Towards Skill Determination in 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 recent work on assessing relative skill from video applicable to a variety of tasks [2]. In this work we formulate the problem as pairwise (who's better?) and overall (who's best?) ranking of video collections, using supervised deep ranking. We propose a novel loss function that learns discriminative features when a pair of videos exhibit variance in skill, and learns shared features when a pair of videos exhibit comparable skill levels. Results demonstrate our method is applicable across tasks, with the pairwise precision ranging from 70% to 83% for four datasets. We see this work as effort toward the automated organization of how-to video collections 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 organized collections normally contain a mixture of contributors with different levels of expertise. The querying person needs to decide who is better and who to learn from. Furthermore, the number of *how-to* videos is only likely to increase, fueled 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 on-line resource.

In our recent work [2], we present the first general method to determine skill for a variety of tasks ranging from surgery to drawing and rolling pizza dough from their video recordings, alongside pairwise skill annotations for three datasets, two of which are newly recorded.

2. Tasks and Datasets

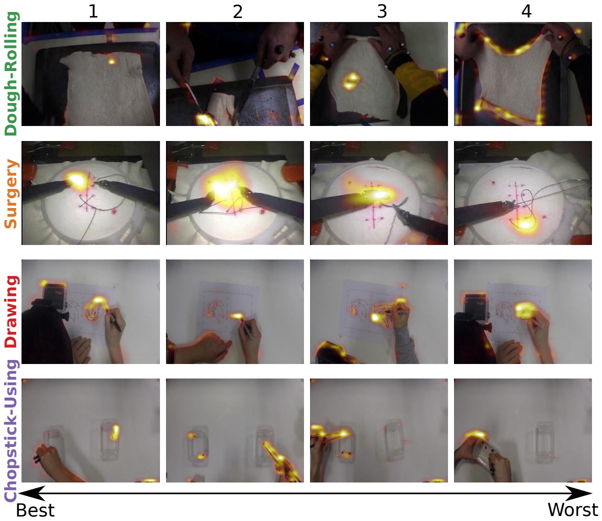

Figure 1. Spatial activations for sample frames at varying ranks.

For evaluation we conduct experiments on tasks from four datasets - two published and two newly recorded (Fig. 1). The first is a surgical dataset from the JIGSAWS dataset [3]. Three other datasets containing daily living tasks are also used, to demonstrate the generality of the approach. The first is from CMU-MMAC [1] and consists of the dough rolling task from the pizza making activity. We then introduce new datasets for the tasks of Drawing and Chopstick-Using which can be seen in Figure 1. These annotated datasets will combined to form the new EPIC-Skills 2018 dataset which can be found on the authors' web-pages.

3. Learning to Determine Skill

Our method, described in detail in [2], aims to rank the relative skill displayed in individual videos. The method can be seen in Figure 2. We consider all pairs of videos, where the first is showing a higher level of skill Ψ, or their skill is comparable Φ, and divide these into N splits to make use of the entire video sequence (Fig. 2a). We use Temporal Segment Networks (TSN) [5] to model the long range temporal structure of the videos, this divides paired splits into 3 equally sized paired segments (Fig. 2b). TSN then 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 (Fig 2c). Each snippet is then fed into a

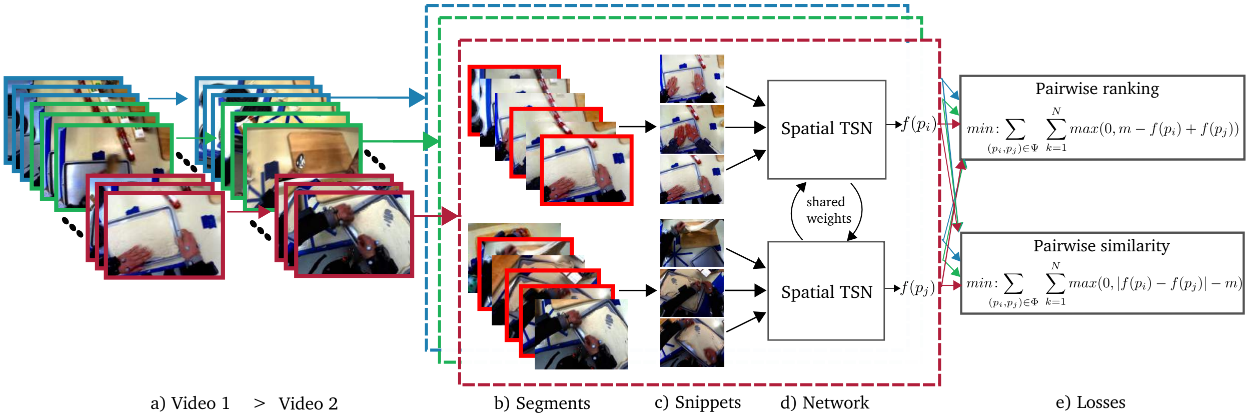
Spatial TSN
Spatial TSN
shared weights
$f(p_i)$
$f(p_j)$
Pairwise ranking
$min\colon \sum_{(p_i,p_j)\in\Psi} \sum_{k=1}^{N} max(0, m - f(p_i) + f(p_j))$
Pairwise similarity
$min\colon \sum_{(p_i,p_j)\in\Phi} \sum_{k=1}^{N} max(0, |f(p_i) - f(p_j)| - m)$
a) Video 1 > Video 2
b) Segments
c) Snippets
d) Network
e) Losses

Figure 2. Training for skill determination

Method	Surgery	Dough-Rolling	Drawing	Chopstick-Using
RankSVM [4]	65.2	72.0	71.5	**76.6**
Yao *et al*. [6]	66.1	78.1	72.0	70.3
Ours	**70.2**	**79.4**	**83.2**	71.5

Table 1. Results of 4-fold cross validation on all datasets, for the baselines and our proposed method.

Siamese architecture of shared weights, for both spatial and temporal streams, of which only the spatial is shown here (Fig 2d). The score from each split is either fed to the proposed loss functions: ranking/similarity which compute the margin ranking loss based on the pair's label (Fig 2e).

4. Results

To evaluate our method we perform four-fold cross validation on each of the four datasets with a pairwise precision evaluation metric. We use two existing ranking methods developed for other applications as baselines as there are no previous general skill determination works. Our first baseline uses RankSVM [4], with pretrained CNN features. The second baseline is Yao *et al*. [6] who originally performed deep ranking on video to determine highlight segments. They use pre-extracted features in a fully connected network to rank segments with a 'highlight score'.

Comparative results are available in Table 1. Our method outperforms both baselines on three of the four tasks. RankSVM performs best on Chopstick-Using. The improvement with our method is most significant in the Drawing task with an improvement of 11.2%.

In Fig. 1 we visualize the frame level spatial activations on example rankings from each dataset using [7]. From Fig.1 we can see that the trained model is picking details that correspond to what a human would attend to. For example, in Dough-Rolling high activations occur on holes in the dough (1, 3), curved or rolled edges (4) and when using a spoon (2). Alternatively, in Surgery, high activations occur when strain is put on the material (1, 2), with abnormal needle passes (3) and when there is loose stitching (4).

5. Conclusion

Our paper [2] presents a method to rank videos based on the skill that subjects demonstrate. Particularly, we propose a pairwise deep ranking model which utilizes both spatial and temporal streams in combination with a novel loss to determine and rank skill. We test this method on four separate datasets, two newly created, and show that our method outperforms the baseline on three out of four datasets, with all tasks achieving over 70% accuracy. Qualitative figures demonstrate the approach's ability to learn task nuances, while using a task-independent, method.

We see our work as a promising step toward the automated and objective organization of *how-to* video collections and as a framework to motivate more work in skill determination from video.

References

[1] F. De la Torre et al. Guide to the carnegie mellon university multimodal activity (CMU-MMAC) database. *Robotics Institute*, page 135, 2008. 1

[2] H. Doughty et al. Who's Better? Who's Best? Pairwise Deep Ranking for Skill Determination. In *CVPR*, July 2018. 1, 2

[3] Y. Gao et al. The JHU-ISI gesture and skill assessment dataset (JIGSAWS): A surgical activity working set for human motion modeling. In *MICCAI*, 2014. 1

[4] T. Joachims. Training linear svms in linear rime. In *Proceedings of the 12th ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 217–226. ACM, 2006. 2

[5] L. Wang et al. Temporal segment networks: Towards good practices for deep action recognition. In *ECCV*, 2016. 1

[6] T. Yao et al. Highlight detection with pairwise deep ranking for first-person video summarization. In *CVPR*, 2016. 2

[7] J. Zhang, Z. Lin, S. X. Brandt, Jonathan, and S. Sclaroff. Top-down neural attention by excitation backprop. In *European Conference on Computer Vision (ECCV)*, 2016. 2