Master Thesis

Skill Level Assessment from Videos with Spatial Attention

（空間的注意を用いたビデオからの技能レベルの評価）

Zhenqiang Li
Advisor: Professor Yoichi Sato

Submission Date: Jan 31st, 2019

Department of Information and Communication Engineering
Graduate School of Information Science and Technology
The University of Tokyo

Advisor

Prof. Yoichi Sato

Abstract

Recent advances in computer vision have made it possible to automatically assess from videos the skills level of humans in performing a task, which breeds many important applications in domains such as health rehabilitation and manufacturing. However, assessing human performance from a video is a challenging task even for humans without expert knowledge in the corresponding domain. Although several works surrounding this topic have been proposed previously, the performance of their approaches are still limited in complicated and general situations and there is much room for improvement. Therefore, the goal of this work is to develop a computer-vision technique, which acts like a human expert in terms of assessing skill levels by learning from a large amount of training data.

The main challenges of skill level assessment lie on two aspects. The first one is the challenge of representation extraction of videos. The key to an accurate skill assessment system relies on an effective feature extraction mechanism, which is expected to offer discriminative representation for videos with different performance levels of a task. The second challenge is about the annotation of training data. Since a set of exact evaluation criteria is not achievable for all tasks at this moment, annotating every video with an exact and objective score is difficult even for an expert, which becomes an impediment for dataset construction.

To address the above challenges, we present an innovative pairwise-ranking-based approach by incorporating the mechanism of spatial attention that performs an import role in human perception process. Previous methods tend to incorporate all the appearance information into extracted feature representation for videos' frames, which may limit their performance since the frames always contain some redundant information and only a part of video regions is critical for skill assessment. Our motivation here is to model human attention in videos which helps to focus more on relevant video regions for better skill assessment. In particular, we leverage the procedure of people forming attention in viewing a video and propose a novel deep model that learns spatial attention automatically from videos in an end-to-end manner. Then, differing from most of the previous works which modeled the task of skill assessment as an element-wise regression problem, we adjust it as a pairwise-ranking problem by grouping all the videos in a dataset into pairs and for each pair, annotating the superiority in skills performed by two videos. This not only alleviates the difficulty of data annotation but also implements one way of data augmentation.

We evaluate our approach on a newly collected dataset of infants grasping task and four existing datasets of hand manipulation tasks. Experiment results demonstrate that state-of-the-art performance can be achieved by considering attention in automatic skill assessment. Furthermore, we analyzed the visualization results of spatial attention maps to reveal how our spatial attention module helps to discriminate different skill levels from videos and why our approach performs differently on different datasets. The analysis results indicate us the possible extension of our method. As the skill level inclines to be contained in the long-range variance of action, how to encode temporal semantics over long videos and capture long-range temporal relationships better is a problem need to be considered in future works.

Contents

List of Figures

List of Tables

1 Introduction

1.1 Overview

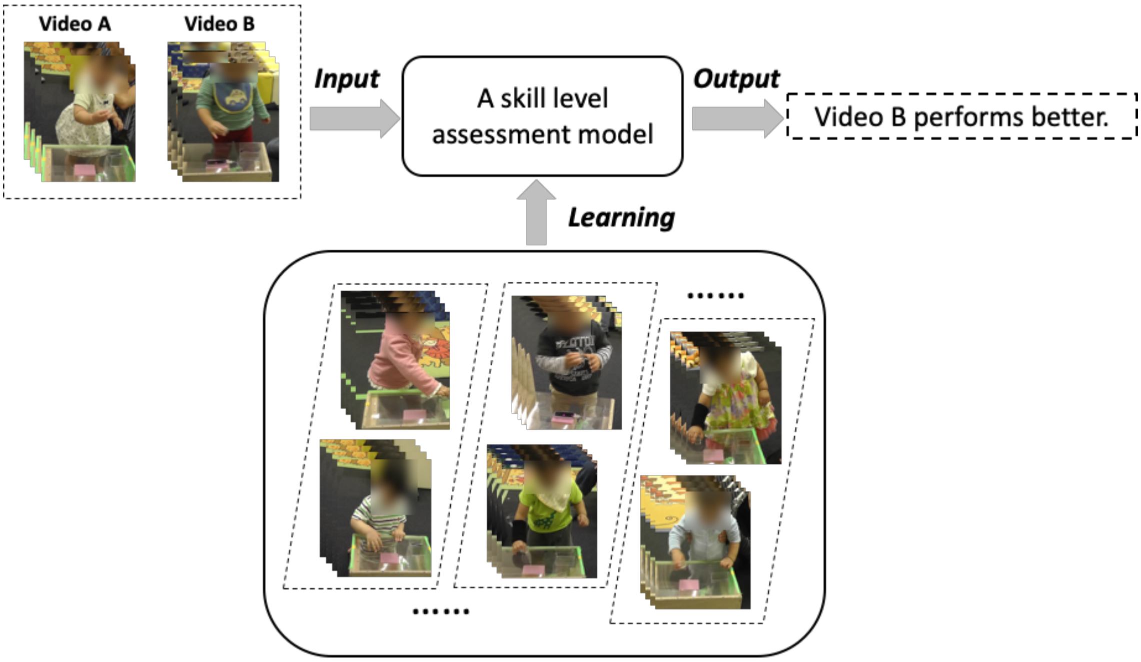


Figure 1.1: The concept of skill assessment model. We attempt to build a skill assessment model, which has the potential to learn the expert knowledge from a large amount videos. Then when given two video performing the same task, e.g., playing basketball, the model could point which one performs better like an expert.

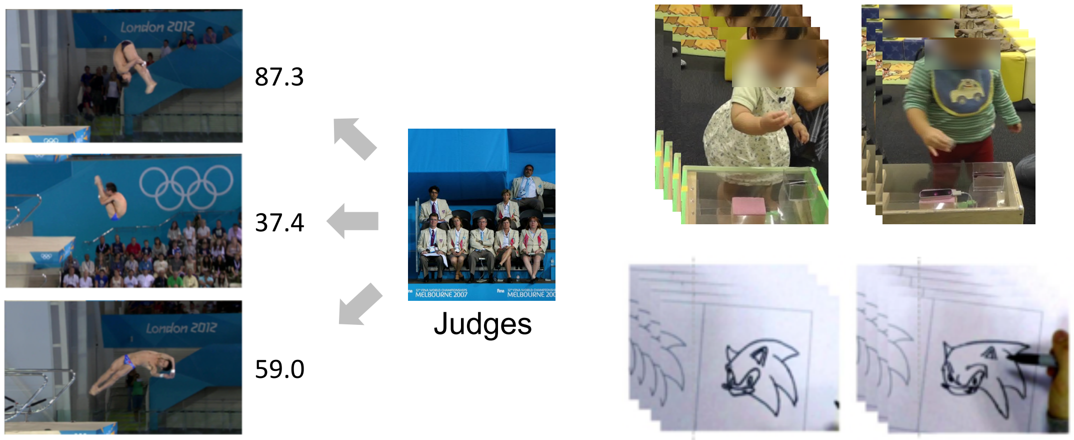
Skill level assessment is a type of evaluation often used to determine the skills and abilities a person has. For example, a pediatrician would assess the motor skills of infants to diagnose their developmental progress, and a factory owner would assess the skills of his employees in order to improve their work performance. While a professional supervisor may find it easy to assess the skills of his apprentice, skill assessment becomes difficult where no professional supervisor exists, e.g., in rural areas.

With recent advances in computer vision, automatic skill assessment from videos is believed to have many potential applications and begins to attract research interests in recent years [IMG03, MVCH14, PVT14, BSPYS17, DDMC17]. In particular, the effectiveness of deep learning techniques in automatic skill assessment has been demonstrated in [DDMC17]. However, in this method, the skill level is determined based on low-level features extracted globally from the whole image, which may be limited in capturing fine details of motion. Skill assessment requires careful observation of the details of the action, of which the location is usually a relatively small region. For example, when experts assess the shooting skill of a basketball player, the hand region rather than the player's face or dressing, should be paid more attention to.

Moreover, the region of attention is not determined solely from the current frame of the video, as knowledge accumulated from previous observations also plays a significant role. Let us consider the procedure of a human expert assessing the skills from a video, which could provide useful inspiration for network design. It has been observed that spatial attention is dependent on the past observation of the video. An expert cannot find an appropriate region to focus only based on the current video frame. Instead, he would first consider the history of this video, and then locates the critical region on the current frame. For instance, when assessing the shooting skill of a basketball player, the attention is first paid to the pose of two hands holding the ball, and when the player shoots, the attention will no longer be paid on the hands but moves to the trajectory of the ball towards the basket. Therefore, a network should find its attention not only based on the low-level information of the current frame, but also the past high-level knowledge about the whole process.

Technically, we propose a deep model for skill assessment, in which a novel recurrent neural network (RNN) based module is developed for estimating spatial attention. At each time step, our model first encodes the appearance and motion information of the input frame with fully convolutional layers. The proposed spatial attention module estimates a spatial attention map by jointly considering the encoded features of the current frame and the skill-related knowledge constructed from previous observation. The encoded features are then filtered via weighted average pooling using the estimated attention map. The attention-filtered features are used as input to a temporal aggregation module (implemented with RNN) which updates the skill-related knowledge temporally. After accepting all the frames, the accumulated information will be used to assess the skill in a video.

To evaluate our approach, we use existing public datasets of hand manipulation tasks as well as a newly collected dataset which records visuomotor skills of infants at different ages (called as “Infant Grasp Dataset"). To alleviate the difficulty of annotation, we annotate the videos pair-wisely and use a pairwise deep ranking technique for training our model. Using pairwise ranking is also a way for data augmentation. The Infant Grasp Dataset contains 4371 video pairs from 94 videos of object grasping task. Experimental results show that our proposed approach not only achieves state-of-the-art performance but also can learn meaningful attention for video-based skill assessment.



(a)

(b)

Figure 1.2: The challenge of annotating training data for skill assessment. (a) In some cases, exact and objective scores are available as the annotation, e.g., the scores given by judges in the diving contest. (b) In most cases, since the absolute criterion is unavailable, it is difficult to annotate the skill in each video with an absolute score.

1.2 Challenges and Contributions

As for the training architecture, one part of past researches on skill assessment was formulated as the element-wise regression frameworks, which raises a challenge for training data collection and annotation. In element-wise regression, the skill is represented as a continuous variable and is estimated from each video. For example, [PVT14] firstly proposed a regression framework for learning to assess the quality of human-based actions from videos. They represented human actions using spatiotemporal pose features and learned a linear SVM regression model to predict the scores of actions. However, training a regression model often requires a large number of annotations of absolute scores, which requires a lot of annotation effort and is sometimes unavailable for tasks that have no clear definition of performance levels. As shown in Figure 1.2, for the task of diving, the absolute scores given by judges in contests could be utilized as the annotation. However, for the other tasks (e.g., drawing and playing basketball) where the specific criterion for quantifying the skill is unavailable, it is difficult to annotate the skill in each video with an absolute score. Because of this challenge of annotation, a skill assessment model based on the element-wise regression framework becomes hard to be trained.

Another challenge is the extraction of effective feature representation of the videos which represent variant performance levels of a task. In the previous work [DDMC17], each input video is randomly sampled into a fixed number of frames. From each frame, one feature with a global receptive field is extracted through the global pooling operation and one score is generated based on this feature. The average of the

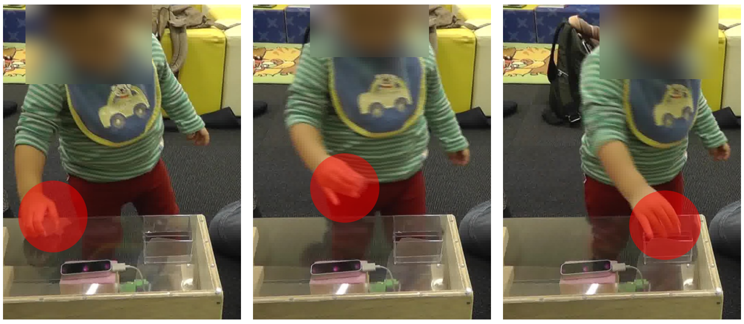


Figure 1.3: The challenge of extracting representation from videos for skill assessment. When assessing skills in videos, rather than processing all information in each frame, human tends to focus on the task-related significant regions (spatial attention).

scores from sampled frames is utilized to estimate the skill in the whole video. This means that this approach treats every information in spatial fairly and also does not model the relationship of information in temporal explicitly, although human tends to pay more attention to the task-related information in spatial and connect the information in temporal when assessing skill from a video. Figure 1.3 shows a group of frames sampled from a video along with the task-related significant regions, which keep consistent temporally.

Main contributions of this paper are summarized as follows:

- We propose a novel attention-based deep learning approach for skill assessment from videos. To the best of our knowledge, this is the first work to consider attention mechanism in skill assessment.
- We designed a new spatial attention module which is distinctive from previous attention modules and adaptive to the task of skill assessment from videos.
- We collect and annotate a new dataset for skill assessment, which records object-grasping task of infants at different ages.
- Our method has achieved state-of-the-art performance in skill assessment. The visualization results also proved the effectiveness of our spatial attention module.

1.3 Thesis Outlines

The rest of this thesis is organized as follows. In Chapter 2, we first provide an overview of recent related works on skill assessment, attention mechanism, and deep ranking framework. We present our approach in Chapter 3, our model architecture includes three different parts for feature encoding, attention pooling, and temporal aggregation. In Chapter 4, we evaluate our method and show its superiority over other baseline methods. As a discussion, in Chapter 5, we show more visualization results to reveal how the learned spatial attention helps discriminate different skill levels from videos and why our approach performs differently on different datasets. Finally, Chapter 6 summarizes this thesis.

2 Related Work

In this section, we review past skill assessment works in video, both for certain specific tasks like sports or surgical tasks and general skill assessment works. We also introduce the attention mechanism in video representation. Finally, we relate this work to deep ranking approaches, on which our method is based.

2.1 Skill assessment

There are only a few works aiming at skill assessment and most of which focus on surgical tasks, probably due to the intensive needs in this area [MVCH14, SBP$^+$14, ZL11, ZL15, ZSB$^+$15]. For example, Zia *et al.*utilize the repetitive nature of surgical tasks and rely on the entropy of repeated motions to identify different skill levels. In [MVCH14], a combination of video and kinematic data is used to rank skill in two surgical tasks. Sharma *et al.*[SBP$^+$14] use motion textures to predict a measure of skill specific to the surgical domain called ”the OSATS criteria”.

There are also some works on sports skill assessment [BSPYS17, ÇAWS13, IMG03, JPDK03, PVT14, PM17]. For instance, [BSPYS17] proposed a pairwise deep ranking model to assess the skill of a basketball player from first-person videos (Figure 2.1). The goal of this work is to predict a performance measure customized to a particular evaluator from a first-person basketball video. So the model is trained by pairs of first-person basketball videos which are weakly labeled by an evaluator. To leverage the spatiotemporal visual semantics information provided by a first-person video, the model exploits a convolutional LSTM network to detect atomic basketball events and produce spatiotemporal features for assessment. Moreover, to address the issue of severe global motions in first-person videos, the network is trained to be able to zoom-in to the salient regions. However, the method proposed in this paper is specifically designed for the basketball game, which is limited and hard to generalize to other tasks. [ÇAWS13, PM17] focus on the quality of motion, and [PVT14] (Figure 2.2) uses a regression framework and pose-based feature for learning to assess the quality of human actions from videos. These works also have the limit in generalizability since either appearance information or motion information is lost. Neither quality of motion nor appearance on its own is an essential condition to determine skill level [DDMC17].

A number of previous methods only regards the skill level in a coarse manner: In [ZSB$^+$15] the level of skill is only split into novice and expert. These works

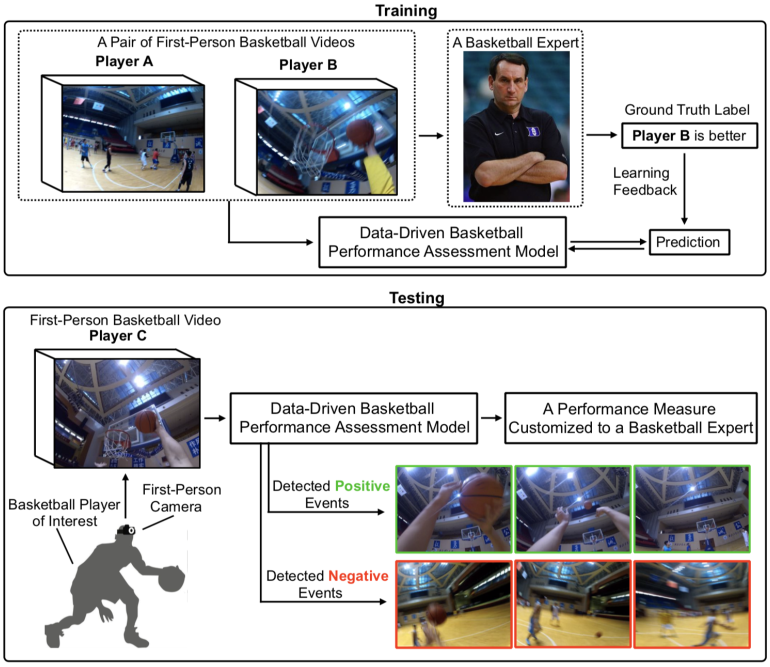

Training
A Pair of First-Person Basketball Videos
Player A
Player B
A Basketball Expert
Ground Truth Label
Player B is better
Learning Feedback
Data-Driven Basketball Performance Assessment Model
Prediction
Testing
First-Person Basketball Video
Player C
Data-Driven Basketball Performance Assessment Model
A Performance Measure Customized to a Basketball Expert
Basketball Player of Interest
First-Person Camera
Detected Positive Events
Detected Negative Events

Figure 2.1: [BSPYS17] proposed a model to assess a basketball player's performance based on an evaluator's criterion from an unscripted first-person basketball video of a player. The model is trained by pairs of weakly labeled first-person basketball videos. During testing, the model predicts a performance measure customized to a particular evaluator from a first-person basketball video. The model is specificiallt designed for the basketball game and is limited to generalize to other tasks.

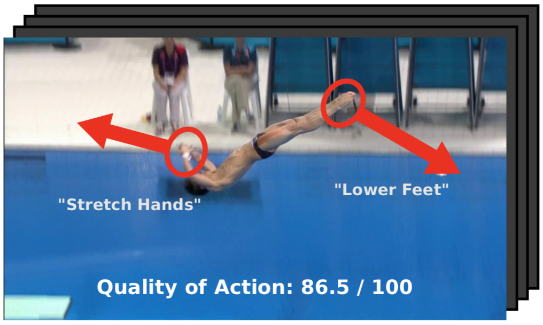


Figure 2.2: [PVT14] introduced a learning framework for assessing the quality of human actions from videos. This approach works by training a regression model from spatiotemporal pose features to scores obtained from expert judges. Interpretable feedback on how people can improve their action could also be provided by this approach. However, because only pose-based features are utilized, prediction results are severely affected by the accuracy of pose estimation.

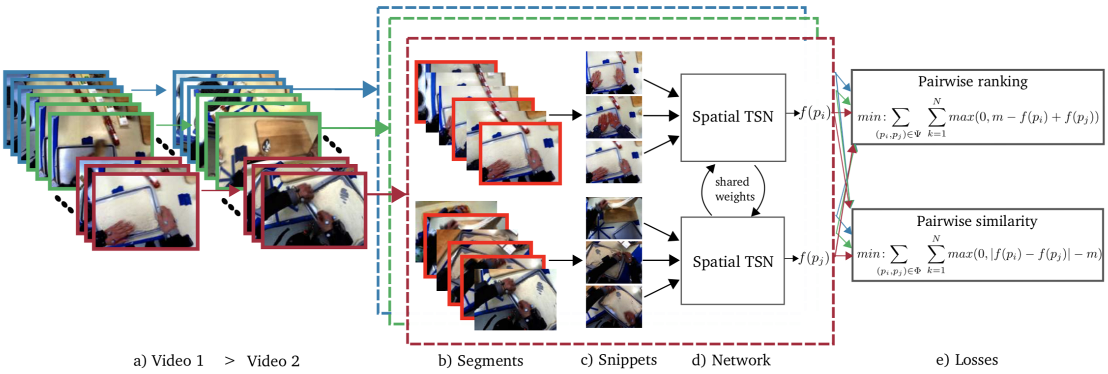

Spatial TSN
$f(p_i)$
shared weights
Spatial TSN
$f(p_j)$
Pairwise ranking
$min: \sum_{(p_i,p_j)\in\Psi} \sum_{k=1}^{N} max(0, m - f(p_i) + f(p_j))$
Pairwise similarity
$min: \sum_{(p_i,p_j)\in\Phi} \sum_{k=1}^{N} max(0, |f(p_i) - f(p_j)| - m)$
a) Video 1 > Video 2
b) Segments
c) Snippets
d) Network
e)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

Figure 2.3: [DDMC17] proposed a two-stream pairwise deep ranking model with a newly designed loss function for skill assessment. However, the method is purely bottom-up, without using the high-level information related to task or skill to guide the bottom-up feed-forward process, which causes all the information including redundancy are included in the extracted feature for each frame.

[GVR⁺14, ZSB⁺15] determine skill labels by participants' previous experience, but not their performance in individual videos. In this work we aim to rank the skill in each video instead of classifying the videos as expert or novice.

Perhaps the most similar work with ours is [DDMC17], in which a two-stream pairwise deep ranking model with a newly designed loss function is used for skill assessment (Figure 2.3). However, their method is purely bottom-up, without using the high-level information related to task or skill to guide the bottom-up feed-forward process. This may result in worse performance since the bottom-up feature extracted from each frame will contain not only the skill-related information but also the redundant information of the variance in background or regions irrelevant to the task. In this work, we use the attention mechanism to guide the bottom-up feed-forward process. Our attention is learned from both low-level information globally extracted from each frame and skill-related information accumulated in previous observation, which is able to dynamically generate spatial attention maps to guide the bottom-up features, thus achieving a better performance.

2.2 Attention mechanism

Evidence from human perception process shows the importance of attention mechanism [MHG⁺14], which uses top information to guide bottom-up feed forward process [WJQ⁺17]. Recently, tentative efforts have been made towards applying attention into deep neural network for various tasks like image recognition [HSS17, WJQ⁺17, WPLSK18], visual question answering (VQA) [DAZ⁺17, LYBP16], image and video captioning [AHB⁺18, CZX⁺17, CZ18, GGZ⁺17] and visual attention prediction [HCK⁺17, SCD⁺17, HCLS18, HCLS19].

2.2.1 Attention mechanism in image representation

There have been several attempts to incorporate attention processing to improve the performance of Convolutional Neural Networks (CNNs) in large-scale classification tasks. Wang *et al.*[WJQ$^+$17] proposed Residual Attention Network which uses an encoder-decoder style attention module. By refining the feature maps, the network not only performs well but is also robust to noisy inputs. Hu *et al.*[HSS17] introduced a compact module to exploit the inter-channel relationship. In their Squeeze-and-Excitation networks (SENet), they use global average-pooled features to compute channel-wise attention. Woo *et al.*[WPLSK18] argued that the way to infer channel attention in SENet is suboptimal and it is insufficient to only consider channel attention. To address this, they proposed a Convolutional-Block-Attention module (CBAM), in which both the channel attention and the spatial attention are exploited based on not only average-pooled features but also max-pooled ones. Not only in image recognition task, the CBAM module also performed its effectiveness in image detection task.

In the tasks of image captioning and video question answering (VQA), the modules for referring attention are usually combined with a Recurrent Neural Network (RNN) architecture, since the generation of sentences requires dynamic modulation of attention. In [CZX$^+$17], the spatial and channel-wise attention was incorporated in a CNNs+LSTM architecture to encode visual attention in deep feature maps, which helps the LSTMs to access accurate visual information for dynamic sentence generation. Anderson *et al.*[AHB$^+$18] leveraged a combined bottom-up and top-down attention mechanism to VQA. The bottom-up mechanism is based on Faster R-CNN that is designed for object detection and can propose image regions, each with an associated feature vector, while the top-down mechanism estimates feature weightings. Differing from previous works, the attention is calculated at the level of objects rather than gridded image regions, and the top-down attention is dynamically updated based on the states of RNN layers.

2.2.2 Attention mechanism in video representation

In video representation, the majority of works utilize the attention mechanism in action recognition [LWH$^+$17, LGG$^+$18, LYD$^+$18, YSLZ18] and action localization [CZM17]. In [SLX$^+$17], an end-to-end spatiotemporal attention model is used to recognize action using skeleton data. They use LSTM-like structure and construct joint selection gates and frame selection gates for modeling spatial and temporal attention. Girdhar *et al.*[GR17] propose an attentional pooling method based on second-order pooling for action recognition. Both saliency-based and class-specific attention are considered in their model, however, the attention is learned statically from each frame where no temporal information between frames is considered. [DWQ17] incorporates a pose-based attention mechanism into recurrent networks to learn complex temporal motion structures for action recognition. The overall

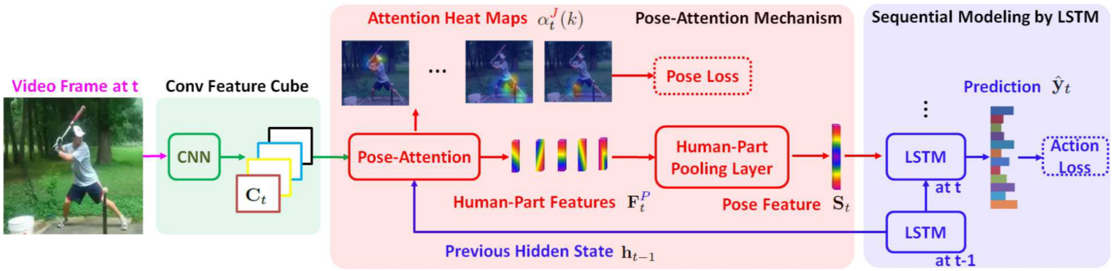

Attention Heat Maps $\alpha_t^J(k)$
Pose-Attention Mechanism
Sequential Modeling by LSTM
Video Frame at t
Conv Feature Cube
...
Pose Loss
Prediction $\hat{\mathbf{y}}_t$
CNN
$\mathbf{C}_t$
Pose-Attention
Human-Part Pooling Layer
LSTM
Action Loss
at t
Human-Part Features $\mathbf{F}_t^P$
Pose Feature $\mathbf{S}_t$
LSTM
Previous Hidden State $\mathbf{h}_{t-1}$
at t-1

Figure 2.4: [DWQ17] proposed an end-to-end Recurrent Pose-Attention Network (RPAN), in which a pose-based attention mechanism is incorporated to learn complex temporal motion structures from redundant frames for action recognition. At each time step, the pose attention module produces several human-part-related features from the convolutional feature cube under the guidance of the previous hidden state of LSTM. Since the pose attention module is regularized by ground-truth poses, the generalization of this network is restricted to conditions where poses are available as ground-truth.

framework of this network is illustrated in Figure 2.4. Although the temporal relationship of attention in context is modeled by a recurrent structure, the model cannot generalize into the situation where the pose is unavailable from appearance.

In this work, we propose a new attention module for skill assessment, which makes use of skill-related knowledge to guarantee the model to focus on regions that are highly correlated with the task. Moreover, as the attention in a frame is not independent of the attention in context, we incorporate the recurrent networks into our model to leverage the relationship of attention in continuous frames. Additionally, differing from previous works, our approach is able work in general conditions even though human body is invisible in frames since our attention module has no requirement for human poses.

2.3 Deep ranking

The most widely used ranking formulation is pairwise ranking. It was originally designed to learn search engine retrieval functions from click-through data, however, it has been adopted in other ranking applications such as relative attributes in images [PG11]. Ranking aims to minimize the number of incorrectly ordered pairs of elements, by training a binary classifier to decide which element in a pair should be ranked higher. This formulation was first used by Joachims *et al.*in RankSVM, where a linear SVM is used to learn a ranking. Firstly in [BSR$^+$05], the ranking has also been integrated into deep learning frameworks. [YMR16] uses a pairwise deep ranking framework for highlight detection in egocentric videos. The framework firstly exploits a spatial and a temporal deep CNN (DCNN) architecture to predict a highlight score for each segment which is split from a video. Then the framework takes a pair of highlight and non-highlight video segments and optimizes the DCNN architectures by a ranking loss to output a higher score of highlight segment than that of non-highlight one. In this work, we also use pairwise deep ranking as the training scheme, not only to ease the difficulty in ground truth labeling but also to provide augmented data (pairs) for training our model.

3 Proposed Method

In this section, we first describe the overview of the framework for skill assessment. Then we introduce the details of each part, especially the proposed spatial attention module which integrates temporal evolution patterns into the estimation of the spatial attention. We also describe the pairwise ranking scheme for training our model.

3.1 Model architecture

Our goal is to learn models for skill assessment in different tasks. Given a video recording a whole progress of finishing a certain task, our model estimates a score to assess the skill performed in the video. Figure 3.1 depicts the architecture of our model.

As is done in [DDMC17], we split the video into N segments, and select one frame randomly in each segment to form a sparse representation of the whole video. At every time step $t \in \{1, \cdots, N\}$, the feature encoding module extracts deep features from a single RGB image and a corresponding stacked optical flow image like [SZ14]. The spatial attention module estimates an attention map based on the low-level deep features and a high-level skill-related vector generated from the top temporal aggregation module. An attended vector at each time step is then obtained by pooling the deep features with weights derived from the attention map and fed into the top temporal aggregation module, which aggregates vectors temporally and outputs a final score. We illustrate the details of each module in the following sections.

3.2 Feature encoding

From each of the t-th segment, the module takes an RGB image I_t and the corresponding stacked optical flow images O_t as input, since appearance and motion are both important for skill assessment. The feature encoding module first extracts two deep features $F_{S,t}$ and $F_{T,t}$ from I_t and O_t by feeding them into two ResNet101 networks respectively. The ResNet101 networks are pre-trained on ImageNet [DDS$^+$09] and then fine-tuned on UCF101 [SZS12] in a two-stream framework

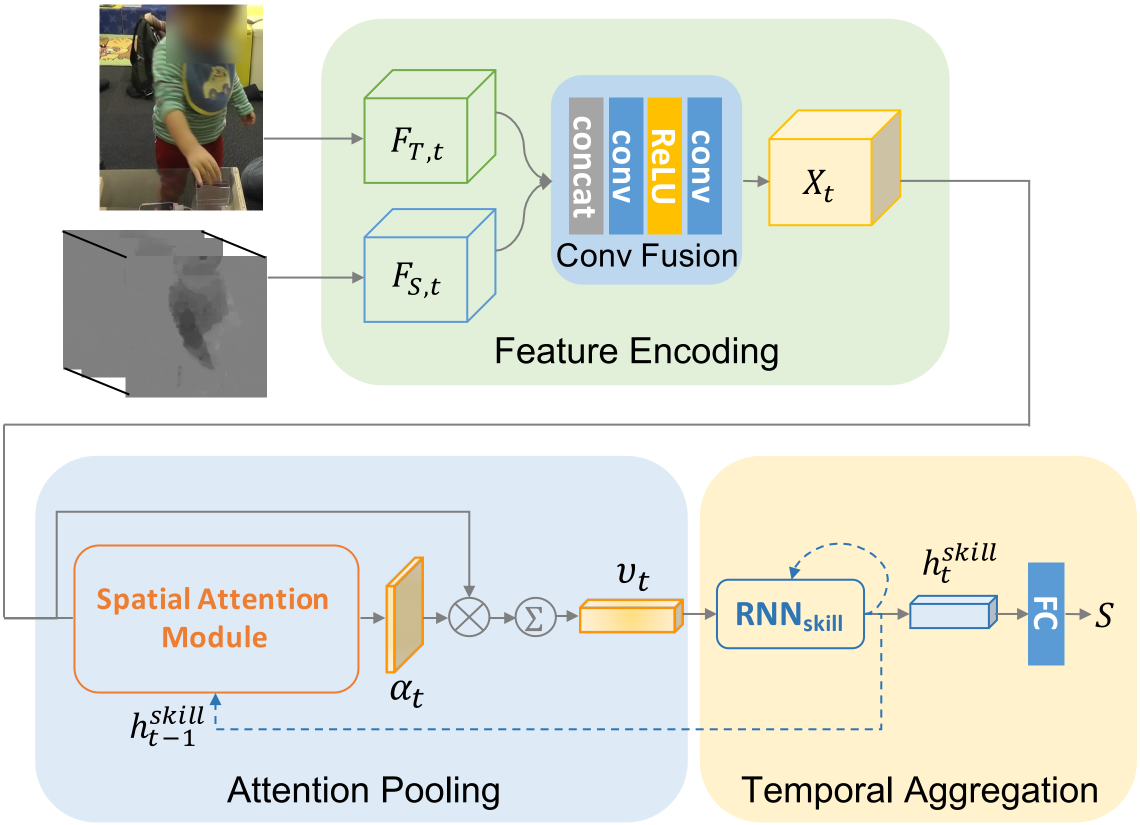


Figure 3.1: The illustration of our skill assessment framework. At every time step, the network takes an RGB image and the corresponding stacked optical flow images as input and firstly represents them as deep features. The spatial attention module is then used to generate an attention map, by integrating the global information from the deep features and the skill-related information from the top part of the framework. Meanwhile, the temporal evolution of spatial attention is also incorporated implicitly in the module. An attended feature vector is then generated by weighted pooling the deep feature according to the estimated attention map. The feature vector is forwarded to an RNN (RNN_{skill}) for temporal aggregation. The output of RNN_{skill} at the final time step is used to yield a ranking score.

for action recognition[SZ14]. As [HCLS18, GRG+17], we extract the deep features from the last convolution layer of the 5th convolutional block.

Then we build a two-layer convolution network for spatiotemporal feature fusion, which takes $F_{S,t}$ and $F_{T,t}$ as input and outputs the fused spatiotemporal deep representation $X_t \in \mathbb{R}^{C\times H\times W}$.

$$X_t = f_{conv2}(ReLU(f_{conv1}([F_{S,t}; F_{T,t}]))) \qquad (3.1)$$

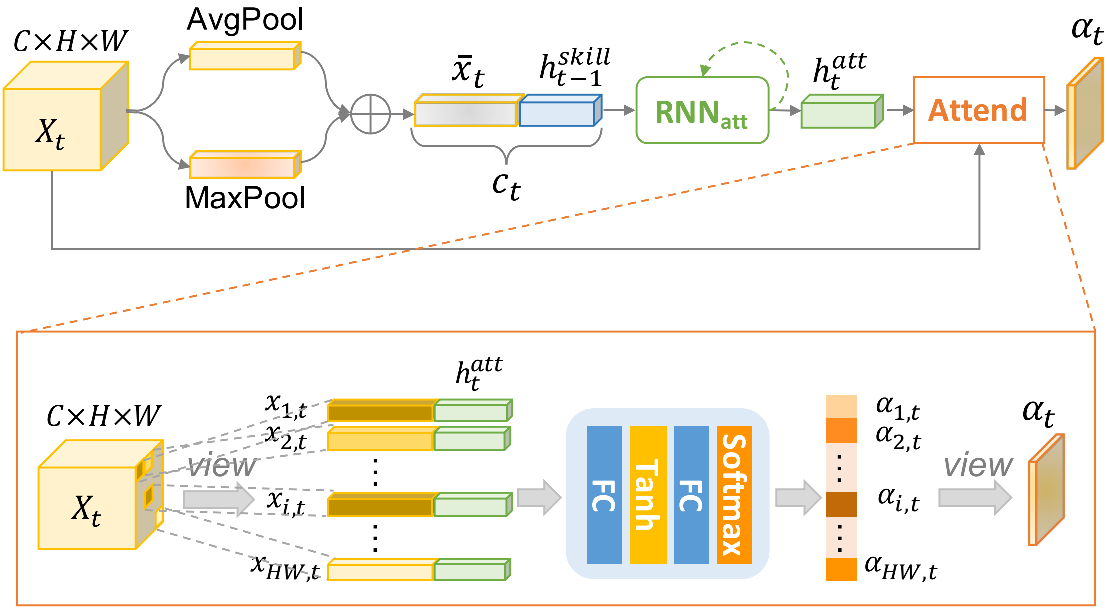


Figure 3.2: The details of the spatial attention module. To estimate the skill-related significance for the regional vectors at different locations, the module incorporates not only low-level information $\bar{x}_t$ globally extracted from the deep feature maps, but also information for skill assessment h_{t-1}^{skill} which is accumulated by a high-level RNN (RNN_{skill} in Figure 3.1). We also use an RNN (RNN_{att}) to learn the evolution pattern of attention and its output is utilized to estimate attention weights for all locations in deep feature map X_t.

3.3 Attention pooling

We aim to apply attention to guide the bottom-up feed-forward process. In our work, this is done by our proposed attention pooling module that dynamically adjusts spatial attention based on both the low-level information and the high-level knowledge of the skill. Specifically, at each time step, the attention pooling module accepts two inputs: the frame's deep feature maps as the low-level information, and the *skill-related vector* as the high-level information. The two inputs will be explained in details as follows.

To extract a compact low-level representation vector from deep feature maps, following [WPLSK18], we firstly squeeze spatial information of the deep feature map X_t by performing average-pooling and max-pooling. As a result, two features are generated and summed together to form a highly abstract low-level representation vector $\bar{x}_t \in \mathbb{R}^C$ for X_t as following:

$$\bar{x}_t = AvgPool(X_t) + MaxPool(X_t) \tag{3.2}$$

The high-level representation vector h_{t-1}^{skill} and the low-level representation vector $\bar{x}_t$ both serve as a basis for estimating attention map. We name this part as *spatial* *attention module*, and show its details (its details are shown)in Figure 3.2. Briefly speaking, we concatenate the two vectors together, getting a vector c_t for estimating the attention map.

$$c_t = Concat[\bar{x}_t; h_{t-1}^{skill}] \tag{3.3}$$

A Recurrent Neural Network (RNN) [CGCB14] (called RNN_{att}) is adopted to learn the transition pattern of attention at different time steps. The output of h_t^{att} at time step t is a vector integrating the current low-level information and the skill-related high-level information:

$$h_t^{att} = RNN_{att}(c_t) \tag{3.4}$$

Given the output vector h_t^{att} and the deep feature maps X_t, the RNN model generates an attention weight $a_{i,t}$ for each spatial location i of the deep features $x_{i,t}$ at all $H \times W$ locations of feature maps and normalize them by $softmax$ activation function. With this procedure, the attention on each spatial location will be guided by the accumulated information h_t^{att}, which leads to a better decision on the importance of each specific location.

$$a_{i,t} = \omega_a^T [tanh(W_{xa} x_{i,t} + b_{xa} + W_{ha} h_t^{att} + b_{ha})], i = 1, 2, ..., H \times W, \tag{3.5}$$

$$\alpha_t = softmax(a_t) \tag{3.6}$$

We call this part as *Attend* part in the attention pooling module (Figure 3.2).

The attended image feature which will be used as input to the final RNN for skill determination is calculated as a convex combination of feature vectors at all locations:

$$v_t = \sum_{i=1}^{H \times W} \alpha_{i,t} x_{i,t} \tag{3.7}$$

3.4 Temporal aggregation

In [DDMC17, SZ14, WXW+16], at every time step, after obtaining the high-level representation for actions of one snippet, a network based on MLP is used to estimate a score for skill determination or action recognition. However, in skill assessment, the skill level is determined by the temporal evolution of actions so it is hard to judge only from short video clips captured at one moment. For this reason, we choose to use a recurrent neural network (RNN) to aggregate the changed action information temporally and estimate the score for skill level at the final step.

Given an input sequence $\mathbf{x} = (x_1, ..., x_T)$, a standard RNN computes the hidden vector sequence $\mathbf{h} = (h_1, ..., h_T)$ and output vector sequence $\mathbf{y} = (y_1, ..., y_T)$ by iterating the following equations from $t = 1$ to T:

$$h_t = \sigma\left(W_{ih}x_t + W_{hh}h_{t-1} + b_h\right), \tag{3.8}$$
$$y_t = W_{hy}h_t + b_y, \tag{3.9}$$

where the W terms denote weight matrices (e.g. W_{ih} is the weight matrix between input and hidden vector), the b terms denote bias vectors (e.g. b_h is the bias vector for hidden state) and σ is the hidden layer activation function, typically the logistic sigmoid function.

The standard RNN suffers from the gradient vanishing problem due to its insufficient, decaying error back flow. To address this problem, the Long Short Term Memory (LSTM) architecture [HS97] was proposed. Unlike standard RNNs, LSTMs use memory cells to store and output information, allowing it to better discover long-range temporal relationships. The hidden layer H of the LSTM is computed as follows:

$$i_t = \sigma\left(W_{xi}x_t + W_{hi}h_{t-1} + b_i\right), \tag{3.10}$$
$$f_t = \sigma\left(W_{xf}x_t + W_{hf}h_{t-1} + b_f\right), \tag{3.11}$$
$$\tilde{c_t} = tanh\left(W_{xc}x_t + W_{hc}h_{t-1} + b_c\right), \tag{3.12}$$
$$c_t = f_t \odot c_{t-1} + i_t \odot \tilde{c_t}, \tag{3.13}$$
$$o_t = \sigma\left(W_{xo}x_t + W_{ho}h_{t-1} + b_f\right), \tag{3.14}$$
$$h_t = o_t \odot tanh\left(c_t\right), \tag{3.15}$$

where σ is the logistic sigmoid function, $\odot$ denotes an element-wise production, and $i, f, \tilde{c_t}, c$ and o are respectively the *input gate*, *forget gate*, *cell candidate state*, *cell state*, and *output gate* activation vectors. By default, the value stored in the LSTM cell c is maintained unless it is added to by the input gate i or diminished by the forget gate f . The output gate o controls the emission of the memory value from the LSTM cell.

Gated recurrent unit (GRU), which was proposed in 2014 [CGCB14], is an simpler variant of LSTM. A GRU maintains two gates, reset gate and update gate to control the information flow. When the response of the reset gate is close to 0, the new gate n_t is forced to ignore the previous hidden state and reset with the current input only. The update gate controls how much information from the previous hidden state will be carried over to the current hidden state h_t. The hidden state acts in a way similar to the memory cell in LSTM. For a GRU layer, the recursive computations of activations of the units are as follows:

$$r_t = \sigma\left(W_{xr}x_t + W_{hr}h_{t-1} + b_r\right), \tag{3.16}$$
$$z_t = \sigma\left(W_{xz}x_t + W_{hz}h_{t-1} + b_z\right), \tag{3.17}$$
$$n_t = tanh\left(W_{xn}x_t + b_{xn} + r_t \odot \left(W_{hn}h_{t-1} + b_{hn}\right)\right), \tag{3.18}$$
$$h_t = z_t \odot h_{t-1} + \left(1 - z_t\right) \odot n_t. \tag{3.19}$$

We aggregate the feature vectors temporally using an RNN. The output at the final temporal step of the RNN (RNN_{skill}) is followed by a fully connected layer (FC) to get the final output score $S \in \mathbb{R}$.

$$h_t^{skill} = RNN_{skill}(\{v_1, v_2, \cdots, v_t\}) \tag{3.20}$$

$$S = FC(h_N^{skill}) \tag{3.21}$$

Here the skill-related vector h_t^{skill} is the high-level information about the video's skill level. Besides being used to predict the final output score of the video, the skill-related vector is also fed into the attention pooling module mentioned previously.

3.5 Training and implementation details

We use a pairwise ranking framework [YMR16, DDMC17] for training, which requires only pairwise annotations to assess the skill of videos performing the same actions. To be more precise, given $M = m(m-1)/2$ pairs of videos formed from a set of m videos $\{V_1, V_2, \cdots, V_m\}$, the annotation only needs to rank the relative skill level of each video other than giving an exact score of a video:

$$P(V_i, V_j) = \begin{cases} 1 & V_i \text{ performs better than } V_j \\ -1 & V_j \text{ performs better than } V_i \\ 0 & \text{no skill preference} \end{cases} \tag{3.22}$$

Since $P(V_i, V_j) = -P(V_j, V_i)$, we only need to annotate one score for one pair of videos. In our pairwise ranking framework, two videos in one pair are fed into a Siamese architecture consisting of two same models with shared weights.

Assume the output of our model is $S(\cdot)$, denote Ψ as all pairs of videos in (a) training set that contains skill preference, and let the video pairs $\langle V_i, V_j \rangle$ in Ψ to be $\forall V_i, V_j \in \Psi, P(V_i, V_j) = 1$, the model learns to minimize the following loss function:

$$L = \sum_{V_i, V_j \in \Psi} max(0, -S(V_i) + S(V_j) + \epsilon) \tag{3.23}$$

$S(V_i)$, $S(V_j)$ depict the predicted skill measure for videos 1 and 2 respectively. ϵ denotes margin, which is incorporated to adjust the distance between the predicted scores for (of) the two videos. In this work, we use $\epsilon = 0.5$ in all experiments.

We use PyTorch [PGC$^+$17] to implement our framework. The optical flow images for motion input are extracted by TV-L^1 algorithm [ZPB07]. For the dataset of Infant-Grasp, the optical flow images are extracted with the original frame rate and for the other datasets, we use the frame rate of 10-fps since motion is slow in these videos. The deep features in Figure 3.1 is extracted from the output of the 5-th convolution block ($conv5\_3$) of ResNet101 [HZRS16]. The input images are resized to 448×448, so the size of deep features extracted from ResNet is $2048 \times 14 \times 14$. The conv-fusion module consists of 2 convolution layers, in which the first layer followed by ReLU activation. The first layers have 512 kernels with a size of 2×2, and the second convolution layer has a kernel size of 1 with 256 output channels. The dimensions of parameters $\{\omega_a, W_{xa}, b_{xa}, W_{ha}, b_{ha}\}$ in attend part of spatial attention module are set as $\{1 \times 32, 32 \times 512, 32 \times 1, 32 \times 128, 32 \times 1\}$.The RNN for both attention pattern learning and temporal aggregation are implemented with a 1-layer Gated Recurrent Unit (GRU) [CGCB14] whose hidden state size is set as 128. We uniformly split each video into 25 segments, and sample one frame randomly from each split during training. The last frame of each segment is utilized for (to test) testing our model. We use stochastic gradient descent with a momentum of 0.9 to optimize our model. We set learning rate as 5E-4 for the Infants-grasp dataset, and 1e-3 when for other datasets. All weight decays are set as 1e-3. As [DDMC17], our model is trained and tested separately on different datasets.

4 Experiments

We evaluate our method on our newly collected dataset as well as four public datasets. Similarly to [DDMC17], we report the results yielded by four-fold cross-validation, and for each fold, we use ranking accuracy as the evaluation metric. Ranking accuracy is defined as the percentage of correctly ranked pairs among all pairs in the validation set.

4.1 Datasets

4.1.1 Infant Grasp Dataset

Since the related public datasets are either small in size (e.g., up to 40 videos [DDMC17, GVR$^+$14]) or unsuitable for manipulation skill assessment (e.g., comparing skill between different diving actions [PVT14]), we construct a larger dataset for infant grasp skill assessment. The dataset consists of 94 videos, and each video contains a whole process of an infant grasping a transparent block and putting it into a specified hole. The videos were originally captured for analyzing visuomotor skill development of infants at different ages. Figure 4.1 shows representative frames selected from a pair of videos. The length of each video ranges from 80 to 530 frames with a frame rate of 60fps. This dataset is expected to be of great importance not only to the computer vision community but also to the developmental psychology community.

To annotate the dataset, we asked 5 annotators from the field of developmental psychology to label each video pair by deciding which video in a pair shows a better skill than the other or there is no obvious difference in skill. We form 4371 pairs out of 94 videos, among which 3318 pairs have skill preference (76%).

4.1.2 Public datasets

We also evaluate our method using public datasets of another four manipulation tasks: Chopstick-using, Dough-Rolling, Drawing [DDMC17], and Surgery [GVR$^+$14]. We select some representative frames from each dataset and show them in Figure 4.2. The Chopstick-using dataset contains 40 videos with 780 total pairs. The number of pairs of video with skill preference is 538 (69% of total pairs). The Dough-Rolling

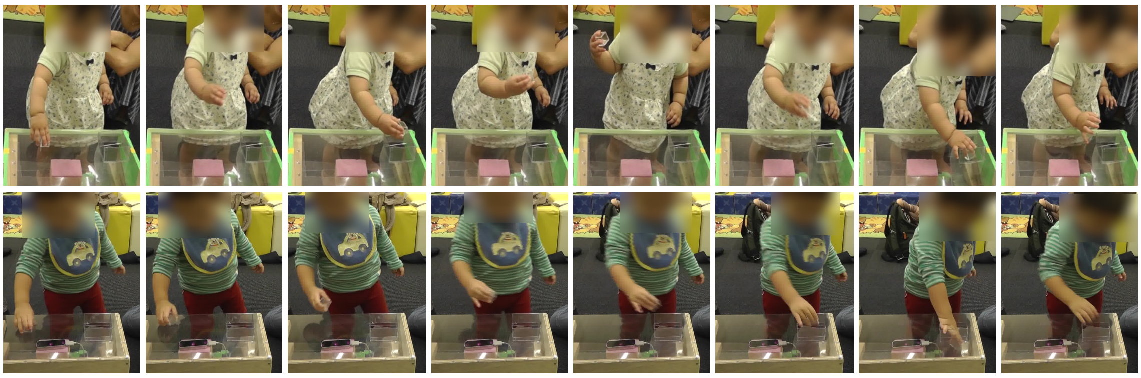


Figure 4.1: Image examples of two videos showing different skill levels in our Infant Grasp Dataset. The skill level in the bottom row is better than in the top row because the action of putting is not continuous in the top row. Infants' faces are blurred for privacy.

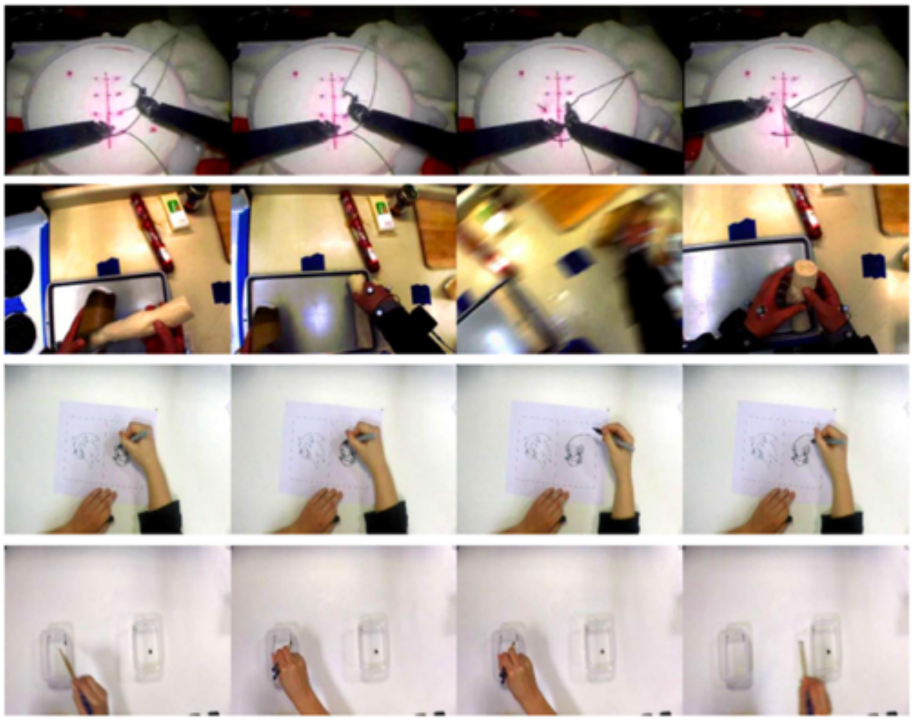


Figure 4.2: Image examples of four public datasets for skill assessment or determination. From top to bottom line, the example images are from Surgery, Dough-Rolling, Drawing and Chopstick-Using datasets respectively.

dataset selects 33 segments about the task of pizza dough rolling from the kitchen-based CMU-MMAC dataset[DlTHB$^+$08] and 538 pairs of videos are annotated with skill preference (69%). The Drawing dataset consists of two sub-dataset and 40 videos in total, among which 380 pairs are formed and 247 pairs show skill preference (65%). The Surgery dataset contains three sub-datasets of three different kinds of surgery tasks: 36 videos of Knot-Tying task, 28 videos of Needle-Passing and 39 videos of Suturing. Each sub-dataset contains a maximum of 630, 378, 701 pairs respectively, and since the annotation is given by a surgery expert using a standard and structured method, more than 90% of pairs contains the difference in skill level. Following [DDMC17], we train and test the 3 sub-datasets of the Surgery dataset together using one model. Same is done for the Drawing dataset.

4.2 Baseline methods

We compare with several baseline methods to validate the effectiveness of our proposed approach. Because previous works on skill assessment are scarce, two ranking-based approaches were introduced as baseline methods for skill assessment in [DDMC17]. The first baseline is **RankSVM** [Joa02], which has been commonly used in ranking problems [PG11]. The second baseline is the pairwise deep ranking method (**Yao *et al.*** [YMR16]) used for video ranking, although it was originally developed for a different purpose (highlight detection). Moreover, we compare our method with **Doughty *et al.*** [DDMC17] which is the most relevant work with ours. They use the TSN [WXW$^+$16] with a modified ranking loss function for skill assessment. Considering there is no previous work adopting attention mechanism into skill assessment, to also validate the effectiveness of the spatial attention module in our model, we construct a competitive baseline method which replaces our spatial attention module with a state-of-the-art attention model [WPLSK18] (**CBAM Atten.**). We implement the spatial and channel attention modules of [WPLSK18], and the attention map is obtained by feeding the encoded feature X_t into channel attention module and spatial attention module successively.

4.3 Performance comparison

We compare our method with other baseline methods and the quantitative results are shown in Table 4.1. Our method achieves best performance on all datasets and outperforms the state-of-the-art method [DDMC17] by a large margin, demonstrating the importance of adopting attention mechanism in video-based skill assessment. Our method also outperforms the CBAM Atten. that also uses attention mechanism. This validates the effectiveness of our proposed spatial attention module which considers not only the appearance of current frame but also the past high-level knowledge about the whole task process. Notably, we evaluate the CBAM

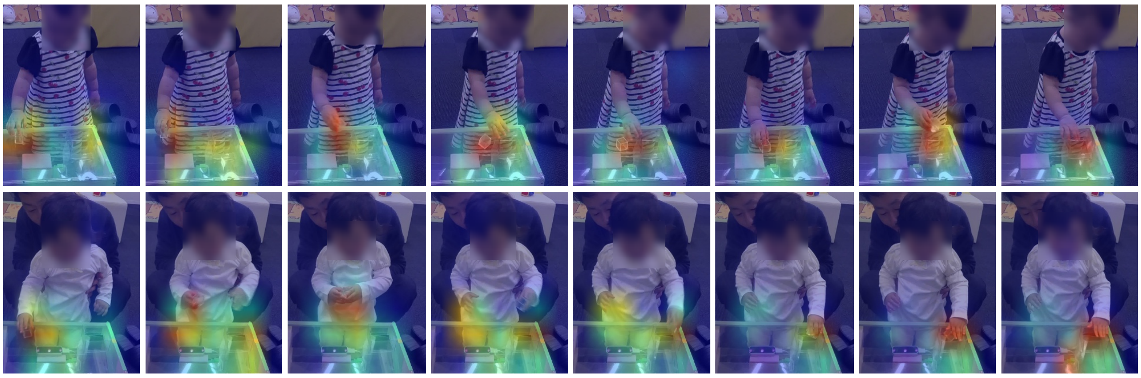


Figure 4.3: Visualization of attention maps learned from our model on the Infant-Grasp dataset.

Accuracy(%)	Chopstick-Using	Surgery	Drawing	Rough-Rolling	Infant-Grasp
RankSVM [Joa02]	76.6	65.2	71.5	72.0	N/A
Yao *et al.*[YMR16]	70.3	66.1	71.5	78.1	N/A
Doughty *et al.* [DDMC17]	71.5	70.2	83.2	79.4	80.3
CBAM Atten. [WPLSK18]	82.0	68.6	84.1	80.8	83.8
Ours	**85.5**	**73.1**	**85.3**	**82.7**	**86.1**

Table 4.1: Performance comparison with baseline methods. Ranking accuracy is used as the evaluation metric.

attention module in an architecture same with us, which mean all the parts of the architecture are kept invariant except the attention module.

We visualize the attention maps generated by our model. Figure 4.3 shows attention maps on our Infant-Grasp dataset. We can clearly see that the spatial attention focus on the critical regions of all images. In the first row, the spatial attention falls on the infant's hand at first, and when the infant accidentally drops the block (4^{th} image), our model successfully locates attention on the dropped block, instead of on the hand continuously. In the second row, the infant first grabs the block with her right hand, then passes the block to her left hand to put it to the destination. Our attention module successfully locates the correct task-related hand to emphasize. We think the reason why our attention model is adaptive to the shifting attention might be that the previously accumulated knowledge about the task is leveraged to focus on the critical events.

We also compare the attention maps of three different datasets obtained by our proposed model and the baseline attention model of CBAM [WPLSK18] in Figure 4.4. With our model, the attention is always paid onto the discriminative skill-related region, while CBAM often locates salient but task-unrelated regions. In the first example, the skill level of the chopstick-using task is determined by whether the

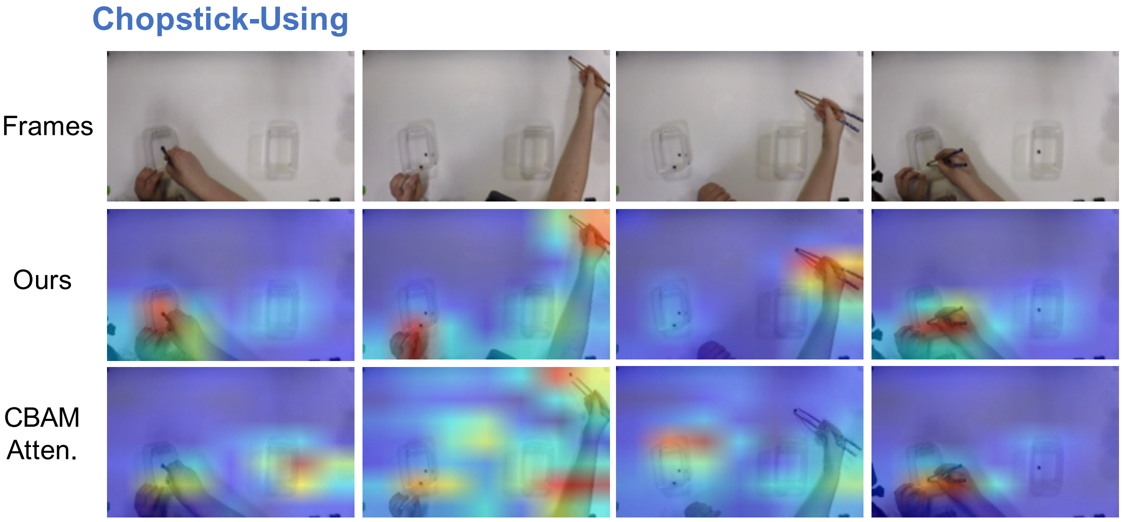


Figure 4.4: Qualitative comparison of the output attention maps of our method and the CBAM [WPLSK18] on the Chopsticks-Using dataset. It can be seen that our method successfully finds the skill-related regions.

bean could be picked up with the chopstick, and its relevant regions are successfully highlighted by our model. In the second example on Drawing dataset, our model focuses on both the drawing hand and the picture drawn by the subject (last image). In the last example on Surgery dataset, our method can locate both robotic arms in performing the task as well as the suture being manipulated.

4.4 Ablation study

To validate the effectiveness of each component of our model, we conduct ablation study on each dataset with the following baselines:

- No RNN_{att}: The component of RNN_{att} designed for learning attention evolution patterns is replaced by one fully-connected layer, which also accepts $\bar{x}_t$ and h^{skill} as input.
- $\bar{x}_t$ based Atten.: The RNN_{att} accepts only $\bar{x}_t$ as input without the concatenation with h^{skill}. We build this baseline to see the influence of the high-level skill-related knowledge in skill assessment.
- h^{skill} based Atten.: The RNN_{att} accepts only h^{skill} as input. We build this baseline to see the influence of low-level features.

Table 4.2 shows quantitative results of different subsets of our full model. The results validate our thought that the evolutionary patterns of attention maps (RNN_{att}), the low-level appearance-based information ($\bar{x}_t$) and the high-level task-related knowledge (h^{skill}) are all important for manipulation-skill assessment.

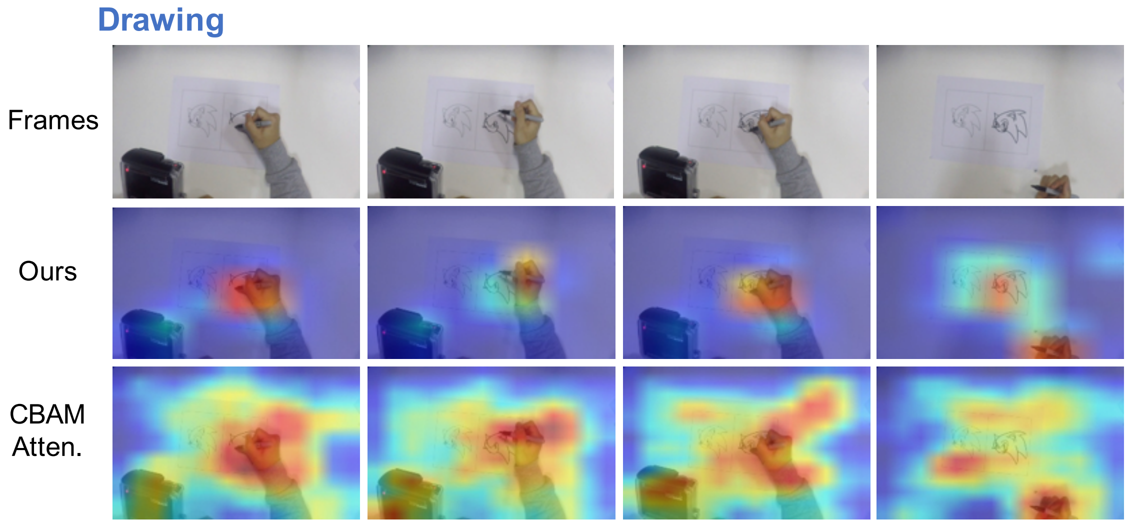


Figure 4.5: Qualitative comparison of the output attention maps of our method and the CBAM [WPLSK18] on the Drawing dataset. It can be seen that our method successfully finds the skill-related regions while the attention maps obtained by CBAM tend to only highlight the salient regions.

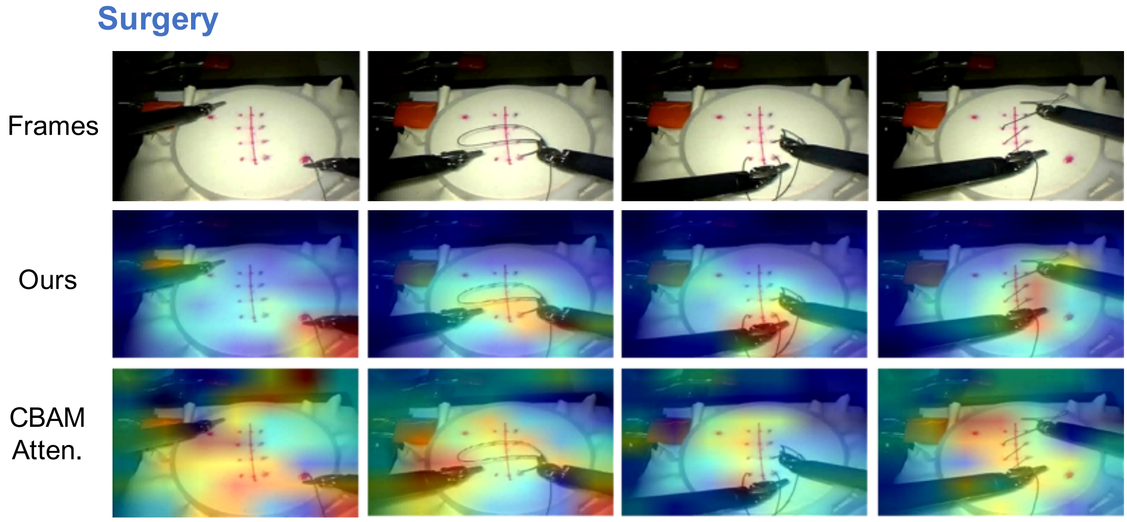


Figure 4.6: Qualitative comparison of the output attention maps of our method and the CBAM [WPLSK18] on the Surgery dataset.

Acc(%)	Chopstick-Using	Surgery	Drawing	Rough-Rolling	Infant-Grasp
No RNN_{att}	84.1	70.8	82.4	82.0	85.1
$\bar{x}_t$-based Atten.	84.1	70.1	83.4	81.8	85.3
h^{skill}-based Atten.	82.8	69.1	84.8	81.6	84.7
Full model	**85.5**	**73.1**	**85.3**	**82.7**	**86.1**

Table 4.2: Ablation study for different components of our model. Ranking accuracy is used as the evaluation metric.

We also train our model with randomized labels to examine whether our model can learn meaningful skill-related attention. As shown in Figure 4.7, when the model is trained with randomized labels, the spatial attention maps become meaningless compared with the one trained with real labels. This indicates that our model is able to discover meaningful skill-related regions rather than just highlighting the salient regions regardless of the underlying skills.

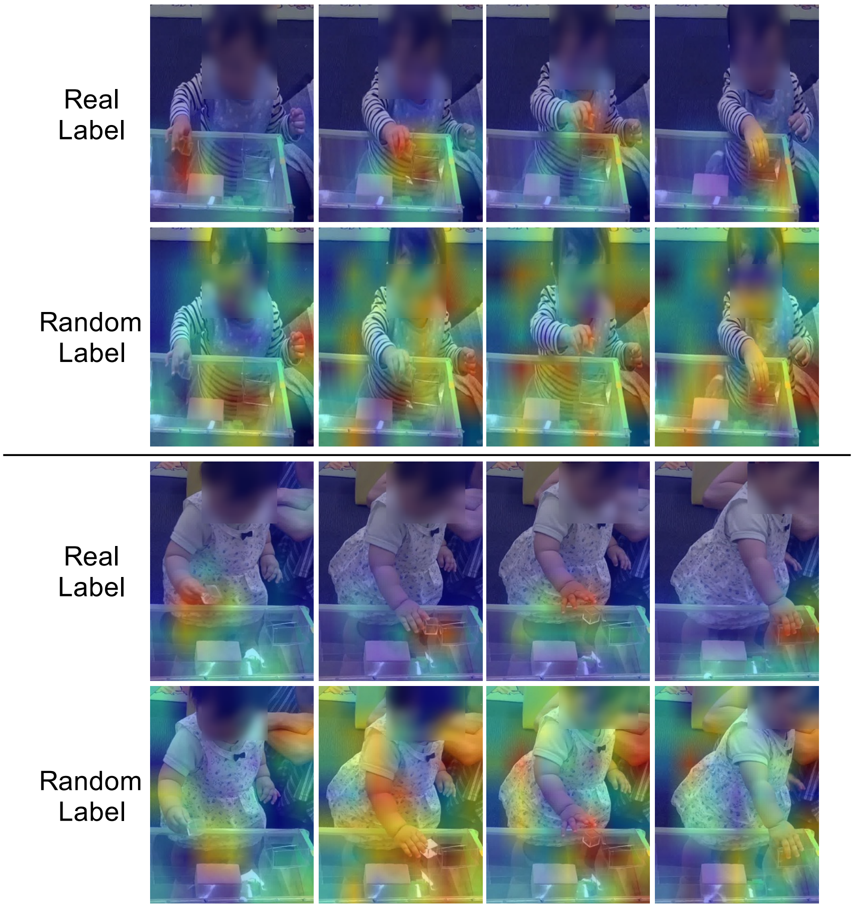


Figure 4.7: Attention map visualization between our model trained with real ground truth labels and randomized labels. The attention maps tend to be chaotic when the model is trained with randomized labels, which shows our model's ability to discover the correct skill-related regions as attention maps.

5 Discussion

5.1 Visualization of spatial attention with skill comparison

In the section of experiments, we have shown that by modeling human attention in assessing skills from a video, our proposed method achieves the best performance of skill assessment compared with previous methods. Here we show more details about how the learned spatial attention helps to discriminate different skill levels from videos.

Two video pairs from Infant Grasp dataset, each having comparably better and worse skill levels, are shown in Figure 5.1. The task for infants at different ages in this dataset is to grab a block given by an adult at the left side of a table and drop it into a hole at the right side of the table. For the first pair, the performance in the second row is considered worse than that in the first row since the block is dropped during being moved. For the second pair, the video in the third row performs better than the video in the fourth row according to the output scores of our model. We consider the reason is that the block in the fourth row is not successfully put into the specified hole after being shaken several times, while the task is successfully executed in the third row. Notably, the attention is always focused on the skill-related hand, even though the action of manipulating the block and performance are variant.

5.2 Analysis of experimental results on Surgery dataset

Among the five datasets, the performance on the Surgery dataset is the worst, no matter by our model or by the baseline approaches. After careful analysis, we consider the reason as the sparsely sampled inputs of the model, which is commonly used in models for action recognition[WXW$^+$16, GRG$^+$17], video classification[MLS17] and previous work of skill determination[DDMC17]. To comparing with the previous work[DDMC17], we sample 25 frames from each video and input them to models, even though most of the videos last for more than one minute. Figure 5.2 shows two pairs of surgery videos that are incorrectly ranked by our model, from which we can see that our model could pay attention to the significant regions, e.g., the robot arms or the suture line. However, it is even difficult for a human to assess the skill from the sparsely sampled frames since the skill of manipulating the suture is revealed by continuous action. For example, a worse manipulation is represented by inserting the suture in a hole several times before succeeding, while only one time could be sampled. This means that although the skill-related attentions could be captured by our model, the ability of the model is still limited to the lost information caused by the sparse sampling of videos. We plan to incorporate the analysis of the variation of actions in continuous frames to our future work.

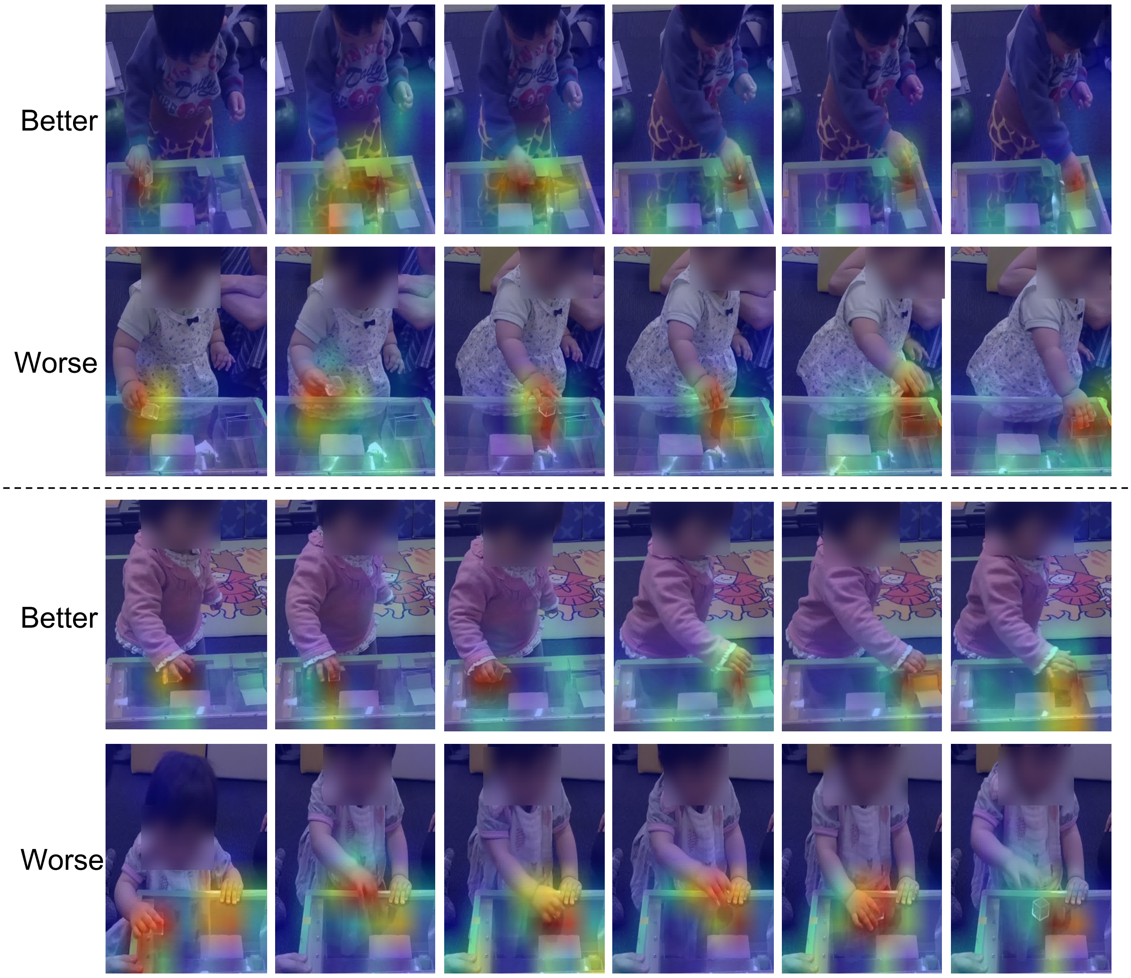


Figure 5.1: Visualization of attention maps with skill comparison on the Infant-Grasp Dataset. We select two correctly ranked pairs and show the main frames with attention maps predicted by our model. The ground-truth labels are listed at left. The estimated attention maps are always able to focus on the skill-related hand even though the actions and performance of infants are variant.

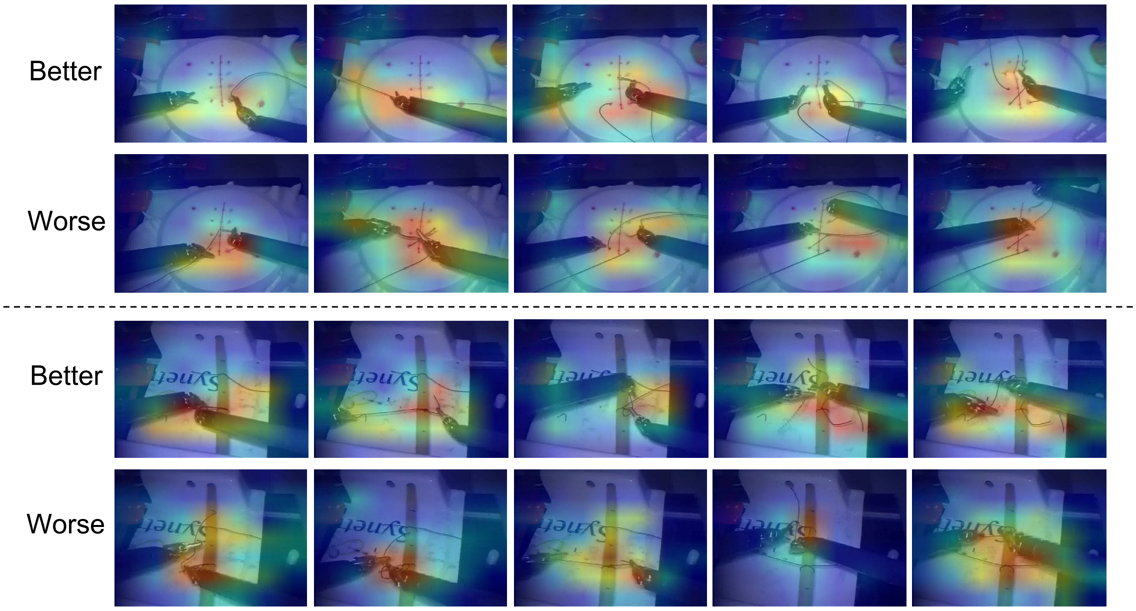


Figure 5.2: Visualization of attention maps with skill comparison on Surgery Dataset. We select two pairs that are ranked incorrectly by our model and show the main frames with attention maps predicted by our model. The ground-truth labels are listed at left. Although our model could capture the task-related significant regions, it is still difficult to distinguish which video performs better in a pair of videos.

5.3 Limitations

As analyzed in 5.2, the sparsely-sampling scheme utilized in our architecture restricts the model to capture the key actions in continuous frames. Although at every time step, the network takes the motion information in the form of stacked optical flow as input, it is still difficult to encode the temporal semantics of actions beyond the scope of a stacked optical flow. Then, to integrate information over time and capture the long-range temporal structure in each video, we adopt a framework based on a recurrent neural network (RNN). However, even though we choose an improved RNN structure, the gated recurrent unit (GRU), it still suffers from the problem of gradient vanishing, especially for long sequences. In our architecture, the processed videos are all no longer than two minutes and they are sampled into sequences with the length of 25 for inputting into GRU. It is essential to consider how to process videos with much longer length since many tasks are recorded in longer videos and the skills tend to be included in long-range temporal structures of videos. We list the limitations of our approach as below.

- The sparsely-sampling scheme cannot guarantee all the temporal semantics of the key actions in continuous frames could be captured, which limits the performance of our approach in analyzing delicate actions such as actions in surgery.
- The RNN structure restricts our approach to process videos with much longer length due to the problem of gradient vanishing.

6 Conclusion and Future work

In this thesis, we present a novel approach for skill level assessment in videos with spatial attention. Since assessing the skill levels from videos is challenging even for humans without expert knowledge in the corresponding domain, a system that could assess skill from videos automatically by learning from a large amount of training data will be helpful in many scenes such as health rehabilitation and surgery skill training. However, previous works cannot solve this task perfectly. The challenges lie on two main aspects: (1) how to extract effective representation for skill assessment from videos; (2) how to collect more training dataset.

To solve the first challenge, we proposed to leverage the mechanism of spatial attention, which plays an important role in human perception process. Our approach could dynamically estimate spatial attention, the regions most relevant to the performed task, by exploiting not only the low-level (appearance-motion) information in each frame but also high-level skill-related information. We show that through the spatial attention module, the redundant information is removed and effective and discriminative representation becomes achievable, which could capture more crucial skill-related information. For the second challenge, to alleviate the difficulty of annotation, we adopted a pair-wise ranking architecture to train our model, in which only pair-wise ranking labels are required as annotations rather than exact scores. Additionally, we also collected and annotated a new dataset which contains the largest number of videos performing the same action compared with existing datasets for skill assessment. Experiments demonstrated that our proposed approach achieved state-of-the-art performance on both our new dataset and existing datasets. The visualization of spatial attention maps also proved the effectiveness of our spatial attention approach. As a discussion, we showed more visualization results to reveal how the learned spatial attention helps discriminate different skill levels from videos and why our approach performs differently on different datasets.

We also analyzed the limitations of our method and will tackle those in our future work. Some limitations are caused by the limited size of existing datasets since our approach is based on deep learning modules which perform better when trained by a large amount of data empirically. Although the pair-wise ranking annotation and training architecture play roles in data augmentation, we think the number of video pairs is still insufficient. We believe a more sophisticated approach for skill assessment will benefit from a dataset with a larger number of videos.

Another limitation is that our approach is weak in processing long videos because the recurrent neural framework adopted in our model tends to suffer from the gradient vanishing problem. Additionally, the sparsely-sampling scheme also limits the performance of our approach in analyzing delicate actions such as suturing in surgery since the key actions incline to be contained in continuous frames. We will delve deeper into the way of extracting representation for videos, especially for long videos, e.g., exploiting temporal attention to remove redundant information in temporal and leveraging long-term feature banks to discover long-range temporal relationships.

Acknowledgments

I would first like to express my gratitude to my supervisor Prof. Yoichi Sato for the continuous support of my Master's study and research. He is patient and has immense knowledge in computer vision. He offered us plenty of free space for self-exploration but can always lead me to the right directions when I was lost. Two years ago, I was a freshman and standing at the door of research and now I finished my initial walk on the way of research. This progress would have been impossible without the aid and guidance of him.

Besides my supervisor, I am profoundly grateful to the rest of the paper's co-authors: Minjie Cai and Yifei Huang. They are not only my co-workers but also friends. The discussion with them always inspired me whenever I met a problem or came up with a new idea. Their excellent ability in research and composing paper also encourage me to keep improving myself and going forward.

My sincere thanks also go to the research group of Prof. Minagawa at Keio University, for providing a data source for our dataset.

I would like to thank all of my labmates in Sato Laboratory and all the friends I have met in the two years, for the stimulating discussions, and for all the fun we have had in the last two years. Without their company, I could not have a wonderful life in the past two years.

Finally, I must express my very profound gratitude to my parents and my old friend Yibo Ma, for providing me with unfailing support and continuous encouragement throughout my years of study and through the process of researching and writing this thesis. This accomplishment would not have been possible without them. Thank you.

References

[AHB+18] Peter Anderson, Xiaodong He, Chris Buehler, Damien Teney, Mark Johnson, Stephen Gould, and Lei Zhang. Bottom-up and top-down attention for image captioning and visual question answering. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2018.

[BSPYS17] Gedas Bertasius, Hyun Soo Park, Stella X. Yu, and Jianbo Shi. Am i a baller? basketball performance assessment from first-person videos. In *IEEE International Conference on Computer Vision (ICCV)*, 2017.

[BSR+05] Chris Burges, Tal Shaked, Erin Renshaw, Ari Lazier, Matt Deeds, Nicole Hamilton, and Greg Hullender. Learning to rank using gradient descent. In *Proceedings of the 22nd international conference on Machine learning*, pages 89–96. ACM, 2005.

[ÇAWS13] Oya Çeliktutan, Ceyhun Burak Akgul, Christian Wolf, and Bülent 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.

[CGCB14] Junyoung Chung, Caglar Gulcehre, KyungHyun Cho, and Yoshua Bengio. Empirical evaluation of gated recurrent neural networks on sequence modeling. *arXiv preprint arXiv:1412.3555*, 2014.

[CZ18] Shi Chen and Qi Zhao. Boosted attention: Leveraging human attention for image captioning. In *European Conference on Computer Vision (ECCV)*, 2018.

[CZM17] Lei Chen, Mengyao Zhai, and Greg Mori. Attending to distinctive moments: Weakly-supervised attention models for action localization in video. In *Computer Vision Workshop (ICCVW), 2017 IEEE International Conference on*, 2017.

[CZX+17] Long Chen, Hanwang Zhang, Jun Xiao, Liqiang Nie, Jian Shao, Wei Liu, and Tat-Seng Chua. Sca-cnn: Spatial and channel-wise attention in convolutional networks for image captioning. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2017.

[DAZ+17] Abhishek Das, Harsh Agrawal, Larry Zitnick, Devi Parikh, and Dhruv Batra. Human attention in visual question answering: Do humans and deep networks look at the same regions? *Computer Vision and Image Understanding*, 163:90–100, 2017.

[DDMC17] Hazel Doughty, Dima Damen, and Walterio Mayol-Cuevas. Who's better, who's best: Skill determination in video using deep ranking. *arXiv preprint arXiv:1703.09913*, 2017.

[DDS+09] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *IEEE Conference on Computer Vision and Pattern Recognition*, 2009.

[DlTHB+08] Fernando De la Torre, Jessica Hodgins, Adam Bargteil, Xavier Martin, Justin Macey, Alex Collado, and Pep Beltran. Guide to the carnegie mellon university multimodal activity (cmu-mmac) database. *Robotics Institute*, page 135, 2008.

[DWQ17] Wenbin Du, Yali Wang, and Yu Qiao. Rpan: An end-to-end recurrent pose-attention network for action recognition in videos. In *IEEE International Conference on Computer Vision*, 2017.

[GGZ+17] Lianli Gao, Zhao Guo, Hanwang Zhang, Xing Xu, and Heng Tao Shen. Video captioning with attention-based lstm and semantic consistency. *IEEE Transactions on Multimedia*, 19:2045–2055, 2017.

[GR17] Rohit Girdhar and Deva Ramanan. Attentional pooling for action recognition. In *Advances in Neural Information Processing Systems*, pages 34–45, 2017.

[GRG+17] Rohit Girdhar, Deva Ramanan, Abhinav Gupta, Josef Sivic, and Bryan Russell. Actionvlad: Learning spatio-temporal aggregation for action classification. In *IEEE conference on computer vision and pattern recognition*, 2017.

[GVR+14] Yixin Gao, S Swaroop Vedula, Carol E Reiley, Narges Ahmidi, Balakrishnan Varadarajan, Henry C Lin, Lingling Tao, Luca Zappella, Benjamın Béjar, David D Yuh, et al. Jhu-isi gesture and skill assessment working set (jigsaws): A surgical activity dataset for human motion modeling. In *MICCAI Workshop: M2CAI*, volume 3, page 3, 2014.

[HCK+17] Yifei Huang, Minjie Cai, Hiroshi Kera, Ryo Yonetani, Keita Higuchi, and Yoichi Sato. Temporal localization and spatial segmentation of joint attention in multiple first-person videos. In *Computer Vision Workshop (ICCVW), 2017 IEEE International Conference on*, 2017.

[HCLS18] Yifei Huang, Minjie Cai, Zhenqiang Li, and Yoichi Sato. Predicting gaze in egocentric video by learning task-dependent attention transition. *arXiv preprint arXiv:1803.09125*, 2018.

[HCLS19] Yifei Huang, Minjie Cai, Zhenqiang Li, and Yoichi Sato. Mutual context network for jointly estimating egocentric gaze and actions. *arXiv preprint arXiv:1901.01874*, 2019.

[HS97] Sepp Hochreiter and Jürgen Schmidhuber. Long short-term memory. *Neural computation*, 9(8):1735–1780, 1997.

[HSS17] Jie Hu, Li Shen, and Gang Sun. Squeeze-and-excitation networks. *arXiv preprint arXiv:1709.01507*, 7, 2017.

[HZRS16] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *IEEE conference on computer vision and pattern recognition*, 2016.

[IMG03] Winfried Ilg, Johannes Mezger, and Martin Giese. Estimation of skill levels in sports based on hierarchical spatio-temporal correspondences. In *Joint Pattern Recognition Symposium*, pages 523–531. Springer, 2003.

[Joa02] Thorsten Joachims. Optimizing search engines using clickthrough data. In *Proceedings of the eighth ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 133–142. ACM, 2002.

[JPDK03] Marko Jug, Janez Perš, Branko Dežman, and Stanislav Kovačič. Trajectory based assessment of coordinated human activity. In *International Conference on Computer Vision Systems*, pages 534–543. Springer, 2003.

[LGG+18] Zhenyang Li, Kirill Gavrilyuk, Efstratios Gavves, Mihir Jain, and Cees GM Snoek. Videolstm convolves, attends and flows for action recognition. *Computer Vision and Image Understanding*, 166:41–50, 2018.

[LWH+17] Jun Liu, Gang Wang, Ping Hu, Ling-Yu Duan, and Alex C Kot. Global context-aware attention lstm networks for 3d action recognition. In *IEEE Conference on Computer Vision and Pattern Recognition*, 2017.

[LYBP16] Jiasen Lu, Jianwei Yang, Dhruv Batra, and Devi Parikh. Hierarchical question-image co-attention for visual question answering. In *Advances In Neural Information Processing Systems*, pages 289–297, 2016.

[LYD+18] Dong Li, Ting Yao, Lingyu Duan, Tao Mei, and Yong Rui. Unified spatio-temporal attention networks for action recognition in videos. *IEEE Transactions on Multimedia*, 2018.

[MHG+14] Volodymyr Mnih, Nicolas Heess, Alex Graves, et al. Recurrent models of visual attention. In *Advances in neural information processing systems*, pages 2204–2212, 2014.

[MLS17] Antoine Miech, Ivan Laptev, and Josef Sivic. Learnable pooling with context gating for video classification. *arXiv preprint arXiv:1706.06905*, 2017.

[MVCH14] Anand Malpani, S Swaroop Vedula, Chi Chiung Grace Chen, and Gregory 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.

[PG11] Devi Parikh and Kristen Grauman. Relative attributes. In *IEEE International Conference on Computer Vision (ICCV)*, 2011.

[PGC⁺17] Adam Paszke, Sam Gross, Soumith Chintala, Gregory Chanan, Edward Yang, Zachary DeVito, Zeming Lin, Alban Desmaison, Luca Antiga, and Adam Lerer. Automatic differentiation in pytorch. 2017.

[PM17] Paritosh Parmar and Brendan Tran Morris. Learning to score olympic events. In *Computer Vision and Pattern Recognition Workshops (CVPRW), 2017 IEEE Conference on*. IEEE, 2017.

[PVT14] Hamed Pirsiavash, Carl Vondrick, and Antonio Torralba. Assessing the quality of actions. In *European Conference on Computer Vision*, 2014.

[SBP⁺14] Yachna Sharma, Vinay Bettadapura, Thomas Plötz, Nils Hammerla, Sebastian Mellor, Roisin McNaney, Patrick Olivier, Sandeep Deshmukh, Andrew McCaskie, and Irfan Essa. Video based assessment of osats using sequential motion textures. Georgia Institute of Technology, 2014.

[SCD⁺17] Ramprasaath R Selvaraju, Michael Cogswell, Abhishek Das, Ramakrishna Vedantam, Devi Parikh, Dhruv Batra, et al. Grad-cam: Visual explanations from deep networks via gradient-based localization. In *ICCV*, pages 618–626, 2017.

[SLX⁺17] Sijie Song, Cuiling Lan, Junliang Xing, Wenjun Zeng, and Jiaying Liu. An end-to-end spatio-temporal attention model for human action recognition from skeleton data. In *AAAI*, 2017.

[SZ14] Karen Simonyan and Andrew Zisserman. Two-stream convolutional networks for action recognition in videos. In *Advances in neural information processing systems*, pages 568–576, 2014.

[SZS12] Khurram Soomro, Amir Roshan Zamir, and Mubarak Shah. Ucf101: A dataset of 101 human actions classes from videos in the wild. *arXiv preprint arXiv:1212.0402*, 2012.

[WJQ⁺17] Fei Wang, Mengqing Jiang, Chen Qian, Shuo Yang, Cheng Li, Honggang Zhang, Xiaogang Wang, and Xiaoou Tang. Residual attention network for image classification. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2017.

[WPLSK18] Sanghyun Woo, Jongchan Park, Joon-Young Lee, and In So Kweon.

Cbam: Convolutional block attention module. In *European Conference on Computer Vision (ECCV)*, 2018.

[WXW$^+$16] Limin Wang, Yuanjun Xiong, Zhe Wang, Yu Qiao, Dahua Lin, Xiaoou Tang, and Luc Van Gool. Temporal segment networks: Towards good practices for deep action recognition. In *European Conference on Computer Vision*, 2016.

[YMR16] Ting Yao, Tao Mei, and Yong Rui. Highlight detection with pairwise deep ranking for first-person video summarization. In *IEEE Conference on Computer Vision and Pattern Recognition*, 2016.

[YSLZ18] Shiyang Yan, Jeremy S Smith, Wenjin Lu, and Bailing Zhang. Hierarchical multi-scale attention networks for action recognition. *Signal Processing: Image Communication*, 61:73–84, 2018.

[ZL11] Qiang Zhang and Baoxin 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.

[ZL15] Qiang Zhang and Baoxin 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.

[ZPB07] Christopher Zach, Thomas Pock, and Horst Bischof. A duality based approach for realtime tv-l 1 optical flow. In *Joint Pattern Recognition Symposium*, pages 214–223. Springer, 2007.

[ZSB$^+$15] Aneeq Zia, Yachna Sharma, Vinay Bettadapura, Eric L Sarin, Mark A Clements, and Irfan 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.

List of Publications

1. Zhenqiang Li, Yifei Huang, Minjie Cai, and Yoichi Sato, “Manipulation-skill Assessment from Videos with Spatial Attention Network, ” Submitted to IEEE Conference on Computer Vision and Pattern Recognition(CVPR), 2019
2. Zhenqiang Li, Yifei Huang, Minjie Cai, and Yoichi Sato, “Pairwise performance assessment from videos using appearance and body pose,” In Extended Abstract of Meeting on Image Recognition and Understanding (MIRU), Aug. 2018
3. Yifei Huang, Minjie Cai, Zhenqiang Li, and Yoichi Sato, “Predicting Gaze in Egocentric Video by Learning Task-dependent Attention Transition,” In Proceedings of the European Conference on Computer Vision(ECCV), 2018