Who's Better, Who's Best: Skill Determination in Video using Deep Ranking

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

Abstract

This paper presents a method for assessing skill of performance from video, for a variety of tasks, ranging from drawing to surgery and rolling dough. 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 utilise a two-stream Temporal Segment Network to capture both the type and quality of motions and the evolving task state. Results demonstrate our method is applicable to a variety of tasks, with the percentage of correctly ordered pairs of videos ranging from 70% to 82% for four datasets. We demonstrate the robustness of our approach via sensitivity analysis of its parameters.

We see this work as effort toward the automated and objective organisation of how-to videos and overall, generic skill determination in video.

1. Introduction

How-to videos on sites such as Youtube and Vimeo, have enabled millions to learn new skills by observing others more skilled at the task. From drawing to cooking and repairing household items, learning from videos is nowadays a commonplace activity. However, these loosely organised collections normally contain a mixture of contributors with different levels of expertise. The querying person, who has the least expertise, needs to decide who is better and who to learn from. While popularity scores can sometimes help, these are prone to subjective ratings and worse, cheating. Furthermore, the number of *how-to* videos is only likely to increase, fuelled by more cameras recording our daily lives. An intelligent agent that is able to assess the skill of the subject, or rank the videos based on the skill displayed, would enable us to delve into the wealth of this online resource.

In this work, we attempt to determine skill for a variety of tasks from their video recordings. We base this work on two assumptions, first - for tasks where novice human observers *consistently* label one video as displaying *more skill* than another, there is enough information in the visual signal to automate that decision; and second - the same framework for determining skill can be used for a variety of tasks ranging from surgery to drawing and rolling pizza dough.

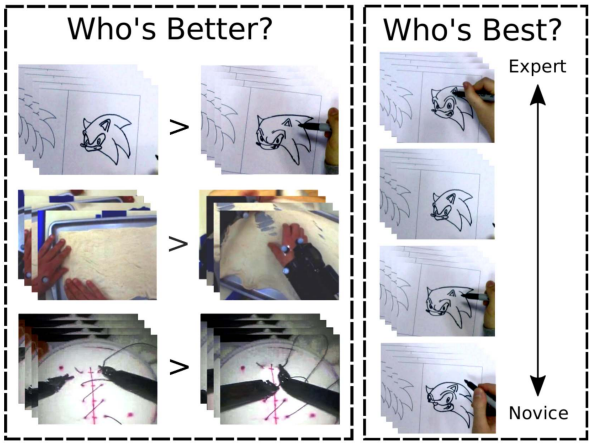

Figure 1: Determining skill in video. **Who's Better?** (Left): pairwise decisions of videos containing the same task, performed with varying levels of skill. **Who's Best?** (Right): ranking formed from pairwise decisions.

We propose to determine skill using a pairwise deep ranking model, which characterises the relative differences in performance between a pair of videos. Each pair consists of two videos: one ranked higher than the other *by human annotators* in terms of the skill displayed (Fig 1). We use a Siamese architecture where *each stream* is made up of a two-stream convolutional neural network (2S-CNN) composed of a spatial and temporal stream. The rationale being that the spatial stream is expected to capture the method and outcomes, while the temporal stream is expected to capture the technique or motion quality. By assigning videos a relative score of skill for the given task, we can create a *skill ranking* for a set of videos.

We evaluate our approach on four datasets (two public); one surgical - in line with previous methods, another on rolling pizza dough, as well as two newly introduced datasets for the tasks of drawing and using chopsticks. Our proposed method outperforms the baseline on each dataset, achieving a correctly ordered video pair accuracy of 75.3% in the surgical tasks, 78.2% in the dough rolling task, 82.1%

for the drawing tasks and 70.0% in the chopstick-using task.

Our main contributions are as follows: i) We present the first general method to determine skill in videos which works both on surgical tasks and a variety of daily tasks. ii) Using this method, we rank the skill displayed in the videos for the task. iii) We present pairwise skill annotations for three datasets, two of which are newly recorded.

2. Related Work

Skill Determination. There have been few prior works on automatically determining skill from videos. The majority of these works are focused on surgical tasks [12, 17, 18, 26, 27, 28, 29, 30], due to the intensive training needs in this area. For instance, Zhang et al. [17] use motion textures to predict the OSATS criteria: a measure of skill specific to the surgical domain. In [28], Zia et al. rely on the repetitive nature of surgical tasks, using the entropy of these repeated motions to identify different skill levels. Malpani et al. [12] use a combination of video and kinematic data to rank performance in two surgical tasks. However, they decompose each task into a sequence of actions, and design specific features for performance evaluation of surgical manoeuvres, which makes this less applicable to non-surgical tasks. Generally, the high speciality of the tasks and methods involved in surgery make these approaches difficult to generalise outside the surgical domain.

Many of these methods also lack a fine-grained approach to identifying skill, instead splitting participants into novices, intermediates and experts [29]. Often this is done by the participant's previous experience, instead of their performance in individual videos. We aim however, to rank the performances in all videos. Thereby, identifying a relative score for each performance, instead of classifying a video, or all of a participant's videos, as expert or novice.

Zhang et al. [27] use relative Hidden Markov Models to evaluate human motion skill by obtaining a ranking between input pairs. However, the main focus in this paper is again motion skill in surgical training tasks. This work is somewhat limited by the ground truth data: the assumption is that a video recorded at a later date will contain better performance than a participant's earlier recording. Thus, skill is only compared within a participant's performances.

There is also some skill assessment work in the domain of sport [1, 2, 7, 9, 14, 15, 16]. However, many of these works are not generalisable to domains outside sports as they either craft features specific to a sport, such as basketball [1, 9], or focus on quality of motion [2, 7, 14, 15]. The most relevant of these works is from Pirsiavash et al. [16], who present a general method for assessing the quality of actions. This is done by estimating human body pose with a skeleton model in order to predict the score of actions, again in sports videos. However, quality of motion on its own is not an essential condition to determine skill. For example, moving a brush in an artistic manner is not a sufficient measure for painting skills.

Video Representation. Recent approaches utilising deep learning to extract features, particularly CNNs, have shown great success in tasks, including image classification [10], object detection [6] and action recognition [19].

A common consideration for these works is how to represent the video as an input to the network. Many mainstream CNNs [19, 21] focus on appearances and short term motions, ignoring any long range temporal structure.

In this paper, we utilise the recent Temporal Segment Networks (TSN) architecture [23], to model long range temporal structure. This architecture achieves state of the art performances for action recognition on UCF101 [20] and HMDB51 [11]. TSN decomposes the input video into uniformly sized segments, sampling a snippet from each segments as input to a CNN. The long range temporal structure is modelled by forming a consensus from these snippets, before the result is fed into the loss layer during training. This method has the advantage of enabling end-to-end learning for long video sequences with a relatively low cost in terms of time and computing resources.

Deep Ranking. Deep learning models have also been used in learning to rank, where the popular pairwise learning to rank model [3] is common. The method aims to minimise the average number of incorrectly ordered pairs of elements in a ranking, by training a binary classifier which can decide which element in a pair should be ranked higher.

For example, Wang et al. [22] use a triplet network for learning fine grained image similarity. Given a query image and a pair of images, one more similar to the query than the other, the training aims to learn a mapping such that more similar images have smaller distances between their features.

Yao et al. [24] use a pairwise deep ranking model to perform highlight detection in egocentric videos by using pairs of highlight and non-highlight segments.

However, to our knowledge, no previous work has used deep learning to determine skill or indeed rank videos in terms of skill. We next present our method for determining skill using deep ranking.

3. Learning to Determine Skill

In this section we first give an overview of the skill determination problem. We then present the two-stream CNN we use to determine skill, followed by the pairwise deep ranking model used to train both streams.

3.1. Definition

Our goal is to learn models for ranking skill. Given a task, we have a set of K videos $P = \{p_k, 1 \leq k \leq K\}$, from multiple people, each performing the task one or more

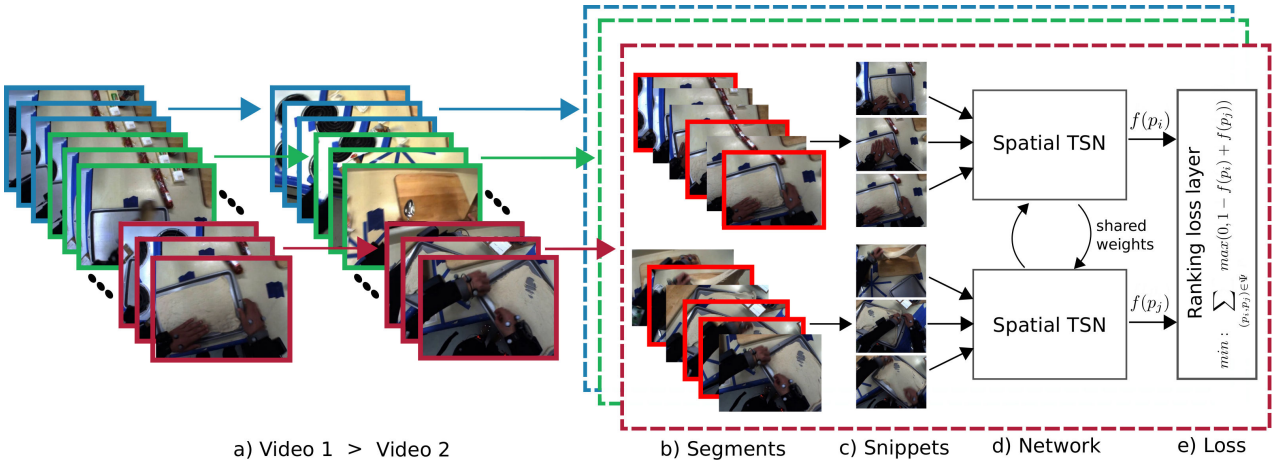

Figure 2: Training for skill determination. a) We consider a pair of videos, with one showing a higher level of skill $E(\cdot) > 1$, and divide these into N splits for data augmentation. b) Paired splits are then divided up into 3 equally sized paired segments as in [22]. c) The TSN selects a snippet randomly from each segment. For the spatial network this is a single frame, for the temporal network this is a stack of 5 dense horizontal and vertical flow frames. d) Each of the snippets are fed into a Siamese architecture of shared weights, for both spatial and temporal streams, of which only the spatial is shown here. e) The loss function computes the margin ranking loss of the pair.

times. We consider each video independently, even if performed by the same person. We assume that people differ in the skill they display in each video, even across multiple runs, and we are thus interested in ranking relative skill per video instead of accumulating a score per person.

For each pair of videos, we use novice human annotations to determine skill (ref Sec. 4.1). Assume $E(p_i, p_j)$ is the decision by human annotators where

$$E(p_i, p_j) = \begin{cases} 1 & p_i \text{ shows higher skill than } p_j \\ -1 & p_j \text{ shows higher skill than } p_i \\ 0 & \text{no skill preference} \end{cases} \quad (1)$$

Note that according to Eq. 1, $E(p_i, p_j) = -E(p_j, p_i)$, we thus need to only obtain one annotation for each pair.

We check all annotations for triangular consistency. Assume $E(p_i, p_j) > 0$ and $E(p_j, p_k) > 0$, we check for $E(p_i, p_k) < 0$, which would show inconsistency in annotations. We do this by creating a directed graph with P nodes and edges $(p_i \rightarrow p_j) \, \forall \, 1 \leq i, j \leq K$ where $E(p_i, p_j) > 0$. Cycles in the graph would indicate a triangular inconsistency, which we manually resolve.

3.2. Two-Stream CNN for Skill Determination

Tasks differ in how skill can be demonstrated. In this respect we identify two main sources of relevant information. The first is the quality and type of motions used. The second is the effect on the environment captured through the appearance of the task. We thus utilise two stream convolutional neural networks (2S-CNN) for skill determination. Specifically, we base our method on Temporal Segment Networks (TSN) [23], which utilise the BN-Inception network architecture [8]. We select TSN due to their ability to model long range temporal structure. This makes the approach suitable for determining skill for tasks regardless of the length of the videos used.

In training TSN, as in [23], we uniformly divide each input video sequence into three segments, then randomly sample a single short snippet from each of these segments (Fig 2b,c). For each iteration in training, our 2S-CNN outputs a preliminary prediction of skill for each snippet. This decision is then pooled across the three snippets, creating a score per input video. The output to the loss function (Fig. 2e), in both the spatial and temporal streams, is then the consensus between selected snippets.

3.3. Pairwise Deep Ranking

As we want to determine relative skill of users in different videos, we use the pairwise approach for learning to rank. To do this we build a Siamese version of the two-stream TSN described in Section 3.2, with the weights shared across both sides of the Siamese network (Fig. 2d). Given a pair of videos, where the first video is ranked higher than the second in terms of skill, we want the Siamese network to output a higher score for the first. Formally, we have a set of pairs $\Psi = \{(p_i, p_j); E(p_i, p_j) = 1\}$ (ref Eq. 1). These two videos are fed into the separate, but identical, TSNs which form the Siamese network (Fig 2a). Assuming the TSN outputs $f(\cdot)$, our goal is to learn the function f such that we determine skill, where

$$f(p_i) > f(p_j) \quad \forall (p_i, p_j) \in \Psi \quad (2)$$

To gain an overall rank for all videos, we use a margin loss layer to evaluate the loss for each pair. The loss function we use is an approximation to 0-1 ranking error loss that has been used successfully for other applications [24, 22];

$$min: \sum_{(p_i,p_j)\in\Psi} max(0, 1 - f(p_i) + f(p_j)) \quad (3)$$

During training, this loss function evaluates the violation of the ranking of each pair and back-propagates the gradient through the network, in order to learn common features which indicate subjects in the videos display more skill.

3.4. Data Augmentation

Traditionally, 2S-CNN are used for action recognition [19], thus the whole length of the video needs to be considered once to recognise the undertaken action. In this work, we are examining skill, which could be understood from all (or any) parts of the video sequence. To increase the size of our training data and make the most of the whole extent of the video sequence, we augment the training data by splitting videos into N uniform splits (Fig. 2a). We assume that two videos of the same task have comparable speed, and thus compare the temporal splits across a pair of videos in order. Assume p_i^k is the k^{th} split of video p_i, we extend the skill annotations such that,

$$E(p_i^k, p_j^k) = E(p_i, p_j) \quad \forall k = 1 \cdots N \quad (4)$$

In our experiments $N = 7$ was tested. By pairing corresponding splits, we ensure the two videos are compared in a similar stage of the task performance, and therefore more discriminative features are likely to be learnt.

3.5. Skill for a Test Video

Following training, the learnt 2S-CNN weights are used to evaluate the skill for test videos of the same task. In testing, we uniformly sample σ snippets from each video p_i, again as in [23]. Each snippet p_{ij} $1 \leq j \leq \sigma$ is then fed into the spatial and temporal TSN independently. The output for each snippet is a score $f(p_{ij})$ for both spatial $f_s(p_{ij})$ and temporal $f_t(p_{ij})$ streams. To fuse the spatial and temporal networks for all snippets we take the weighted average of the outputs,

$$f(p_i) = \frac{1}{\sigma} \sum_{j=1}^{\sigma} \alpha f_s(p_{ij}) + (1 - \alpha) f_t(p_{ij}) \quad (5)$$

where α reflects the fusion weighting between spatial and temporal information, and σ is the number of snippets.

An overall ranking for a test set is achieved by ordering all test videos in a descending order based on $f(p_i)$.

4. Tasks and Datasets

For evaluation we conduct experiments on tasks from four datasets - two published and two newly recorded (Fig. 3). The first is a surgical dataset. Three other distinct datasets containing daily living tasks are also used, to demonstrate the generality of the approach. Here we detail the four datasets, followed by the skill annotations for these datasets.

Surgery. We use the published JHU-ISI Gesture and Skill Assessment Working Set (JIGSAWS) dataset [5]. In this dataset, three surgical procedures are performed using the da Vinci surgical system by 8 surgeons with varying levels of experience. In total, JIGSAWS consists of 36 trials of Knot Tying, 28 trials of Needle Passing and 39 trials of Suturing. This dataset contains stereo recordings, from which we use only the right-hand video of each sequence.

Dough-Rolling. We use the kitchen-based CMU-MMAC dataset [4], and select the dough rolling task from the pizza making activity, as this exhibits varying levels of performance across participants. In total, 33 Dough-Rolling videos from 33 distinct participants are manually extracted.

Drawing. We introduce a new dataset for drawing, captured using a stationary camera at a resolution of 1920x1080 and a frame rate of 60 fps. Participants were given a reference image to copy. Two reference images were used; a cartoon of Sonic the Hedgehog and a gray-scale photograph of a hand. Similarly to the **Surgery** tasks, both tasks were performed five times each, by four participants.

Chopstick-Using. We also introduce a new dataset for using chopsticks, captured using the same setup as the **Drawing** dataset. Participants were presented with two identical tubs, one containing 4 coffee beans. Each participant was tasked with moving as many of the beans as possible from one tub to the other using chopsticks. Eight participants were recruited, each repeated the task 5 times and were limited to one minute per trial.

4.1. Skill Annotation

Only the JIGSAWS dataset came with skill scores. This was annotated by a surgery expert, out of a maximum score of 30. In this section, we explain how we obtained skill ranking for the remaining three datasets using *Amazon Mechanical Turk (AMT)*.

We determine the ground truth relative ranking of video pairs using a similar method to [13], where the authors demonstrate crowdsourcing yields reliable pairwise comparison for skill in surgical tasks. We asked AMT workers to watch pairs of videos simultaneously and select the video displaying the higher level of skill for the given task. Each worker was presented with 5 pairs of videos per HIT from the same task. One of these pairs was a quality control pair -

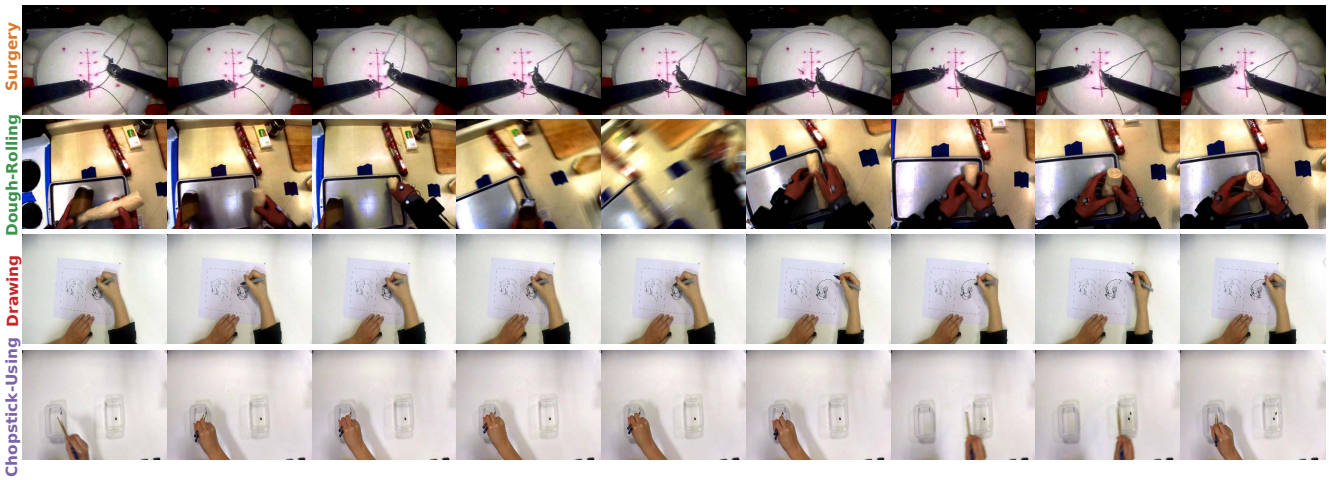

Figure 3: Sample sequences from the four tasks.

Task	#Videos	#Max Pairs	#Cons. Pairs	%Cons. Pairs
Surgery (KT)	36	630	596	95%
Surgery (NP)	28	378	362	96%
Surgery (Suturing)	39	741	701	95%
Dough-Rolling	33	528	181	34%
Drawing (Sonic)	20	190	118	62%
Drawing (Hand)	20	190	129	68%
Chopstick-Using	40	780	536	69%

Table 1: For the four datasets: #videos, #of pairs $(n)(n-1)/2$ with number of strict consistent pairs in annotations and their percentages. KT=Knot Tying, NP=Needle Passing

a pair with an obvious difference in the skill of the two subjects. Annotators were asked for strict preferences per pair. We then check for consensus between different annotators for skill annotation. Each video pair was annotated by four different workers. Only pairs of videos for which *all* annotators agreed on their skill order are considered. These were used for both training and testing subsequently.

We further check for consistent strict pairings. An inconsistent set of strict pairings is one that contains a triangular inconsistency as explained in Section 3.1. For instance the ranking of videos p_i, p_j and p_k would be inconsistent if the set of ordered pairs contained the pairs $p_i > p_j$, $p_j > p_k$ and $p_k > p_i$. Only a single triangular inconsistency was found in all AMT annotations for the three tasks. This was in the **Dough-Rolling** task and was excluded from training and testing. Similarly, we take the skill scores from the **Surgery** dataset and compute all strict consistent pairs. Table 1 presents results on these consistent pairings.

From Table 1, the **Surgery** dataset has a high number of consistent pairs ($> 95\%$). The pairs in this dataset come from the scores of a single expert, available with the JIGSAWS dataset, therefore inconsistent pairs only arise when the scores for two videos are identical. For the other tasks, we use the judgements from multiple AMT workers. Dough-Rolling has the lowest percentage of consistent strict pairs, as many pairs were considered comparable in skill by human annotators. This is likely due to the nature of the task. Many subjects do manage a similar level of performance, resulting in fewer strict orderings. For the newly-recorded datasets of Drawing and Chopstick-Using, the number of strict consistent pairs is $60-70\%$.

We will publish the annotations and consistent pairings for all datasets towards a combined dataset for skill determination in video. *Link available with publication*

4.2. Alternative Measures of Skill

One naive way to approach measuring skill is to use time of completion, as finishing a task faster (or slower) could imply a higher level of skill. However, by examining the JIGSAWS dataset we can see that time is not sufficient. When comparing the ranking given by the time of completion against the ranking for the score achieved in each video we can see that, although there is some correlation between score and time in the Knot Tying task ($\rho = 0.72$), there is little correlation in the Needle Passing task ($\rho = 0.23$) and the Suturing task ($\rho = 0.34$). Therefore, although time could be useful in some cases, we conclude it is not a general method for skill determination.

The JIGSAWS dataset also highlights the problem with classifying users into the categories of novice, intermediate and expert as a ground-truth of skill. In the Knot Tying task, 3 participants achieve the highest score of 22 in one trial. Each of these three participants falls into a different skill category.

We thus do not further consider time of completion or participant skill categories, and present results for ranking collections of videos using the supervised pairings from Section 4.1.

Task	Baseline	Siamese TSN			Siamese TSN with additional pairs			α
		Spatial	Temporal	Two-stream	Spatial	Temporal	Two-stream	
Surgery	50%	66.5%	**74.4%**	**74.4%**	66.5%	**75.3%**	**75.3%**	0.0
Dough-Rolling	50%	73.9%	**76.7%**	75.4%	77.0%	76.1%	**78.2%**	0.6
Drawing	50%	75.6%	76.5%	**77.4%**	76.7%	79.0%	**82.1%**	0.3
Chopstick-Using	50%	67.7%	67.4%	**68.1%**	66.8%	69.8%	**70.0%**	0.1

Table 2: Results of 4-fold cross validation on all datasets, for baseline and the proposed method with and without augmentation pairs (best performance in bold). For all datasets, temporal or two-stream outperform spatial and baseline results.

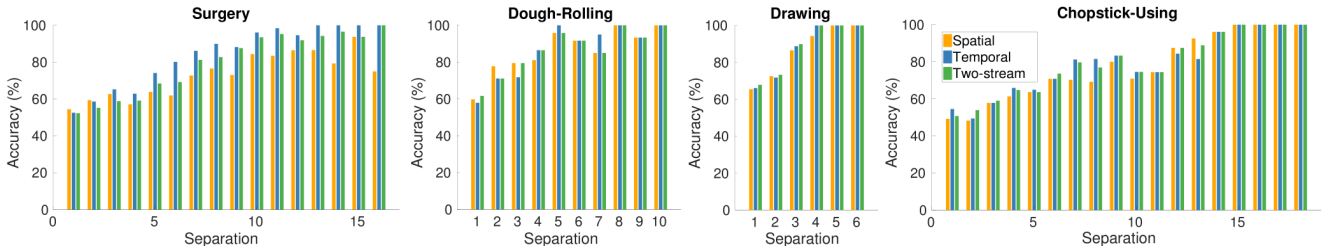

Figure 4: The accuracy of each ordered pair by separation between videos in a pair for each task, using the data augmentation from Section 3.4 with $\alpha = 0.4$, $\sigma = 25$. The accuracy consistently increases as tested pairs are further in the ground-truth ranking for all datasets.

5. Experiments

For all datasets, we use a four-fold cross validation to report results. For each fold, the pairs between three quarters of the videos are used in training, and we then test on all remaining pairs. This includes pairs where neither video has been used in a pair for training as well as pairs where one video has been used in training within a different pairing.

5.1. Implementation Details

To extract the optical flow frames for input to the temporal network we use the $TV - L^1$ algorithm [25], as this works with the modified version of Caffe we use [23]. We use mini-batch stochastic gradient descent with a batch size of 128 and a momentum of 0.9. Both sides of each Siamese network are initialised with network weights from pre-trained ImageNet models [10]. In the spatial network the learning rate begins as 0.001 and decreases by a factor of 10 every 1.5K iterations, with the learning process finishing after 3.5K iterations. The temporal network's learning rate is initialised as 0.005, decreasing by a factor of 10 after 10K iterations and after 16K iterations, with learning ending after 18K iterations. The training time for each fold is approximately 3hrs for the spatial and 18hrs for the temporal stream with a NVIDIA TITANX GPU.

To avoid over-fitting we use the same data augmentation techniques as Wang et al. [23] in addition to the technique described in Section 3.4, namely horizontal flipping, corner cropping and scale jittering on the 340x256 pixels RGB and optical flow images. The cropped regions are 224x224 pixels for network training. The batch-normalisation layers throughout the network also prevent over-fitting, as does the dropout layer before the output and segmental consensus layers. The dropout ratios used are 0.8 and 0.7 for the spatial and temporal networks respectively.

5.2. Evaluation Metric

To evaluate our method, we use pairwise precision on the rankings produced by each testing fold. Pairwise precision is defined as the *percentage of correctly ordered pairs*. We say a pair is correctly ordered if for a pair (p_i, p_j) where $E(p_i, p_j) = 1$ in the ground truth, the method outputs $f(p_i) > f(p_j)$.

As there are no generic existing methods for ranking skill nor performing skill determination for non-surgical tasks, we use the average performance of random rankings of the videos in each task as a baseline, i.e. 50% accuracy[1].

5.3. Results

We first test the results of training our Siamese TSN, without the data augmentation described in Section 3.4. This uses the training method described in Section 3.3, with each pair forming a single input into the TSN. The results of 4-fold cross validation on each of the four datasets are shown in Table 2. We report results for $\sigma = 25$ as in [23], and for the best α per dataset (Eq. 5). Below we test the sensitivity of these results to the values of α and σ.

We first note that the baseline of average performance of random rankings is improved upon for every dataset. We also notice that in general the temporal network performs

[1]Each pair in a random ranking has two possibilities: to be correctly or incorrectly ordered. For each possible ranking with performance x, there exists the reverse ranking with performance $1 - x$. Hence baseline of 50%.

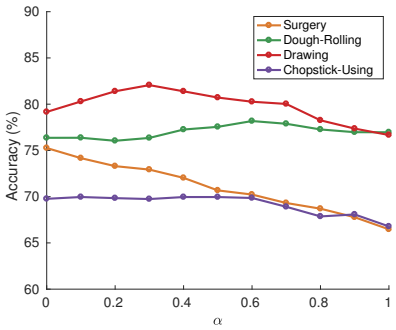

Figure 5: The accuracy for each dataset with different α values. The system is resilient to the parameter value chosen.

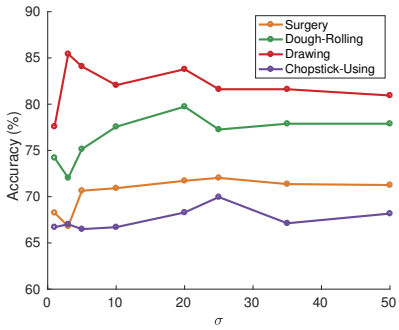

Figure 6: The accuracy achieved when changing the number of snippets used in testing.

better than the spatial network, except for the Chopstick-Using task, where the results are comparable (67.7% vs. 67.4%). This implies the motions performed are more important for determining skill than the current state of the task (captured in the spatial stream). For this reason the temporal network is weighted higher than the spatial network when fusing, for the majority of datasets (best $\alpha < 0.5$). We note that the largest difference between the two streams is in the Surgery tasks. This is because these require quick smooth motions, putting minimal stress on the surrounding areas. Hence, while the end result of each stage is visually similar, the motions affect the scoring significantly.

We also test the data augmentation technique described in 3.4. From Table 2, we see that our data augmentation for skill videos consistently improves upon the basic Siamese TSN. This is particularly true for the Dough-Rolling (75.4% to 78.2%) and Drawing tasks (77.4% to 82.1%), as these datasets are smaller, resulting in fewer training pairs.

While Table 2 reports results for pairwise comparisons, we also wish to assess which pairs are being correctly ordered. Assume we have consistent annotation pairings resulting in the partial ranking $p_i < p_{i+1} < ... < p_{i+n} < p_j$. It is more important that the pair (p_i, p_j) is correctly ordered by our method, as opposed to the pair (p_i, p_{i+1}). We test this by reporting accuracy as separation increases. For the partial ranking above, we define the separation of the pair (p_i, p_{i+1}) to be 1 and of the pair (p_i, p_j) to be $n + 1$.

From Figure 4 we observe, that as the separation between videos increases so does the accuracy of the correctly ordered pairs, reaching 100% for the furthest pairs.

Fusion Parameter. We assess the sensitivity of our results to the late fusion weighting α in Equation 5. We test α values from 0 to 1 at intervals of 0.1 for all datasets, as shown in Figure 5. This confirms that, for the majority of tasks, the combination of temporal and spatial modalities is useful. The only exception to this is the Surgery task which peaks at $\alpha = 0$, where no information from the spatial network is included. All tasks, except Dough-Rolling, peak with $\alpha < 0.5$, demonstrating the importance of the temporal network. Despite these differences, we observe that the method is fairly resilient to the values of α picked, particularly for the Chopstick-Using and Dough-Rolling tasks.

Number of Snippets in Testing. Previous results are reported for $\sigma = 25$ snippets, similar to [23]. We assess the effect of σ on performance. When $\sigma = 1$, we select one snippet at the centre of the video, for $\sigma = 3$, we select a snippet at the start, one at the centre and one at the end. Snippets are uniformly sampled for $\sigma = 5, 10, 20, 25, 35, 50$. Results are shown in Figure 6.

From Figure 6 we see little improvement after $\sigma = 25$ snippets. We also note reasonable accuracy is achieved with only 5 snippets for all datasets. Interestingly, for Drawing, accuracy is the highest for $\sigma = 3$, with the lowest at $\sigma = 1$. This indicates discriminative information is contained in the snippets from the first and last parts of the video, showing the initiation of drawing as well as its completion. This is in contrast with the Surgery tasks, as mistakes will have been corrected by the last snippet meaning little useful information is contained there.

Consensus Function. Previously, in [23] mean was used as the segmental consensus function to fuse the results from the $\sigma = 25$ snippets in testing. We compare the results of four consensus functions for skill determination; min, mean, median and max in Figure 7, with $\alpha = 0.4, \sigma = 25$. From these results we see that mean is the best segmental consensus function overall, performing the best in the Surgery tasks and 2nd in all others. Intuitively, min seems a reasonable function to use for skill, as this would define skill as lack of mistakes. However, we see that it is too prone to outliers, only outperforming all others in Chopstick-Using.

Multi-Task Datasets. In all previous results, we train the model on one dataset and on all tasks within the dataset. For example, for Drawing, training will include pairings from the Sonic and the hand drawings. We evaluated training on only the Hand-Drawing task, and test on both the Hand-Drawing and the Sonic-Drawing separately. For single-

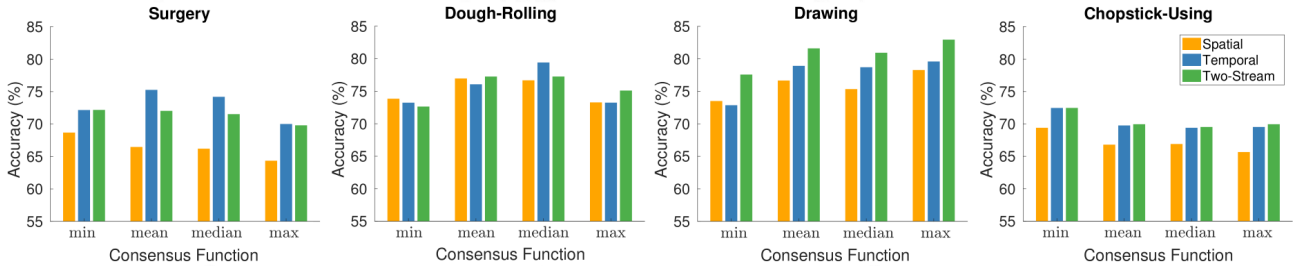

Figure 7: The accuracy of each dataset using four functions for segmental consensus, with $\alpha = 0.4$ and $\sigma = 25$.

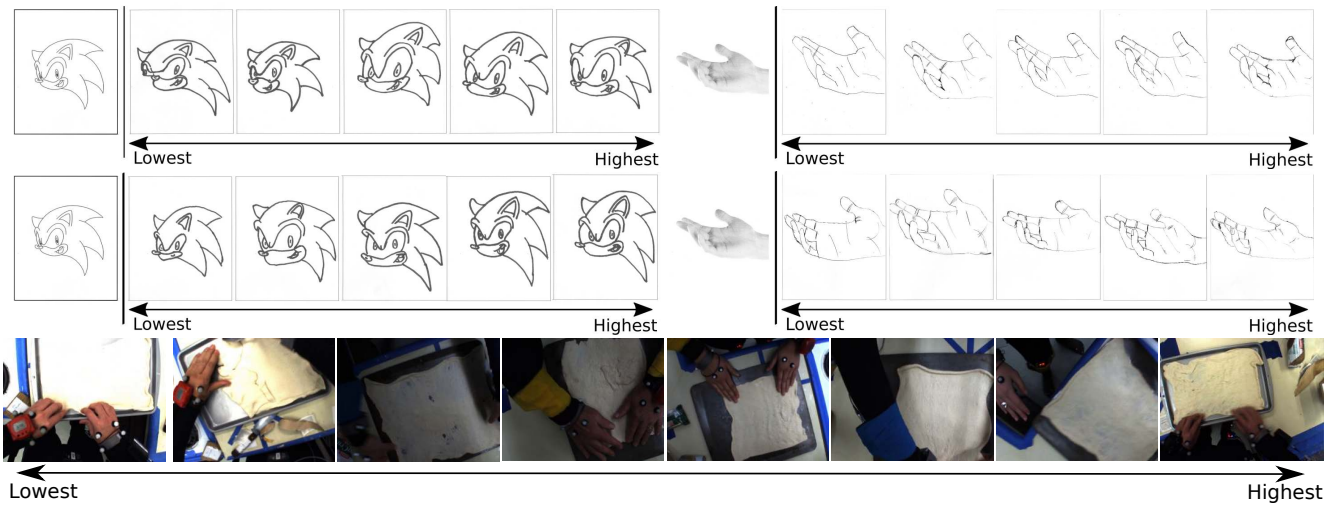

Figure 8: Ranking examples from the Drawing and Dough-Rolling tasks from multiple folds.

task testing results are comparable to multi-task training (77.0%). However, when the task is not included in training, e.g. training on hand and testing on Sonic, we note a significant drop in performance (46.0%). We notice similar drops when training on one dataset and testing on another. The generality of the trained model to new tasks within the same dataset or other datasets is a topic for future research.

5.4. Example Rankings

We demonstrate qualitative results of our method using sample rankings in Fig. 8. For Drawing, generally the results from the higher ranked videos have a better resemblance to the reference image. In the Sonic-Drawing rankings we see visible improvement in the shape and expression of Sonic from left to right. In the Hand-Drawing task the method manages to lowly rank the videos in which the participants draw the perspective incorrectly or add little detail. In Figure 8, we can also demonstrate in increase in skill from lowest to highest ranked videos for the Dough-Rolling task. For example, the second image shows dough which has been badly folded over to patch a large hole. The third image shows an improvement, with the dough containing only several small holes. Again the fourth image demonstrates more skill, as holes have been patched less obviously.

Clearly, a single frame from the video is not sufficient to visualise skill or ranking. Rankings are thus demonstrated by video on the authors' webpage for all tasks.

6. Conclusion

In this paper we have presented a method to rank videos based on the skill subjects demonstrate. Particularly, we have proposed a pairwise deep ranking model which utilises both spatial and temporal streams to determine and rank skill. We have tested this method on four separate datasets, two newly created, and shown that our method out performs the baseline on each dataset, with all tasks achieving over 70% accuracy. Furthermore, we have explored the best way to form a consensus for skill from a video and examined the method's resistance to changes in parameters.

We see our work as an encouraging step toward the automated and objective organisation of *how-to* video collections and as a framework to motivate more work in skill determination from video. Further work involves exploring trained fusion between the two streams of the network, as well as testing on other and across datasets.

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. 2

[2] O. Çeliktutan, C. B. Akgul, C. Wolf, and B. Sankur. Graph-based analysis of physical exercise actions. In *Proceedings of the 1st ACM international workshop on Multimedia indexing and information retrieval for healthcare*, pages 23–32. ACM, 2013. 2

[3] G. Chechik, V. Sharma, U. Shalit, and S. Bengio. Large scale online learning of image similarity through ranking. *Journal of Machine Learning Research*, 11(Mar):1109–1135, 2010. 2

[4] 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. 4

[5] Y. Gao, S. Vedula, C. Reiley, N. Ahmidi, B. Varadarajan, H. Lin, L. Tao, L. Zappella, B. Béjar, D. Yuh, et al. The JHU-ISI gesture and skill assessment dataset (JIGSAWS): A surgical activity working set for human motion modeling. In *Medical Image Computing and Computer-Assisted Intervention (MICCAI)*, 2014. 4

[6] R. Girshick, J. Donahue, T. Darrell, and J. Malik. Rich feature hierarchies for accurate object detection and semantic segmentation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 580–587, 2014. 2

[7] W. Ilg, J. Mezger, and M. Giese. Estimation of skill levels in sports based on hierarchical spatio-temporal correspondences. In *Joint Pattern Recognition Symposium*, pages 523–531. Springer, 2003. 2

[8] S. Ioffe and C. Szegedy. Batch normalization: Accelerating deep network training by reducing internal covariate shift. In *Proceedings of the 32nd International Conference on Machine Learning (ICML-15)*, pages 448–456, 2015. 3

[9] M. Jug, J. Perš, B. Dežman, and S. Kovačič. Trajectory based assessment of coordinated human activity. In *International Conference on Computer Vision Systems*, pages 534–543. Springer, 2003. 2

[10] A. Krizhevsky, I. Sutskever, and G. E. Hinton. Imagenet classification with deep convolutional neural networks. In *Advances in neural information processing systems*, pages 1097–1105, 2012. 2, 6

[11] H. Kuehne, H. Jhuang, E. Garrote, T. Poggio, and T. Serre. HMDB: a large video database for human motion recognition. In *Computer Vision (ICCV), 2011 IEEE International Conference on*, pages 2556–2563. IEEE, 2011. 2

[12] 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. 2

[13] A. Malpani, S. S. Vedula, C. C. G. Chen, and G. D. Hager. A study of crowdsourced segment-level surgical skill assessment using pairwise rankings. *International journal of computer assisted radiology and surgery*, 10(9):1435–1447, 2015. 4

[14] G. I. Parisi, S. Magg, and S. Wermter. Human motion assessment in real time using recurrent self-organization. In *Robot and Human Interactive Communication (RO-MAN), 2016 25th IEEE International Symposium on*, pages 71–76. IEEE, 2016. 2

[15] P. Parmar and B. T. Morris. Learning to score olympic events. *arXiv preprint arXiv:1611.05125*, 2016. 2

[16] H. Pirsiavash, C. Vondrick, and A. Torralba. Assessing the quality of actions. In *European Conference on Computer Vision*, pages 556–571. Springer, 2014. 2

[17] Y. Sharma, V. Bettadapura, T. Plötz, N. Hammerla, S. Mellor, R. McNaney, P. Olivier, S. Deshmukh, A. McCaskie, and I. Essa. Video based assessment of osats using sequential motion textures. In *International workshop on modeling and monitoring of computer assisted interventions (M2CAI)-workshop*, 2014. 2

[18] Y. Sharma, T. Plötz, N. Hammerld, S. Mellor, R. McNaney, P. Olivier, S. Deshmukh, A. McCaskie, and I. Essa. Automated surgical osats prediction from videos. In *Biomedical Imaging (ISBI), 2014 IEEE 11th International Symposium on*, pages 461–464. IEEE, 2014. 2

[19] 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, 4

[20] K. Soomro, A. Roshan Zamir, and M. Shah. UCF101: A dataset of 101 human actions classes from videos in the wild. In *CRCV-TR-12-01*, 2012. 2

[21] D. Tran, L. Bourdev, R. Fergus, L. Torresani, and M. Paluri. Learning spatiotemporal features with 3d convolutional networks. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 4489–4497, 2015. 2

[22] 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, 3, 4

[23] 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. 2, 3, 4, 6, 7

[24] T. Yao, T. Mei, and Y. Rui. Highlight detection with pairwise deep ranking for first-person video summarization. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 982–990, 2016. 2, 4

[25] C. Zach, T. Pock, and H. Bischof. A duality based approach for realtime tv-l 1 optical flow. In *Joint Pattern Recognition Symposium*, pages 214–223. Springer, 2007. 6

[26] Q. Zhang and B. Li. Video-based motion expertise analysis in simulation-based surgical training using hierarchical dirichlet process hidden markov model. In *Proceedings of the 2011 international ACM workshop on Medical multimedia analysis and retrieval*, pages 19–24. ACM, 2011. 2

[27] Q. Zhang and B. Li. Relative hidden markov models for video-based evaluation of motion skills in surgical training.

IEEE transactions on pattern analysis and machine intelligence, 37(6):1206–1218, 2015. 2

[28] A. Zia, Y. Sharma, V. Bettadapura, E. Sarin, T. Ploetz, M. Clements, and I. Essa. Automated video-based assessment of surgical skills for training and evaluation in medical schools. *International journal of computer assisted radiology and surgery*, 11(9):1623, 2016. 2

[29] A. Zia, Y. Sharma, V. Bettadapura, E. L. Sarin, M. A. Clements, and I. Essa. Automated assessment of surgical skills using frequency analysis. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 430–438. Springer, 2015. 2

[30] 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. 2