joints (knees) have for these classes.

6. Conclusion

We have presented a new dataset for video recognition of squat form that captures a diversity of users and backgrounds, and a new method based on temporal distance matrices that shows favourable performance with respect to existing approaches. By using the 3D pose, normalizing for user-specific characteristics and using a translation and rotation invariant representation, we show that our approach generalizes much better to other real world data. Although we have focused on squat form detection as a problem, our proposed approach is amenable to any human-centric video classification problem.

References

[1] S. Abu-el haija, J. Lee, P. Natsev, and G. Toderici. YouTube-8M : A Large-Scale Video Classification Benchmark. CoRR, abs/1609.0, 2016.

[2] S. Bianco, F. Tisato, D. Dipartimento, and S. Comunicazione. Karate Moves Recognition from Skeletal Motion. Three-Dimensional Image Processing (3DIP) and Applications 2013, 8650:1–10, 2013.

[3] J. Carreira and A. Zisserman. Quo Vadis, action recognition? A new model and the kinetics dataset. Proceedings - 30th IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2017, 2017-Janua:4724–4733, 2017.

[4] O. Çeliktutan, C. B. Akgül, C. Wolf, and B. Sankur. Graph-Based Analysis of Physical Exercise Actions. MIIRH '13 Proceedings of the 1st ACM international workshop on Multimedia indexing and information retrieval for healthcare, pages 23–32, 2013.

[5] C. Cheron, Guilhem and Laptev, Ivan and Schmid. P-CNN : Pose-based CNN Features for Action Recognition. The IEEE International Conference on Computer Vision (ICCV), pages 3218–3226, 2015.

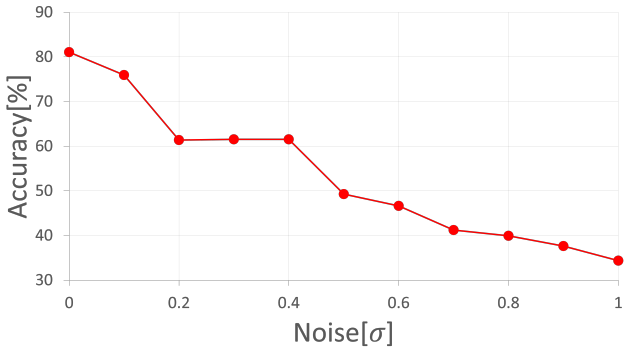
[6] H. Doughty, D. Damen, and W. Mayol-Cuevas. Who's Better, Who's Best: Skill Determination in Video using Deep Ranking. CoRR, abs/1703.0:6057–6066, 2017.

[7] C. Feichtenhofer, A. Pinz, and A. Zisserman. Convolutional Two-Stream Network Fusion for Video Action Recognition. The IEEE Conference on Computer Vision

Video 1 Video 2

Figure 11: Failure cases. In the left video, Warped Back is detected even though the back is in fact round. This is mad difficult because there is no keypoint in the middle of the back. In the right video, Good Squat is the result even though the knees are bending inwards. This is the result of failing to correctly place the right knee key point at the crucial frame when the knee bends inwards.

Figure 12: Result of adding noise to the pose estimation. Accuracy gets worse as the noise becomes bigger.



and Pattern Recognition (CVPR), pages 1933–1941, 2016.

[8] B. Ghanem, J. C. Niebles, C. Snoek, F. C. Heilbron, H. Alwassel, V. Escorcia, R. Krishna, S. Buch, and C. D. Dao. The ActivityNet Large-Scale Activity Recognition Challenge 2018 Summary. CoRR, abs/1808.0, 2018.

[9] T. P. H. Kuehne, H. Jhuang, E. Garrote and T. Serre. HMDB : A Large Video Database for Human Motion Recognition. 2011 International Conference on Computer Vision, pages 2556–2563, 2011.

[10] J. He, Kaiming and Zhang, Xiangyu and Ren, Shaoqing and Sun. Deep Residual Learning for Image Recognition. In The IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pages 770–778, 2016.

[11] C. Ionescu, D. Papava, V. Olaru, and C. Sminchisescu. Human3.6M: Large Scale Datasets and Predictive Methods for 3D Human Sensing in Natural Environments. IEEE Transactions on Pattern Analysis and Machine Intelligence, 36:1325–1339, 2013.

[12] U. Iqbal, M. Garbade, and J. Gall. Pose for Action - Action for Pose. Proceedings - 12th IEEE International Conference on Automatic Face and Gesture Recognition, FG 2017 - 1st International Workshop on Adaptive Shot Learning for Gesture Understanding and Production, ASL4GUP 2017, Biometrics in the Wild, Bwild 2017, Heteroge, pages 438–445, 2017.

[13] A. Kanazawa, M. J. Black, D. W. Jacobs, and J. Malik. End-to-end Recovery of Human Shape and Pose. In The IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pages 7122–7131, 2018.

[14] W. Kay, J. Carreira, K. Simonyan, B. Zhang, C. Hillier, S. Vijayanarasimhan, F. Viola, T. Green, T. Back, P. Natsev, M. Suleyman, and A. Zisserman. The Kinetics Human Action Video Dataset. 2017.

[15] V. Kazemi, M. Burenius, H. Azizpour, and J. Sullivan. Multi-view Body Part Recognition with Random Forests. BMVC, pages 48.1–48.11, 2014.

[16] M. Loper, N. Mahmood, J. Romero, G. Pons-Moll, and M. J. Black. SMPL : A Skinned Multi-Person Linear Model. ACM Transactions on Graphics (TOG), 34(6), 2015.

[17] D. C. Luvizon, D. Picard, and H. Tabia. 2d/3d pose estimation and action recognition using multitask deep learning. In The IEEE Conference on Computer Vision and Pattern Recognition (CVPR), June 2018.

[18] J. Martinez, R. Hossain, J. Romero, and J. J. Little. A Simple Yet Effective Baseline for 3d Human Pose Estimation. Proceedings of the IEEE International Conference on Computer Vision, 2017-Octob:2659–2668, 2017.

[19] D. Mehta, H. Rhodin, D. Casas, P. Fua, O. Sotnychenko, W. Xu, and C. Theobalt. Monocular 3D human pose estimation in the wild using improved CNN supervision. Proceedings - 2017 International Conference on 3D Vision, 3DV 2017, pages 506–516, 2018.

[20] F. Moreno-noguer and I. D. Rob. 3D Human Pose Estimation from a Single Image via Distance Matrix Regression. The IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2017.

[21] B. X. Nie, C. Xiong, and S. C. Zhu. Joint action recognition and pose estimation from video. Proceedings of the IEEE Computer Society Conference on Computer

Vision and Pattern Recognition, 07-12-June:1293–1301, 2015.

[22] G. I. Parisi, S. Magg, and S. Wermter. Human Motion Assessment in Real Time using Recurrent Self-Organization. 2016 25th IEEE International Symposium on Robot and Human Interactive Communication (RO-MAN), pages 71–76, 2016.

[23] P. Parmar and B. T. Morris. Learning to Score Olympic Events. In IEEE Computer Society Conference on Computer Vision and Pattern Recognition Workshops, volume 2017-July, pages 76–84, 2017.

[24] H. Pirsiavash, C. Vondrick, and A. Torralba. Assessing the Quality of Actions. Computer Vision – ECCV 2014, 8694, 2014.

[25] A. I. Popa, M. Zanfir, and C. Sminchisescu. Deep multitask architecture for integrated 2D and 3D human sensing. Proceedings - 30th IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2017, 2017-Janua:4714–4723, 2017.

[26] M. D. Rodriguez, J. Ahmed, and M. Shah. Action MACH A Spatio-temporal Maximum Average Correlation Height Filter for Action Recognition. 2010.

[27] H. Sak, A. Senior, and F. Beaufays. Long Short-Term Memory Recurrent Neural Network Architectures for Large Scale Acoustic Modeling. IEEE Access, 6:15733–15742, 2018.

[28] R. R. Selvaraju, M. Cogswell, A. Das, R. Vedantam, D. Parikh, and D. Batra. Grad-CAM: Visual Explanations from Deep Networks via Gradient-Based Localization. Proceedings of the IEEE International Conference on Computer Vision, 2017-Octob:618–626, 2017.

[29] A. Simonyan, Karen and Zisserman. Two-Stream Convolutional Networks for Action Recognition in Videos. Advances in Neural Information Processing Systems 27, pages 568–576, 2014.

[30] K. Soomro, A. R. Zamir, and M. Shah. UCF101: A Dataset of 101 Human Actions Classes From Videos in The Wild. CoRR, (November), 2012.

[31] D. Tran, L. Bourdev, R. Fergus, L. Torresani, and M. Paluri. Learning spatiotemporal features with 3D convolutional networks. Proceedings of the IEEE International Conference on Computer Vision, 2015 Inter:4489–4497, 2015.

[32] X. Wang, R. Girshick, A. Gupta, and K. He. Non-local Neural Networks. In 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition, pages 7794–7803, 2018.

[33] M. D. Zeiler. ADADELTA: An Adaptive Learning Rate Method. CoRR, abs/1212.5, 2012.