Skeleton-Based Methods for Speaker Action Classification on Lecture Videos

Fei Xu(✉), Kenny Davila, Srirangaraj Setlur, and Venu Govindaraju

Department of Computer Science and Engineering, University at Buffalo, Buffalo, NY, USA
{fxu3,kennydav,setlur,govind}@buffalo.edu

Abstract. The volume of online lecture videos is growing at a frenetic pace. This has led to an increased focus on methods for automated lecture video analysis to make these resources more accessible. These methods consider multiple information channels including the actions of the lecture speaker. In this work, we analyze two methods that use spatio-temporal features of the speaker skeleton for action classification in lecture videos. The first method is the AM Pose model which is based on Random Forests with motion-based features. The second is a state-of-the-art action classifier based on a two-stream adaptive graph convolutional network (2S-AGCN) that uses features of both joints and bones of the speaker skeleton. Each video is divided into fixed-length temporal segments. Then, the speaker skeleton is estimated on every frame in order to build a representation for each segment for further classification. Our experiments used the AccessMath dataset and a novel extension which will be publicly released. We compared four state-of-the-art pose estimators: OpenPose, Deep High Resolution, AlphaPose and Detectron2. We found that AlphaPose is the most robust to the encoding noise found in online videos. We also observed that 2S-AGCN outperforms the AM Pose model by using the right domain adaptations.

Keywords: Action classification · Lecture video analysis · Pose estimation · Lecture video dataset

1 Introduction

Today, the number of online lecture videos is growing faster than ever. These videos are becoming a valuable educational resource for both teachers and students globally. Easy and efficient access to specific topics within the massive amount of lecture videos is enabled by applications for summarization, navigation, and retrieval of lecture video content. Methods for automated lecture video analysis facilitate these applications by extracting information from multiple channels including scene text, supplementary lecture materials, audio, transcriptions and speaker actions. In particular, many lecture videos feature handwriting of content on traditional whiteboards/chalkboards, and speaker actions such as writing or erasing can be used to facilitate the extraction of such content [1].

A. Del Bimbo et al. (Eds.): ICPR 2020 Workshops, LNCS 12664, pp. 250–264, 2021.
https://doi.org/10.1007/978-3-030-68799-1_18

In these scenarios, the development of accurate classifiers for speaker actions in lecture videos is important for advancing the field of lecture video analysis.

In this work, we analyze two methods that use spatio-temporal features of the speaker skeleton for action classification in lecture videos. The first method is based on our previous work, AM Pose [1], which uses Random Forests and motion-based features to classify speaker actions for further lecture video summarization on the AccessMath dataset [2]. The second method is a generic action classifier [3] which uses a two-stream adaptive graph convolutional network (2S-AGCN).

We adopt a generic framework which splits the video into small temporal segments, which we call action segments, for speaker action classification (See Fig. 1). The pose of the speaker is estimated on every frame using one of four state-of-the-art methods for pose estimation: OpenPose [4], Deep High Resolution (DHR) pose estimator [5], Detectron 2 [6], and AlphaPose [7]. Then, all poses are normalized and a feature representation is built for every action segment according to the corresponding action classifier model.

Our experiments are based on two lecture video datasets for which the speaker actions have been annotated. The first one is the AccessMath dataset [2], which has linear algebra lectures by one speaker. We significantly extended this dataset by adding videos from multiple online sources to built the second dataset. These annotations will be made publicly available.

During the course of this study, we attempt to answer the following research questions: **RQ1.** What is the performance of different state-of-the-art pose estimators when applied to online lecture videos? **RQ2.** How can we adapt generic skeleton-based action classification methods to the lecture video domain? **RQ3.** Can the 2S-AGCN model perform better than the AM Pose method and to what degree?

2 Background

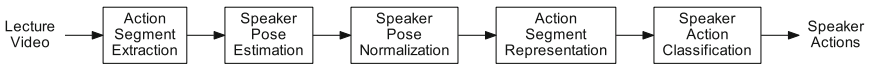


Fig. 1. Summary of our lecture video speaker action classification framework. The process starts with a lecture video which is logically divided into action segments. The pose of the speaker is estimated and normalized in every video frame. A representation is built for all speaker poses in each action segment. A classifier is then used to determine the action of the speaker in every action segment.

Our work addresses the topic of skeleton-based action classification in lecture videos. In this section, we briefly describe vision-based human pose estimation methods which can be used to get the speaker skeleton in a lecture video. We then discuss skeleton-based methods for action classification. Finally, we review recent works on automated lecture video analysis.

2.1 Vision-Based Human Pose Estimator

A vision-based human pose estimator (HPE) is a method used to extract pose features from still images and videos. Chen et al. [8] describe three types of human body models that can be used in HPE: skeleton-based, contour-based, and volume-based. Here, we concentrate on skeleton-based pose representations.

Depending on how the estimation starts, Chen et al. [8] categorize pose estimators into two families: bottom-up and top-down. The bottom-up methods estimate all parts of the human body from input images and assemble them into different skeletons by fitting human body models. One example is OpenPose [4] which detects body parts and represents them using location and direction information stored in 2D vector fields. Full skeletons are produced in a greedy fashion by locally matching connected body parts (e.g. elbow candidates to hand candidates), by maximizing their confidence using the Hungarian method [9]. This is prone to failure when people are partially occluded in the image. Our previous work [1] uses OpenPose for speaker action classification on lecture videos.

The family of top-down methods, first generate bounding boxes of all people using object detectors and then extract human pose representations for each person detected. The Deep High-Resolution (DHR) learning method [5] uses multiple parallel sub-networks to produce high resolution heatmaps of human pose keypoints. The Detectron 2 method [6] uses mask R-CNN [10] to get the bounding boxes of human objects and also to detect human pose key-points within these boxes. The AlphaPose model [7] uses a regional multi-person pose estimation which deals with noisy detections of human objects. It uses a spatial transform network [11] to improve the object proposals that are fed to the single person pose estimator [12]. In general, top-down methods are likely to be slower than bottom-up methods when the image contains a large number of people [8].

Our work focuses on analyzing lecture videos where the upper body of the speaker is visible most of the time. We have considered the four pose estimators described here for speaker pose estimation in our work (see Sect. 5.1).

2.2 Action Classification

Human actions can be represented by sequential human pose changes. Some recent works address video-based action representations in different ways. Nguyen et al. [13] fuse salient maps from multiple prediction models first, and construct video representation using Spatial-Temporal Attention-aware Pooling (STAP). Yi et al. [14] combine appearance and motion saliencies to extract most salient trajectories, which are encoded by Bag of Features (BoF) or Fisher Vector (FV) as action features. In the work by Wang et al. [15], trajectories are constructed based on skeleton joints and projected as 2D images which are used for action classification by Convolutional Neural Network. The works by Yan et al. [16] and Shi et al. [3] construct skeleton-based graph representations which contain the spatial-temporal information for actions in videos. In this work, we consider action classification methods using skeleton-based action representations.

The 2s-AGCN model [3] uses adaptive graph convolutional networks with two spatio-temporal graph representations, one for joints and one for bones. For each representation, three adjacency matrices are required to decide the graph topology. The first matrix represents the physical connections from the skeleton. The second and third matrices make the graph topology adaptive because they are learned from the input data, and they represent the existence and strength of connections between any two joints/bones.

In our previous work, AM Pose [1], we used motion-based features and Random Forests for speaker action classification. These features are generated from joints and bones of the speaker skeleton, and they were designed based on observations of the speaker lecturing behavior. However, the AccessMath dataset [2] on which AM Pose was evaluated, has only one speaker. Our novel expansion includes multiple speakers. It is possible that for the same action some speakers behave differently (e.g. handedness, gestures, etc.). In this work, we take AM Pose as the baseline method, and we compare its classification accuracy against the 2S-AGCN on both the AccessMath dataset as well as our extended dataset.

2.3 Lecture Video Analysis

Lecture video analysis applications include lecture content summarization, indexing, search and navigation. Based on the type of data used for lecture video analysis, we can broadly categorize existing works into three groups.

The first group uses scene text appearing in lecture videos. Ma et al. [17] use various visual features to remove speakers/audience shots and extract slide/board frames from lecture videos for indexing. Davila and Zanibbi [18] extract keyframe-based lecture video summaries from single-shot lecture videos, and propose the Tangent-V visual search engine [19] which can use rendered LaTeX from course notes as query images to search handwritten formulas in these keyframes and vice versa. Kota et al. [20] propose a framework to detect and binarize handwritten whiteboard content for lecture video summarization based on keyframes.

The second group does not rely on scene text. Instead, it uses other information channels from lecture videos such as audio or transcripts. Soares and Barrére [21] use audios of lecture videos to generate fully voiced audio chunks, and the corresponding textual and acoustic features are used with genetic algorithms for video segmentation. Shah et al. [22] use lecture video transcripts and Wikipedia texts to detect the boundaries of lecture topics for video segmentation.

The third group uses both scene text and information from other channels. Some works [23,24] combine the speech from the instructor with the extracted lecture video slides for video retrieval. Xu et al. [1] use speaker action classification for whiteboard lecture video segmentation and keyframe selection. The handwritten content on these keyframes is then extracted for summarization. In this work, we focus on improving speaker action classification on whiteboard and chalkboard lecture videos for more effective lecture video analysis.

3 Methodology

In our current work, we use a generic framework for skeleton-based speaker action classification on lecture videos. The framework (illustrated in Fig. 1) uses small fixed-length temporal segments, called action segments, where we assume that the speaker is performing only one action. The input is a lecture video and the output is the predicted speaker action for every action segment of the video. First, we start by estimating the skeleton-based pose of the speaker on every video frame. We then normalize the pose estimations and create a representation for each action segment according to the selected method for action classification. We consider two skeleton-based speaker action classification models: AM Pose [1] and 2S-AGCN [3].

Assumptions. Each lecture video has only one instructor whose upper body is almost always visible and performs one action at any moment of the lecture. The video is recorded using a stationary camera without focus changing. The audience, if any, is never visible on the video. Finally, the lecture is based on handwritten content on a whiteboard/chalkboard.

3.1 Action Segment Extraction

The basic unit used for classification of actions in our framework is the action segment. In order to train and test our approach, we must extract the action segments of each lecture video. During normal runtime, the input can be simply divided into small sequential temporal segments of fixed length. Following the previous work AM Pose [1], we have currently fixed this length to 15 frames which roughly corresponds to half a second for all videos in our datasets.

In this work, we use 4 overlapping tracks of action segments both for training and testing videos, as illustrated in Fig. 2. The ground truth of the video contains the beginning and ending frames of each speaker action. When we sample a given action segment, we assign to it the label of the majority class of all frames contained within that segment.

3.2 Speaker Pose Estimation

Pose refers to the visual position and orientation of the human body at a given moment and can be represented in multiple ways as described in Sect. 2.1. In this work, we are concerned with skeleton-based pose estimations. Figure 3 shows different examples of speaker poses estimated using AlphaPose [7]. In Fig. 3a and 3c, the speaker face is partially or totally invisible. Figure 3b shows the joints of the upper body enumerated in the COCO-18 format [25]. We consider 4 different pose estimators in this work: OpenPose [4], DHR [5], Detectron2 [6], and AlphaPose [7]. An empirical comparison of these methods is described in Sect. 5.1.

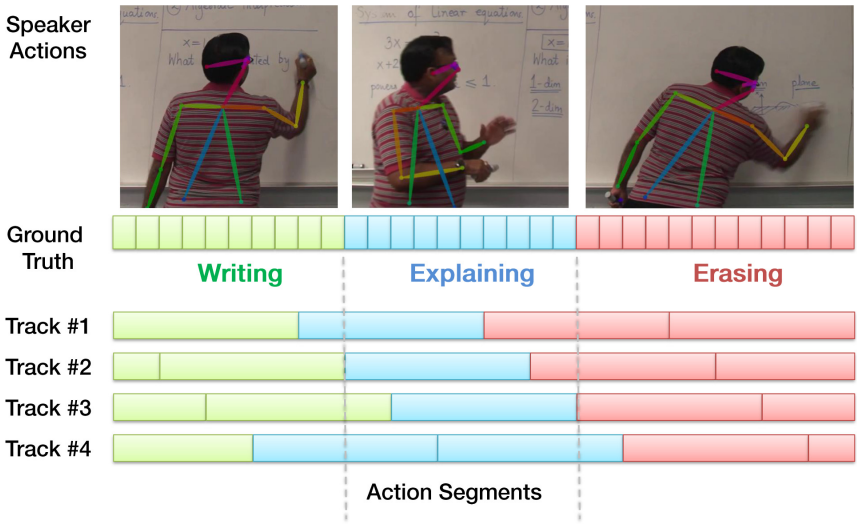
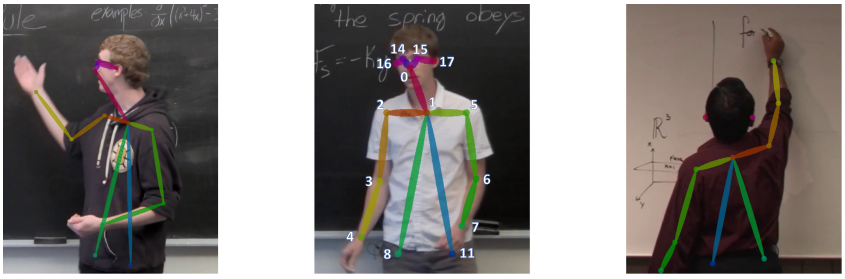


Fig. 2. Action segments used for speaker action classification. The action of the speaker, shown at the top, are labeled by frame ranges. Then, we extract one or more overlapping tracks of action segments of a fixed length (15 frames). The label of an action segment is decided using the majority class of the frames it covers.



(a) Side (b) Front (c) Back

Fig. 3. Examples of skeleton data produced by the AlphaPose system [7] on lecture videos. We illustrate different poses such as the (a) side, (b) front and (c) the back. In (b), we have also included the original joint numbers, based on COCO-18 format [25].

3.3 Speaker Pose Normalization

Lecture videos are recorded in a variety of environments and video resolutions resulting in various ranges for the absolute speaker pose coordinates for each lecture video. In order to improve the robustness of speaker action classification, we normalize these coordinates using an affine transformation (translation and scaling) for each action segment.

First, we recognize that many frames within a given action segment might have invalid or missing pose estimations. This can be due to either detection errors or the speaker being out of the image in that frame. We identify the first valid frame where the pose estimator captures a skeleton, and we use the nose joint as the origin point to translate all coordinates from all valid pose predictions in that segment. For invalid poses, a special value is assigned to all joint coordinates.

The second normalization step is scaling, which follows the same principle as AM Pose [1], which first computes a scaling factor based on the average size of a given normalization bone through the entire video. We extend this idea by considering the average size of two bones instead of just one. In particular, we use the mean of global average distances between the neck to the left and right hip on pose estimators using the COCO-18 format [25] (see Fig. 3). All coordinates are normalized by the scaling factor.

3.4 Action Segment Representation

We consider two representations for speaker action segments in this paper. The first is based on the motion-based heuristic features from AM Pose [1]. The second uses two spatio-temporal graphs, one for joints and one for bones, of the skeletons in the action segment [3].

Motion-Based Feature Representation. As in our previous work, AM Pose[1], we select joints and pairs of joints which are related to the speaker actions. In one action segment, joint-wise features are constructed for every pair of consecutive frames using the means, medians and covariance matrix for joint displacement and means for joint distances. In addition, pair-wise features including means and variances of displacement between two joints are computed on every frame of the segment. For both types of selected features, we omit cases where either of the joints coordinates is invalid and add a confidence value that represents the percentage of valid joints in the action segment. Unlike AM Pose, this work considers lecture videos with left handed speakers as well. For this reason, we consider two versions of this representation. The first one uses joints from the head, torso and right upper limb (RUL) for a total of 71 features and the second considers head, torso and both upper limbs (BUL) for a total of 106 features.

Graph-Based Representation. This is the representation used by Shi et al. [3] in 2S-AGCN, which considers two types of features to represent both spatial and temporal information of action segments. The first one is joints data from pose estimation. While the original 2S-AGCN work assumes 3D poses and up to two people in the image, here we only use 2D pose estimation for a single person. The second feature is the bones in the skeleton, where bones represent physical connections between two joints. For the two joints of a bone, the one closer to the central point is the source joint and the other is the target joint. In our work, the neck is the central joint (point 1 in Fig. 3b). The displacement from source joint to target joint is used to represent the bone as a vector. For each of these

two features, a spatio-temporal graph is constructed for every action segment, where the graph nodes are represented by corresponding joint or bone features, and the graph edges are the spatial and temporal connections between nodes.

Based on the observation that the lower limbs of the speakers are rarely visible in the videos, we consider 3 variations for the spatio-temporal graph of action segment. The first variation, full body, uses all joints on the speaker skeleton, while the other two use upper body joints including only the RUL or BUL.

3.5 Speaker Action Classification

In this work, we consider 8 speaker actions which are relevant to the handwritten content on whiteboards/chalkboards in lecture videos: write, pick eraser, erase, drop eraser, out, out-writing, out-erasing, and explain. Actions related to writing and erasing are key to changes in the content on the board. During the *explain* action, the speaker will mostly move around and use different gestures to emphasize important written content. The *out* action refers to the case when speaker is not visible in the image, and *out-writing/erasing* represent the hardest case to analyze where the speaker is mostly invisible but the handwritten content around the image edge is changed.

We have considered two different approaches for speaker action classification. The first is AM Pose [1] which combines the motion-based feature representation with Random Forests. The second is the 2S-AGCN model [3] which uses adaptive graph convolutional networks with spatio-temporal graph-based representations generated from both joints and bones of speaker skeleton data.

4 Datasets

We use two datasets for our experiments: AccessMath [2] and a novel extension. The speaker actions distributions on these datasets are shown in Table 1.

4.1 AccessMath

AccessMath is a set of 20 linear algebra lecture videos recorded in different classrooms with a single speaker. These videos are 1080p at 29.97 FPS and their average length is 49 min. For lecture video summarization, Davila et al. [18] used 12 videos (around 10 h) from the AccessMath dataset, with 5 videos for training (around 3.5 h) and 7 for testing. In AM Pose [1], we only annotated actions of the speaker for the 5 training videos.

4.2 Extended Dataset

We extended the original AccessMath dataset by annotating the intervals of speaker actions on lecture videos from multiple online sources. This dataset has 28 whiteboard videos and 6 chalkboard videos, for a total of 34. The resolutions

of these lecture videos vary from 480p to 1080p. The video lengths vary in the range of 2.4 to 67.3 min with an average of 29.4 min. Overall, the dataset has 16.7 h of lecture videos. A total of 14 videos are from the AccessMath dataset, including the 5 training videos annotated in AM Pose.

The dataset includes 8 speakers, 6 of whom are right-handed and the other 2 are left-handed. For each speaker, half of the videos are for training and the other half are for testing. In this way, we have 17 training videos and 17 testing videos in total.

Most video frames in this dataset show the speaker lecturing. However, many videos from online sources include opening and ending credits, which are mostly text-based. We annotate such frames as *out* because speakers are not in the image.

These lecture videos are collected from YouTube channels, including the online version of the AccessMath dataset[1]. In order to ensure that all action segments (see Sect. 3.1) represent the same video length, we re-encoded all videos to 29.97 FPS. However, we notice that the original AccessMath videos and their corresponding online versions have different video length and image quality. To ensure that the annotations of the original videos are still compatible with the online versions, we use the video lengths of both versions to automatically synchronize the annotations to the online video. The annotations of the extended dataset will be publicly available[2].

Table 1. Distribution of the classes of Speaker Actions considered in this work. We consider both the AccessMath dataset and the extended dataset.

Action	AccessMath	Extended dataset	
	Training (%)	Training (%)	Testing (%)
Drop Eraser	0.8558	0.7523	0.8347
Erase	7.4659	5.1220	5.9263
Explain	25.7790	43.6089	44.9741
Out	25.6347	15.9167	13.8559
Out Erasing	0.0471	0.0537	0.0184
Out Writing	0.3163	0.2375	0.0022
Pick Eraser	0.8374	0.7371	0.9091
Write	39.0637	33.5719	33.4793

5 Experiments

In order to answer the research questions mentioned in Sect. 1, we ran three experiments: pose estimator selection (**RQ1**), adaptations of the 2S-AGCN [3]

[1] https://www.youtube.com/playlist?list=PLg2YxOqXd_2Ptnj2adKJRngjD1TK7fAo5.

[2] https://kdavila.github.io/lecturemath/.

model for speaker action classification (**RQ2**), and finally a comparison between AM Pose [1] and 2S-AGCN (**RQ3**).

Table 2. Pose Estimator Selection. Cross-validation results for speaker action classification on AccessMath training videos using 4 different pose estimators as well as both the original and the noisier re-encoded videos from YouTube. Action classification is performed with AM POSE [1] using data from upper body and the right upper limb (RUL) or both upper limbs (BUL).

	Original		YouTube	
	RUL	BUL	RUL	BUL
OpenPose 1.5.1 [4]	83.59	84.01	82.72	83.15
DHR [5]	84.65	84.62	82.94	83.09
Detectron2 [6]	84.81	84.90	83.76	83.85
AlphaPose [7]	**85.23**	**85.57**	**84.23**	**84.32**

5.1 Pose Estimator Selection

To determine what is the performance of different state-of-the-art pose estimators on online lecture videos (**RQ1**), we evaluated four models: Openpose 1.5.1 [4], DHR [5], Detectron2 [6] and AlphaPose [7]. In addition, we used two versions of the AccessMath dataset [2] videos: the original and the online (YouTube). We compared the results of the original AM Pose [1] which considers features from RUL of the speaker, and the extended version which includes features from BUL (see Sect. 3.4). Following the experimental methodology from our previous work [1], we used cross-validation over the training videos.

The speaker action classification results for different pose estimators, feature sets and video sources are shown in Table 2. Except in the case when DHR is used on the original AccessMath videos, using features from BUL for action classification performs better than just considering the RUL. We observed that the additional features proposed in this work help to differentiate actions such as *out*, *explain* and *write*.

For all of the pose estimators evaluated here, the speaker action classification accuracy is lower on the YouTube version than on the original. In this case, when parts of the speaker body move too fast, they become blurred on the image. This causes errors in the pose estimation such as missing speaker skeletons. This can likely be explained by the fact that the tested pose estimators are not trained with this kind of image noise. The missing estimations can make the corresponding action segments to be incorrectly classified as *out*, which affects the classification accuracy. We observed that the average size of YouTube videos is 368 MB, and the average size of the originals is 4.28 GB.

We found that AlphaPose performs better than the other three pose estimators in terms of cross-validation action classification accuracy on both versions of the videos. In the following two experiments, we used AlphaPose to generate pose estimations.

5.2 Adaptations of the 2S-AGCN Model for Speaker Action Classification

To determine what would be required to adapt a generic skeleton-based action classifier to the lecture video domain (**RQ2**), we tested different domain adaptations using the 2S-AGCN model [3]. We evaluated the effect of pose normalization (see Sect. 3.3), using three types of pose data: raw pose, pose normalized by translation, and pose normalized by both translation and scaling. We also tested the effectiveness of each of the two streams for the adaptive graph convolutional network (AGCN), and we considered three different graph representations: full skeleton, upper body with RUL only and upper body with BUL. The second representation is used to make a fair comparison between AGCN and the original AM Pose. Same as our previous experiment, we also followed the same cross-validation protocol from AM Pose [1], but we only used the online version of the AccessMath dataset training videos.

In all conditions, the network was trained for a total of 16 epochs, with batch size of 64 and initial learning rate of 0.1 with decay factor of 0.1 every 5 epochs. Note that two streams of AGCN were trained for each cross-validation fold: One for joints, and one for bones. The average of all 5 folds are presented in Table 3. In all cases, using both streams of joints and bones (2S-AGCN) gives better classification result than using either stream alone. As shown in Table 2, using AGCN with joints representation performs better than using AM Pose with motion-based features generated from joints coordinates when the pose normalization is applied. We also observed that normalization helps more on the joint stream than bone stream. Overall, 2S-AGCN using BUL graph representations with pose translation gives the best result. In the third experiment, since we used a larger lecture video dataset with different speakers in different recording environments, we used both translation and scale for pose normalization.

Table 3. Cross-validation results for speaker action classification using the 2S-AGCN model [3]. We compare the effect of different pose normalization settings and 3 different graph representations

		Pose normalization		
Streams	Graph (# Joints)	None	Translation	Translation + Scale
Joints	RUL (12)	83.66	85.20	84.93
	BUL (14)	83.80	85.27	85.13
	Full (18)	84.25	84.84	85.15
Bones	RUL (12)	84.07	83.86	83.85
	BUL (14)	83.21	84.29	84.06
	Full (18)	83.48	83.52	83.39
Joints & Bones	RUL (12)	84.95	85.51	85.26
	BUL (14)	84.52	**85.74**	85.29
	Full (18)	85.31	85.21	85.43

5.3 Classifier: AM Pose Vs 2S-AGCN Model

To determine if the 2S-AGCN model [3] performs better than AM Pose [1] and to what degree (**RQ3**), our final experiment compared the AM Pose and the 2S-AGCN models on the extended dataset (see Sect. 4.2). Results in terms of action classification accuracy for all conditions are shown in Table 4.

The distribution of the speaker actions (see Table 1) shows that the most common actions on both datasets are *write*, *explain*, *out* and *erase*, and the extended dataset has more *explain* and less of the other three actions. We expected that the classification accuracy using the original AM Pose [1] with RUL would drop on the extended dataset as it has two left-handed speakers. However, while using features from BUL helps to classify actions better than using RUL in AM Pose, the overall difference between these two configurations is small. The per-class classification f-measures of *explain*, *out*, *write* and *erase* are 83.39%, 80.18%, 69.18%, and 4.53% for the left-handed speakers, and 83.37%, 88.32%, 87.11% and 77.61% for the right-handed ones respectively. It can be seen that the classifier performs similarly for both left and right-handed speakers on the *explain* action as it is the most common class (44.97%) in the extended dataset. It also predicts most of the *out* actions. Probably, the non-dominant upper limb provides enough information to classify these two actions. At the same time, the accuracies for the *write* and the *erase* actions are lower for left-handed speakers. In particular, the *erase* action is badly predicted, but this only represents 5.9% of the total actions (see Table 1). Overall, there are only around 2.1 h (out of 16.7 in total) of left-handed speaker videos in the extended dataset.

When the action segment contains a transition from one action to another, we label the action segment by the majority rule (see Sect. 3.1). Approximately 9.12% of the action segments in the test set match this condition. We anticipated that the action classifier might give the other action as the prediction for these action segments. Therefore, we introduced another evaluation standard (majority + secondary) which counts the prediction as correct if it is either of these two actions (see third column of Table 4). For all conditions of both methods, the difference between both evaluation standards is consistently around 3.2%. This difference corresponds to trivial action classification errors. We considered them as trivial because the corresponding action segments map to frames which are annotated with either of these two actions. For example, for AM Pose - BUL, we observed that 41.18% of the action segments containing transitions were incorrectly classified, but 83.74% of these mistakes were indeed trivial errors, and they represent 19.90% of the overall error.

Under the second evaluation standard, the incorrectly predicted action segments represent the non-trivial errors. For example, in the condition of AM Pose - BUL, 10.64% of the *write* actions are predicted as *explain* actions, which represents 22.21% of the overall classification error. On the other side, 8.55% of *explain* actions are classified as *write* actions, which is 23.96% of the total error. These confusions are probably related to the fact that sometimes speakers point at some handwritten content during *explain* and mimic *write* actions for emphasizing the content. These challenging errors represent potential areas where the action classification accuracy can be improved.

Table 4. Final Speaker Action Classification results on the LectureMath dataset. Apart from the **majority classification** result, we consider **secondary classification** for the boundary action segment. AM Pose - RUL follows work of [1], and 2S-AGCN considers both joints and bones spatio-temporal features as work of [3]

Method	Evaluation Mode	
	Majority	Majority + Secondary
AM Pose - RUL	83.57	86.77
AM Pose - BUL	83.97	87.16
Joint-AGCN (14)	84.73	87.94
Bone-AGCN (14)	84.19	87.4
2S-AGCN (14)	**85.46**	**88.68**

Overall, 2S-AGCN gives better result than AM Pose using both evaluation standards, and it also performs better than using either joint or bone alone.

6 Conclusion

In this work, we tried to answer three main research questions. First, to determine what is the performance of different state-of-the-art pose estimators on online lecture videos (RQ1), we compared four methods using cross-validation on the original and online (YouTube) versions of the AccessMath training videos [2]. We found that online lecture videos have encoding noise which caused pose estimators to miss or incorrectly predict poses. We also found that using AlphaPose helped to achieve higher speaker action classification accuracy on both versions of the videos.

Second, to determine what would be required to adapt a generic skeleton-based action classifier to the lecture video domain (RQ2), we tested the 2S-AGCN model [3] following the same cross-validation method with different combinations of graph representations and pose normalization methods. We found that using the graph representation of the speaker upper body normalized by translation can give the best classification accuracy on the YouTube versions of the AccessMath training videos. Consistent with the work by Shi et al. [3], we achieved higher classification accuracy by using both streams of AGCN.

Third, to determine if the 2S-AGCN model performs better than AM Pose and to what degree (RQ3), we extended the AccessMath dataset by annotating speakers actions in 29 lecture videos from different online sources. Then, we ran different configurations for both speaker action classification models on this dataset. We found that AM Pose can classify the speaker actions well, even on the larger lecture video dataset. However, the 2S-AGCN model still performs better than AM Pose by a small margin.

In the future, we intend to improve the action classification on noisy lecture videos, especially for the case where the pose estimator incorrectly predicts the speaker pose. We will continue to expand our lecture video dataset by including

more speakers. We also want to use the 2S-AGCN for lecture video summarization similar to AM Pose [1] on the extended dataset. In addition, we would like to adapt this action classification framework to videos where only the hands of the speaker are visible.

Acknowledgement. This material was partially supported by the National Science Foundation under Grants No. 1640867 (OAC/DMR), and No. 1651118 (SBE).

References

1. Xu, F., Davila, K., Setlur, S., Govindaraju, V.: Content extraction from lecture video via speaker action classification based on pose information. In: 2019 International Conference on Document Analysis and Recognition (ICDAR), pp. 1047–1054. IEEE (2019)
2. Davila, K., Agarwal, A., Gaborski, R., Zanibbi, R., Ludi, S.: Accessmath: indexing and retrieving video segments containing math expressions based on visual similarity. In: 2013 Western New York Image Processing Workshop (WNYIPW), pp. 14–17. IEEE (2013)
3. Shi, L., Zhang, Y., Cheng, J., Lu, H.: Two-stream adaptive graph convolutional networks for skeleton-based action recognition. In: 2019 Conference on Computer Vision and Pattern Recognition (CVPR), pp. 12018–12027. IEEE/CVF (2019)
4. Cao, Z., Hidalgo, G., Simon, T., Wei, S., Sheikh, Y.: Openpose: realtime multi-person 2d pose estimation using part affinity fields. arXiv preprint arXiv:1812.08008 (2018)
5. Sun, K., Xiao, B., Liu, D., Wang, J.: Deep high-resolution representation learning for human pose estimation. In: 2019 Conference on Computer Vision and Pattern Recognition (CVPR), pp. 5693–5703. IEEE/CVF (2019)
6. Wu, Y., Kirillov, A., Massa, F., Lo, W.-Y., Girshick, R.: Detectron2 (2019). https://github.com/facebookresearch/detectron2
7. Fang, H.-S., Xie, S., Tai, Y.-W., Lu., C.: Rmpe: regional multi-person pose estimation. In: 2017 International Conference on Computer Vision (ICCV), pp. 2353–2362. IEEE/CVF (2017)
8. Chen, Y., Tian, Y., He, M.: Monocular human pose estimation: a survey of deep learning-based methods. Computer Vision and Image Understanding, pp. 102897 (2020)
9. Kuhn, H.W.: The hungarian method for the assignment problem. Naval Res. Logistics Q. **2**(1–2), 83–97 (1955)
10. He, K., Gkioxari, G., Dollár, P., Girshick, R.: Mask R-CNN. In: 2017 International Conference on Computer Vision (ICCV), pp. 2961–2969. IEEE/CVF (2017)
11. Max Jaderberg, Karen Simonyan, Andrew Zisserman, et al. Spatial transformer networks. In Advances in neural information processing systems, pages 2017–2025, 2015
12. Newell, A., Yang, K., Deng, J.: Stacked Hourglass Networks for Human Pose Estimation. In: Leibe, B., Matas, J., Sebe, N., Welling, M. (eds.) ECCV 2016. LNCS, vol. 9912, pp. 483–499. Springer, Cham (2016). https://doi.org/10.1007/978-3-319-46484-8_29
13. Nguyen, T.V., Song, Z., Yan, S.: Stap: spatial-temporal attention-aware pooling for action recognition. IEEE Trans. Circuits Syst. Video Technol. **25**(1), 77–86 (2014)

14. Yi, Y., Zheng, Z., Lin, M.: Realistic action recognition with salient foreground trajectories. Expert Syst. Appl. **75**, 44–55 (2017)
15. Wang, P., Li, W., Li, C., Hou, Y.: Action recognition based on joint trajectory maps with convolutional neural networks. Knowl.-Based Syst. **158**, 43–53 (2018)
16. Yan, S., Xiong, Y., Lin, D.: Spatial temporal graph convolutional networks for skeleton-based action recognition. In: Thirty-second Conference on Artificial Intelligence (AAAI) (2018)
17. Ma, D., Xie, B., Agam, G.: A machine learning based lecture video segmentation and indexing algorithm. In: Document Recognition and Retrieval XXI, vol. 9021, pp. 90210V. International Society for Optics and Photonics (2014)
18. Davila, K., Zanibbi, R.: Whiteboard video summarization via spatio-temporal conflict minimization. In: 2017 14th IAPR International Conference on Document Analysis and Recognition (ICDAR), vol. 1, pp. 355–362. IEEE (2017)
19. Davila, K., Zanibbi, R.: Visual search engine for handwritten and typeset math in lecture videos and latex notes. In: 2018 16th International Conference on Frontiers in Handwriting Recognition (ICFHR), pp. 50–55. IEEE (2018)
20. Kota, B.U., Davila, K., Stone, A., Setlur, S., Govindaraju, V.: Generalized framework for summarization of fixed-camera lecture videos by detecting and binarizing handwritten content. Int. J. Doc. Anal. Recogn. (IJDAR) **22**(3), 221–233 (2019)
21. Soares, E.R., Barrére, E.: An optimization model for temporal video lecture segmentation using word2vec and acoustic features. In: 25th Brazillian Symposium on Multimedia and the Web, pp. 513–520 (2019)
22. Shah, R.R., Yu, Y., Shaikh, A.D., Zimmermann, R.: Trace: linguistic-based approach for automatic lecture video segmentation leveraging wikipedia texts. In: 2015 IEEE International Symposium on Multimedia (ISM), pp. 217–220. IEEE (2015)
23. Yang, H., Meinel, C.: Content based lecture video retrieval using speech and video text information. IEEE Trans. Learn. Technol. **7**(2), 142–154 (2014)
24. Radha, N.: Video retrieval using speech and text in video. In: 2016 International Conference on Inventive Computation Technologies (ICICT), vol. 2, pp. 1–6. IEEE (2016)
25. Lin, T.-Y., et al.: Microsoft COCO: common objects in context. In: Fleet, D., Pajdla, T., Schiele, B., Tuytelaars, T. (eds.) ECCV 2014. LNCS, vol. 8693, pp. 740–755. Springer, Cham (2014). https://doi.org/10.1007/978-3-319-10602-1_48