Skill Determination from Egocentric Video

Hazel Doughty Dima Damen Walterio Mayol-Cuevas
University of Bristol, Bristol, UK
<Firstname>.<Surname>@bristol.ac.uk

Abstract

In this extended abstract we describe our work on determining relative skill from video [3]. We formulate the problem as pairwise and overall ranking of video collections, and propose a supervised deep ranking model to learn discriminative features between pairs of videos exhibiting different amounts of skill. We test our method on both stationary and egocentric recordings, but note that the egocentric allows for better performance, due to the camera position's closeness to the action as well as information in the head motion. We evaluate on a variety of tasks ranging from drawing to kitchen activities that result in increased ranking accuracy from 73.1% to 78.7% for egocentric viewpoints.

1. Introduction

How-to videos on sites such as Youtube and Vimeo, have enabled millions to learn new skills by observing those more skilled perform the task. From drawing to cooking and repairing household items, learning from videos is nowadays a commonplace activity. With wearable cameras becoming more readily available many of these instructional (*how-to*) videos are recorded from an egocentric perspective. Such a viewpoint offers a more focused view of the task and allows the viewer to gain an immediate understanding by putting themselves in the place of the recorder. However, collections of *how-to* videos tend to be loosely organised without guarantee of the level of skill of the contributors. Therefore, the person wishing to learn has to decide who demonstrates the skill best and who to learn from. Automatic skill determination is crucial for this problem. By ranking videos based on the skill displayed we will not only be able to assess skill in specific training tasks, as has been done in surgery [10], but also assess relative skill in daily living tasks, aiding automated, objective organisation of *how-to* videos.

In our recent work [3] we used stationary footage to determine skill. In this abstract we demonstrate our method is also capable of using egocentric video for skill determination and that egocentric video provides a unique viewpoint with less occlusions of manipulated objects and extra information, allowing us to achieve better results.

2. Related Work

Skill determination in video has received relatively little attention particularly in egocentric video. The majority of existing works in skill determination focus on specfic domains, for instance surgical training [5]. Malpani et al. use hand-crafted features from a combination of video and accelerometer data to assess skill. The features used, such as 'area travelled by instrument tip', are specific to surgical training tasks and are therefore not generalisable to other domains. Also specific to surgical tasks, Zia et al. [10] utilise the structured nature of surgical tasks; using entropy-based features to detect irregularity within performances.

Another domain that has been the focus of previous works is sports [1, 6]. Bertasius et al. [1] assess skill in Basketball from egocentric videos, using a convolutional LSTM network to detect basketball events in the videos, followed by Gaussian mixtures to evaluate skill from these events. As egocentric video is used, the footage directly demonstrates what is happening to the player himself, eliminating the need for tracking used in third person views and reflecting the decisions being made by players.

Similar to our own work, Kim et al. [4] aim to evaluate skill in a more general sense for a variety of tasks. They first obtain action units for each participants performance of a task and evaluate the semantic similarity between the action units of an expert performance and a test video to determine whether the same activity is performed. Our work goes beyond this in two ways. Firstly, we don't rely on having 'expert' performances available during training. Secondly, instead of determining whether the same activity is being performed in relation to the expert, we assume that we have multiple videos of the same activity and aim to rank these videos based on their performance of that same activity.

3. Method

Our method, described in detail in [3], utilises two-stream Temporal Segment Networks (TSN) [9] to model

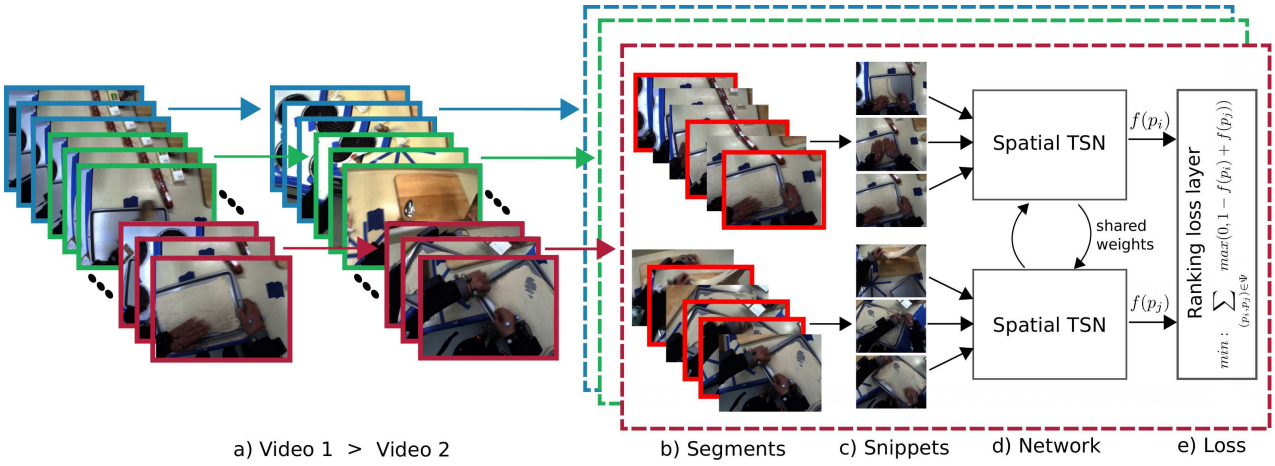

Figure 1: Training for skill determination for the spatial stream. a) Each video in the pair are divided these into N splits for data augmentation. b) Paired splits are then divided up into 3 equally sized paired segments as in [8]. c) The TSN selects a snippet randomly from each segment. d) Each of the snippets are fed into a Siamese architecture with shared weights. e) The loss function computes the margin ranking loss of the pair. The temporal stream works similarly.

the long range temporal structure of the videos. To determine the relative skill levels displayed in different videos a Siamese version of TSN is used, meaning the network is trained on pairs of videos. Each pair of input videos $(p_i, p_j) \in P$ is associated with a label indicating whether p_i displays more skill than p_j or vice-versa. Each input video in the pair is then split into K equally sized segments as in [9], with the TSN randomly selecting a snippet from each segment. We use $K = 3$ in line with [9]. In the spatial network this snippet is a single RGB frame, while the temporal network snippets consist of a stack of 5 optical flow frames in both the horizontal and vertical directions. The six snippets are fed into the Siamese CNN with weights shared between the six identical networks. These networks output a score per snippet, from which consensus scores $f(p_i)$ and $f(p_j)$ are formed for the input videos in the pair (p_i, p_j). To learn a model to determine relative skill between the input videos a margin loss layer is used:

$$\min : \sum_{(p_i,p_j)\in P} max(0, m - f(p_i) + f(p_j)) \qquad (1)$$

where p_i is ranked higher than p_j. During training this loss aims to separate the input videos into the correct ranking by the margin $m = 1$, and back propagates according to the violation of this condition.

In order to make use of the longer nature of videos in skill determination and increase the size of our training data we augment the training data by splitting the videos into N uniform splits before input to the network (Fig 1a). We assume progress through subcomponents of the task is comparable between videos, even if the time taken to complete the task is different. We also assume that an annotation $p_i \succ p_j$ holds for the pairs on subsections of the video pair (p_i, p_j) Thus, for a video pair (p_i, p_j) we compare the pairs $({p_i}^k, {p_j}^k) \quad \forall k = 1...N$.

When testing a video we uniformly sample σ snippets from each video p_i and form ten inputs from crops and flips of the four corners and centre of the images, as in [7]. For each snippet p_{ij} we gain an output score $f(p_{ij})$ for both the spatial network $f_s(p_{ij})$ and temporal network $f_t(p_{ij})$. We then combine these two predictions with a weighted average and combine the predictions of all the snippets to create an overall video prediction using the following equation:

$$f(p_i) = \text{SegmentConsensus}(\alpha f_s(p_{ij}) + (1 - \alpha) f_t(p_{ij})) \qquad (2)$$

In this paper we use mean as the segment consensus function. For an investigation of different functions see [3].

4. Datasets

In our evaluation we compare the results of the skill determination of both stationary and egocentric recordings of three datasets. These datasets consist of the following tasks: rolling out pizza dough, drawing and using chopsticks. Each of the datasets are partitioned in four folds for training and testing and are detailed below.

Drawing This dataset consists of 8 participants drawing copies of two different reference images: a photograph of a hand and a picture of the Sonic the Hedgehog cartoon. Each participant draws the reference images 5 times, resulting in 40 videos. We use the stationary recordings from [3] as well as egocentric recordings from a head-mounted GoPro not previously tested. Both GoPros recorded the drawing tasks at a resolution of 1920x1080 and a frame rate of

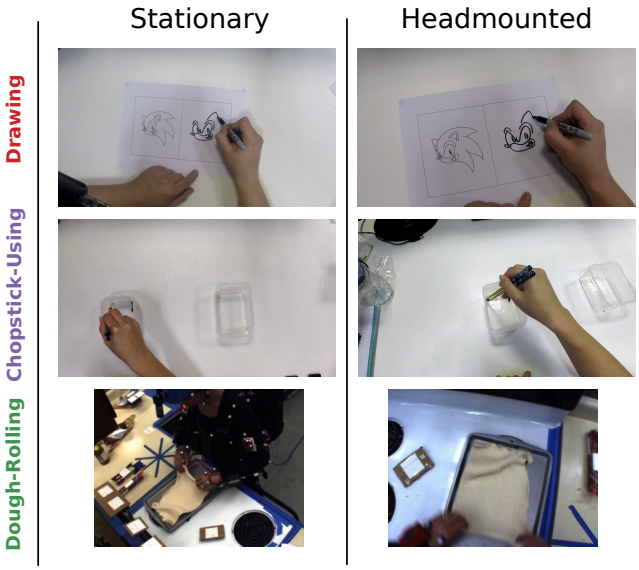

Figure 2: A comparison of the stationary and egocentric viewpoints for each of the three datasets.

60fps.

Chopstick-Using Similarly to the Drawing dataset we compare the stationary recordings tested in [3] as well as egocentric recordings. Again both cameras recorded at a resolution of 1920x1080 and 60fps. In the Chopstick-Using dataset 8 participants each attempt to move four beans between two identical plastic tubs 5 times, totalling 40 videos.

Dough-Rolling This consists of a dough-rolling task from the pizza making activity in the CMU-MMAC dataset [2]. In total there are 33 videos in this dataset from 33 distinct participants. The videos in this dataset consist of a participant opening the dough container and rolling out the dough into a rectangle. The egocentric recording of this data was tested in [3], here we also test a stationary recording of the same activity and same time frame for each participant. Frames from both view points are shown in Figure 2. CMU-MMAC contains five different non-egocentric viewpoints, the one shown in Figure 2 was chosen as it is the least occluded stationary camera during the dough-rolling task.

4.1. Annotations

We assume that provided a human has a good enough view of the task in both videos in a pair they can successfully determine which video they believe displays a higher level of skill and that if multiple people are in unanimous agreement one participant displays more skill than another in a video then that the ranking will hold in another viewpoint. Therefore, we use are annotations from [3] for both egocentric and stationary data. These annotations consisted of Mechanical Turk workers selecting which video in a pair of videos displays more skill. We assume that if all four workers tested for each pair were able to make a unanimous decision then the strict skill pairings also hold for the egocentric versions of the recordings. Thus, the Dough-Rolling task consists of 181 consistent pairs, the Drawing 247 consistent pairs (118 for the Sonic-Drawing task and 129 for the Hand-Drawing task) and the Chopstick-Using 536.

4.2. Different Viewpoints

From Figure 2 we can see that the footage from the stationary and head-mounted cameras are similar for the Drawing and Chopsticks-Using tasks. Both show a clear, un-occluded view of the task in the centre of the image and show only the hands, rather than being third person. Therefore, the main difference between the data in the stationary and egocentric footage for these datasets is that head-mounted data contains motion information for the person's head, although resulting in motion blur, and gives us more information about where the participants attention is focussed. On the other hand, the stationary footage for the Dough-Rolling task is closer to a third person view. Although in the example shown (and many of the other videos) the action takes place near the centre of the image there is more variability in the position the task can take place in and it is not guaranteed to be in the centre of the image or necessarily always in view. Therefore, the differences between the different types of footage for the Drawing and Chopstick-Using datasets still hold here, but with the addition that the egocentric footage guarantees the task is taking place in the centre of the image.

5. Experimental Results

5.1. Evaluation Metric

We use pairwise precision output rankings of each testing fold. This is defined as:

$$\frac{\#\text{Correctly ordered pairs}}{\text{Total \#pairs}} \tag{3}$$

A pair is defined as correctly ordered if for a pair (p_i, p_j) where the annotations contain the preference $p_i \succ p_j$ the method outputs $f(p_i) > f(p_j)$.

For implementation details see [3].

5.2. Results

The results of four-fold cross validation on each of the three datasets for both types of footage. We use $\sigma = 25$ for all results and we report the results for the best α (Eq. 2) for each result. We first note that the egocentric footage provides a much higher accuracy in two out of the three tasks. We also note that this is the same for the individual spatial and temporal results as well as two-stream, the highest result.

Task	Random	Stationary			Egocentric		
		Spatial	Temporal	Two-stream	Spatial	Temporal	Two-stream
Drawing	50	76.7	79.0	**82.1**	72.9	72.3	73.1
Chopstick-Using	50	66.8	69.8	70.0	78.4	73.5	**78.7**
Dough-Rolling	50	52.3	56.0	56.0	77.0	76.1	**78.2**

Table 1: Results of four-fold cross validation on both the egocentric and stationary footage for each of the three datasets. For two out of three datasets the egocentric result outperforms the stationary one.

The largest difference in performance between stationary and egocentric is in Dough-Rolling. Here the egocentric data outperforms the stationary by 22.2%. We believe this is mainly due to the clearer view present in the egocentric data for the Dough-Rolling task. The fact that we train on different image crops and test on the crops of the corners and centre should mitigate the effect of the dough-rolling not necessarily being in the centre, however the clear view of the task from egocentric camera has a large impact on the results.

The Chopstick-Using dataset also displays a large improvement when using an egocentric view. From inspecting the videos for which our method performs much better on the egocentric recordings of the task, we see that the main advantage of the egocentric viewpoint is that it has a much better viewpoint of what is happening while attempting to pick up on the beans, therefore offering more information in relation to the skill of the participant.

Alternatively, drawing performs worse when using the egocentric data as opposed to the stationary data. One issue with a head-mounted camera for this task is that the camera can become so close that it loses information about that task. For instance, one participant mostly looked at the reference image while drawing, which is reflected in the head-mounted recordings. Although this gives us information about where the participant's attention is focussed we lose information by only seeing a small portion of the drawing at a time and not necessarily the hand.

For a further analysis, including analysis of different parameters and consensus functions see [3].

6. Conclusion

In this paper we have discussed our recent work on skill determination. We have shown our method works for both stationary and egocentric data and demonstrated egocentric data provides an opportunity over the stationary data that is advantageous for automatic determination of skill from video.

References

[1] G. Bertasius, S. X. Yu, H. S. Park, and J. Shi. Am i a baller? basketball skill assessment using first-person cameras. *arXiv preprint arXiv:1611.05365*, 2016. 1

[2] F. De la Torre, J. Hodgins, A. Bargteil, X. Martin, J. Macey, A. Collado, and P. Beltran. Guide to the carnegie mellon university multimodal activity (cmu-mmac) database. *Robotics Institute*, page 135, 2008. 3

[3] H. Doughty, D. Damen, and W. Mayol-Cuevas. Who's better, who's best: Skill determination in video using deep ranking. *arXiv preprint arXiv:1703.09913*, 2017. 1, 2, 3, 4

[4] S. T. Kim and Y. M. Ro. Evaluationnet: Can human skill be evaluated by deep networks? *arXiv preprint arXiv:1705.11077*, 2017. 1

[5] A. Malpani, S. S. Vedula, C. C. G. Chen, and G. D. Hager. Pairwise comparison-based objective score for automated skill assessment of segments in a surgical task. In *International Conference on Information Processing in Computer-Assisted Interventions*, pages 138–147. Springer, 2014. 1

[6] P. Parmar and B. T. Morris. Learning to score olympic events. *arXiv preprint arXiv:1611.05125*, 2016. 1

[7] K. Simonyan and A. Zisserman. Two-stream convolutional networks for action recognition in videos. In *Advances in neural information processing systems*, pages 568–576, 2014. 2

[8] J. Wang, Y. Song, T. Leung, C. Rosenberg, J. Wang, J. Philbin, B. Chen, and Y. Wu. Learning fine-grained image similarity with deep ranking. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 1386–1393, 2014. 2

[9] L. Wang, Y. Xiong, Z. Wang, Y. Qiao, D. Lin, X. Tang, and L. Van Gool. Temporal segment networks: towards good practices for deep action recognition. In *European Conference on Computer Vision*, pages 20–36. Springer, 2016. 1, 2

[10] A. Zia, Y. Sharma, V. Bettadapura, E. L. Sarin, and I. Essa. Video and Accelerometer-Based Motion Analysis for Automated Surgical Skills Assessment. *ArXiv e-prints*, Feb. 2017. 1