Assessing the Quality of Actions

Hamed Pirsiavash, Carl Vondrick, and Antonio Torralba

Massachusetts Institute of Technology
{hpirsiav,vondrick,torralba}@mit.edu

Abstract. While recent advances in computer vision have provided reliable methods to recognize actions in both images and videos, the problem of assessing how well people perform actions has been largely unexplored in computer vision. Since methods for assessing action quality have many real-world applications in healthcare, sports, and video retrieval, we believe the computer vision community should begin to tackle this challenging problem. To spur progress, we introduce a learning-based framework that takes steps towards assessing how well people perform actions in videos. Our approach works by training a regression model from spatiotemporal pose features to scores obtained from expert judges. Moreover, our approach can provide interpretable feedback on how people can improve their action. We evaluate our method on a new Olympic sports dataset, and our experiments suggest our framework is able to rank the athletes more accurately than a non-expert human. While promising, our method is still a long way to rivaling the performance of expert judges, indicating that there is significant opportunity in computer vision research to improve on this difficult yet important task.

1 Introduction

Recent advances in computer vision have provided reliable methods for recognizing actions in videos and images. However, the problem of automatically quantifying *how well* people perform actions has been largely unexplored.

We believe the computer vision community should begin to tackle the challenging problem of assessing the quality of people's actions because there are many important, real-world applications. For example, in health care, patients are often monitored and evaluated after hospitalization as they perform daily tasks, which is expensive undertaking without an automatic assessment method.

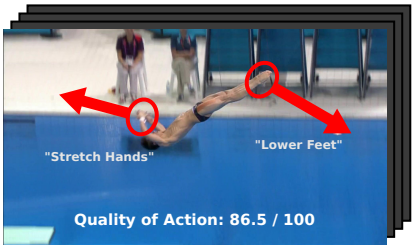


Fig. 1. We introduce a learning framework for assessing the quality of human actions from videos. Since we estimate a model for what constitutes a high quality action, our method can also provide feedback on how people can improve their actions, visualized with the red arrows.

D. Fleet et al. (Eds.): ECCV 2014, Part VI, LNCS 8694, pp. 556–571, 2014.

In sports, action quality assessments would allow an athlete to practice in front of a camera and receive quality scores in real-time, providing the athlete with rapid feedback and an opportunity to improve their action. In retrieval, a video search engine may want to sort results based on the quality of the action performed instead of only the relevance.

However, automatically assessing the quality of actions is not an easy computer vision problem. Human experts for a particular domain, such as coaches or doctors, have typically been trained over many years to develop complex underlying rules to assess action quality. If machines are to assess action quality, then they must discover similar rules as well.

In this paper, we propose a data-driven method to *learn* how to assess the quality of actions in videos. To our knowledge, we are the first to propose a general framework for learning to assess the quality of human-based actions from videos. Our method works by extracting the spatio-temporal pose features of people, and with minimal annotation, estimating a regression model that predicts the scores of actions. Fig.1 shows an example output of our system.

In order to quantify the performance of our methods, we introduce a new dataset for action quality assessment comprised of Olympic sports footage. Although the methods in this paper are general, sports broadcast footage has the advantage that it is freely available, and comes already rigorously "annotated" by the Olympic judges. We evaluate our quality assessments on both diving and figure skating competitions. Our results are promising, and suggest that our method is significantly better at ranking people's actions by their quality than non-expert humans. However, our method is still a long way from rivaling the performance of expert judges, indicating that there is significant opportunity in computer vision research to improve on this difficult yet important task.

Moreover, since our method leverages high level pose features to learn a model for action quality, we can use this model to help machines understand people in videos as well. Firstly, we can provide interpretable feedback to performers on how to improve the quality of their action. The red vectors in Fig.1 are output from our system that instructs the Olympic diver to stretch his hands and lower his feet. Our feedback system works by calculating the gradient for each body joint against the learned model that would have maximized people's scores. Secondly, we can create highlights of videos by finding which segments contributed the most to the action quality, complementing work in video summarization. We hypothesize that further progress in building better quality assessment models can improve both feedback systems and video highlights.

The three principal contributions of this paper revolve around automatically assessing the quality of people's actions in videos. Firstly, we introduce a general learning-based framework for the quality assessment of human actions using spatiotemporal pose features. Secondly, we then describe a system to generate feedback for performers in order to improve their score. Finally, we release a new dataset for action quality assessment in the hopes of facilitating future research on this task. The remainder of this paper describes these contributions in detail.

2 Related Work

This paper builds upon several areas of computer vision. We briefly review related work:

Action Assessment: The problem of action quality assessment has been relatively unexplored in the computer vision community. There have been a few promising efforts to judge how well people perform actions [1–3], however, these previous works have so far been hand-crafted for specific actions. The motivation for assessing peoples actions in healthcare applications has also been discussed before [4], but the technical method is limited to recognizing actions. In this paper, we propose a generic learning-based framework with state-of-the-art features for action quality assessment that can be applied to most types of human actions. To demonstrate this generality, we evaluate on two distinct types of actions (diving and figure skating). Furthermore, our system is able to generate interpretable feedback on how performers can improve their action.

Photograph Assessment: There are several works that assess photographs, such as their quality [5], interestingness [6] and aesthetics [7, 8]. In this work, we instead focus on assessing the quality of human actions, and not the quality of the video capture or its artistic aspects.

Action Recognition: There is a large body of work studying how to recognize actions in both images [9–13] and videos [14–18], and we refer readers to excellent surveys [19, 20] for a full review. While this paper also studies actions, we are interested in *assessing* their quality rather than *recognizing* them.

Features: There are many features for action recognition using spatiotemporal bag-of-words [21, 22], interest points [23], feature learning [24], and human pose based [25]. However, so far these features have primarily been shown to work for recognition. We found that some of these features, notably [24] and [25] with minor adjustments, can be used for the quality assessment of actions too.

Video Summarization: This paper complements work in video summarization [26–31]. Rather than relying on saliency features or priors, we instead can summarize videos by discarding segments that did not impact the quality score of an action, thereby creating a "highlights reel" for the video.

3 Assessing Action Quality

We now present our system for assessing the quality of an action from videos. On a high level, our model learns a regression model from spatio-temporal features. After presenting our model, we then show how our model can be used to provide feedback to the people in videos to improve their actions. We finally describe how our model can highlight segments of the video that contribute the most to the quality score.

3.1 Features

To learn a regression model to the action quality, we extract spatio-temporal features from videos. We consider two sets of features: low-level features that capture gradients and velocities directly from pixels, and high-level features based off the trajectory of human pose.

Low Level Features: Since there has been significant progress in developing features for *recognizing* actions, we tried using them for *assessing* actions too. We use a hierarchical feature [24] that obtains state-of-the-art performance in action recognition by learning a filter bank with independent subspace analysis. The learned filter bank consists of spatio-temporal Gabor-like filters that capture edges and velocities. In our experiments, we use the implementation by [24] with the network pre-trained on the Hollywood2 dataset [32].

High Level Pose Features: Since most low-level features capture statistics from pixels directly, they are often difficult to interpret. As we wish to provide feedback on how a performer can improve their actions, we want the feedback to be interpretable. Inspired by actionlets [25], we now present high level features based off human pose that are interpretable.

Given a video, we assume that we know the pose of the human performer in every frame, obtained either through ground truth or automatic pose estimation. Let $p^{(j)}(t)$ be the x component of the jth joint in the tth frame of the video. Since we want our features to be translation-invariant, we normalize the joint positions relative to the head position:

$$q^{(j)}(t) = p^{(j)}(t) - p^{(0)}(t)$$

where we have assumed that $p^{(0)}(t)$ refers to the head. Note that $q^{(j)}$ is a function of time, so we can represent it in the frequency domain by the discrete cosine transform (DCT): $Q^{(j)} = Aq^{(j)}$ where A is the discrete cosine transformation matrix. We then use the k lowest frequency components to create the feature vector $\phi_j = \left|Q^{(j)}_{1:k}\right|$ where $A_{1:k}$ selects the first k rows of A. We found that only using the low frequencies helps remove high frequency noise due to pose estimation errors. We use the absolute value of the frequency coefficients Q_i.

We compute ϕ_j for every joint for both the x- and y-components, and concatenate them to create the final feature vector ϕ. We note that if the video is long, we break it up into segments and concatenate the features to produce one feature vector for the entire video. This inreases the temporal resolution of our features for long videos.

Actionlets [25] uses a similar method with Discrete Fourier Transform (DFT) instead. Although there is a close relationship between DFT and DCT, we see better results using DCT. We believe this is the case since DCT provides a more compact representation. Additionally, DCT coefficients are real numbers instead of complex, so less information is lost in the absolute value operation.

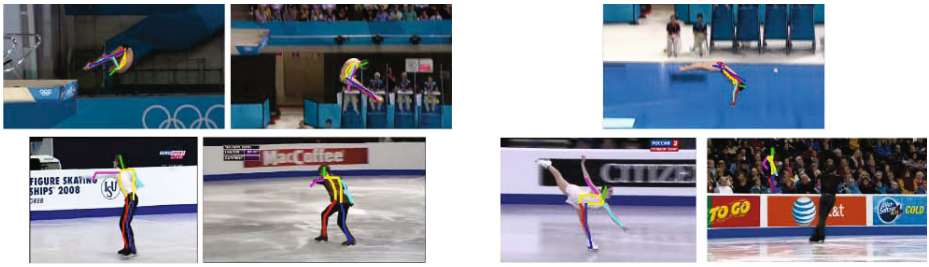


Fig. 2. Pose Estimation Challenges: Some results for human pose estimation on our action quality dataset. Since the performers contort their body in unusual configurations, pose estimation is very challenging on our dataset.

In order to estimate the joints of the performer throughout the video $p^{(j)}(t)$, we run a pose estimation algorithm to find the position of the joints in every frame. We estimate the pose using a Flexible Parts Model [33] for each frame independently. Since [33] finds the best pose for a single frame using dynamic programming and we want the best pose across the entire video, we find the N-best pose solutions per frame using [34]. Then we associate the poses using a dynamic programming algorithm to find the best track in the whole video. The association looks for the single best smooth track covering the whole temporal span of the video. Fig.2 shows some successes and failures of this pose estimation.

3.2 Learning

We then pose quality assessment as a supervised regression problem. Let $\Phi_i \in \mathbb{R}^{k \times n}$ be the pose features for video i in matrix form where n is the number of joints and k is the number of low frequency components. We write $y_i \in \mathbb{R}$ to denote the ground-truth quality score of the action in video i, obtained by an expert human judge. We then train a linear support vector regression (L-SVR) [35] to predict y_i given features Φ_i over a training set. In our experiments, we use libsvm [36]. Optimization is fast, and takes less than a second on typical sized problems. We perform cross validation to estimate hyperparameters.

Domain Knowledge: We note that a comprehensive model for quality assessment might use domain experts to annotate fine-tuned knowledge on the action's quality (e.g., "the leg must be straight"). However, relying on domain experts is expensive and difficult to scale to a large number of actions. By posing quality assessment as a machine learning problem with minimal interaction from an expert, we can scale more efficiently. In our system, we only require a single real number per video corresponding to the score of the quality.

Prototypical Example: Moreover, a fairly simple method to assess quality is to check the observed video against a ground truth video with perfect execution, and then determine the difference. However, in practice, many actions can have

multiple ideal executions (e.g., a perfect overhand serve might be just as good as a perfect underhand serve). Instead, our model can handle multi-modal score distributions.

3.3 Feedback Proposals

As a performer executes an action, in addition to assessing the quality, we also wish to provide feedback on how the performer can improve his action. Since our regression model operates over pose-based features, we can determine how the performer should move to maximize the score.

We accomplish this by differentiating the scoring function with respect to joint location. We calculate the gradient of the score with respect to the location of each joint $\frac{\partial S}{\partial p^{(j)}(t)}$ where S is the scoring function. By calculating the maximum gradient, we can find the joint and the direction that the performer must move to achieve the largest improvement in the score.

We are able to analytically calculate the gradient. Recall that L-SVR learns a weight vector $W \in \mathbb{R}^{k \times n}$ such that W predicts the score of the action quality by the dot-product:

$$S = \sum_{f=1}^{k} \sum_{j=1}^{n} W_{fj} \Phi_{fj}$$

where Φ_{fj} is the fth frequency componenet for the jth joint. After basic algebra, we can compute the gradient of the score S with respect to the location of each joint $p^{(j)}(t)$:

$$\frac{\partial S}{\partial p^{(j)}(t)} = \sum_{f=1}^{k} A_{ft} W_{fj} \cdot \text{sign} \left(\sum_{t'=1}^{T} \left(A_{ft'} (p^{(j)}(t') - p^{(0)}(t')) \right) \right)$$

By computing $\max_{p^{(j)}(t)} \frac{\partial S}{\partial p^{(j)}(t)}$, we can find the joint and the direction the performer must move to most improve the score.[1]

3.4 Video Highlights

In addition to finding the joint that will result in the largest score improvement, we also wish to measure the *impact* a segment of the video has on the quality score. Such a measure could be useful in summarizing the segments of actions that contribute to high or low scores.

We define a segment's impact as how much the quality score would change if the segment were removed. In order to remove a segment, we compute the most likely feature vector had we not observed the missing segment. The key observation is that since we only use the low frequency components in our feature vector, there are more equations than unknowns when estimating the DCT coefficients. Consequently, removing a segment corresponds to simply removing some equations.

[1] We do not differentiate with respect to the head location because it is used for normalization.

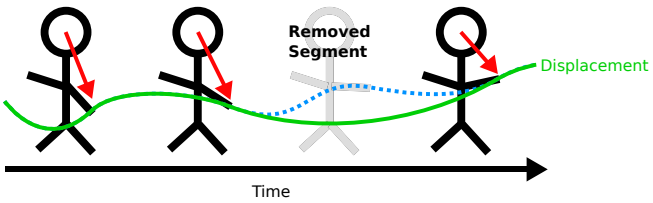


Fig. 3. Interpolating Segments: This schematic shows how the displacement vector changes when a segment of the video is removed in order to compute impact. The dashed curve is the original displacement, and the solid curve is the most likely displacement given observations with a missing segment.

Let $B = A^+$ be the inverse cosine transform where A^+ is the psuedo-inverse of A. Then, the DCT equation can be written as $Q^{(j)} = B^+ q^{(j)}$. If the data from frames u through v is missing, then the inferred DCT coefficients are $\hat{Q}^{(j)} = (B_{\overline{u:v}})^+ q^{(j)}$ where $B_{\overline{u:v}}$ is the sub-matrix of B that excludes rows u through v. The frequency components $\hat{Q}^{(j)}$ are the same dimensionality as $Q^{(j)}$, but they have inferred the missing segment with the most likely joint trajectory. Fig.3 visualizes how the features change with this transformation.

We use $\hat{Q}^{(j)}$ to create the feature vector for the video with the missing segment. Finally, we determine the impact of the missing segment by calculating the difference in scores between the original feature vector and the feature vector with the missing segment.

4 Experiments

In this section, we evaluate both our quality assessment method and feedback system for quality improvement with quantitative experiments. Since quality assessment has not yet been extensively studied in the computer vision community, we first introduce a new video dataset for action quality assessment.

4.1 Action Quality Dataset

There are two primary hurdles in building a large dataset for action quality assessment. Firstly, the score annotations are subjective, and require an expert. Unfortunately, hiring an expert to annotate hundreds of videos is expensive. Secondly, in some applications such as health care, there are privacy and legal issues involved in collecting videos from patients. In order to establish a baseline dataset for further research, we desire freely available videos.

We introduce an Olympics video dataset for action quality assessment. Sports footage has the advantage that it can be obtained freely, and the expert judge's scores are frequently released publicly. We collected videos from YouTube for two categories of sports, diving and figure skating, from recent Olympics and other worldwide championships. The videos are long with multiple instances of actions performed by multiple people. We annotated the videos with the start

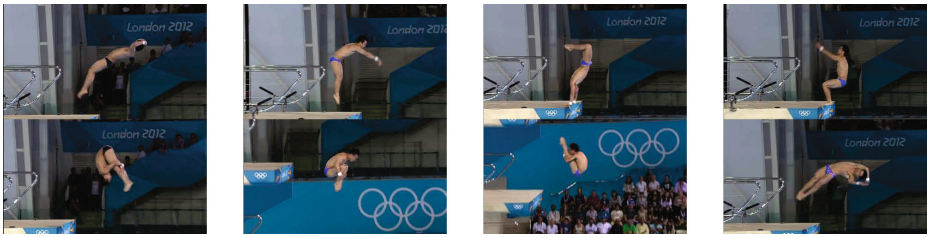


Fig. 4. Diving Dataset: Some of the best dives from our diving dataset. Each column corresponds to one video. There is a large variation in the top-scoring actions. Hence, providing feedback is not as easy as pushing the action towards a canonical "good" performance.

Fig. 5. Figure Skating Dataset: Sample frames from our figure skating dataset. Notice the large variations of routines that the performers attempt. This makes automatic pose estimation challenging.

and end frame for each instance, and we extracted the judge's score. The dataset will be publicly available.

Diving: Fig.4 shows a few examples of our diving dataset. Our diving dataset consists of 159 videos. The videos are slow-motion from television broadcasting channels, so the effective frame rate is 60 frames per second. Each video is about 150 frames, and the entire dataset consists of 25,000 frames. The ground truth judge scores varies between 20 (worst) and 100 (best). In our experiments, we use 100 instances for training and the rest for testing. We repeated every experiment 200 times with different random splits and averaged the results. In addition to the Olympic judge's score, we also consulted with the MIT varsity diving coach who annotated which joints a diver should adjust to improve each dive. We use this data to evaluate our feedback system for the quality improvement algorithm.

Figure Skating: Fig.5 shows some frames from our figure skating dataset. This dataset contains 150 videos captured at 24 frames per second. Each video is almost 4,200 frames, and the entire dataset is 630,000 frames. The judge's score ranges between 0 (worst) and 100 (best). We use 100 instances for training and the rest for testing. As before, we repeated every experiment 200 times with different random splits and averaged the results. We note that our figure skating

Table 1. Diving Evaluation: We show mean rank correlation on our diving dataset. Higher is better. The pose-based features provide the best performance.

Method	STIP	Hierarchical	Pose+DFT	Pose+DCT
SVR	0.07	0.19	0.27	**0.41**
Ridge Reg	0.10	0.16	0.19	0.27

Table 2. Figure Skating Evaluation: We calculate mean rank correlation on our figure skating dataset. Higher is better. The hierarchical network features provide the best results. Although pose based features are not superior, they still enable high level analysis by providing feedback for quality improvement. We believe pose based features can benefit from using a better pose estimation.

Method	STIP	Hierarchical	Pose+DFT	Pose+DCT
SVR	0.21	**0.45**	0.31	0.35
Ridge Reg	0.20	**0.44**	0.19	0.25

tends to be more challenging for pose estimation since it is at a lower frame rate, and has more variation in the human pose and clothing (e.g., wearing skirt).

4.2 Quality Assessment

We evaluate our quality assessment on both the figure skating and diving dataset. In order to compare our results against the ground truth, we use the rank correlation of the scores we predict against the scores the Olympic judges awarded. Tab.1 and Tab.2 show the mean performance over random train/test splits of our datasets. Our results suggest that pose-based features are competitive, and even obtain the best performance on the diving dataset. In addition, our results indicate that features learned to recognize actions can be used to assess the quality of actions too. We show some of the best and worst videos as predicted by our model in Fig.6.

We compare our quality assessment against several baselines. Firstly, we compare to both space-time interest points (STIP) and pose-based features with Discrete Fourier Transform (DFT) instead of DCT (similar to [24]). Both of these features performed worse. Secondly, we also compare to ridge regression with all feature sets. Our results show that support vector regression often obtains significantly better performance.

We also asked non-expert human annotators to predict the quality of each diver in the diving dataset. Interestingly, after we instructed the subjects to read the Wikipedia page on diving, non-expert annotators were only able to achieve a rank correlation of 19%, which is half the performance of support vector regression with pose features. We believe this difference is evidence that our algorithm is starting to learn which human poses constitute good dives. We note, however, that our method is far from matching Olympic judges since they

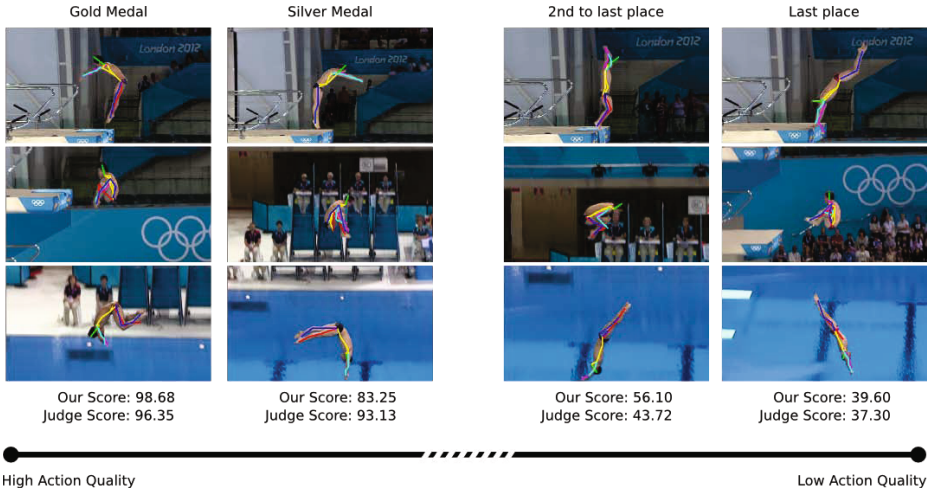


Fig. 6. Examples of Diving Scores: We show the two best and worst videos sorted by the predicted score. Each column is one video with ground truth and predicted score written below. Notice that in the last place video, the diver lacked straight legs in the beginning and did not have a tight folding pose. These two pitfalls are part of common diving advice given by coaches, and our model has learned this independently.

are able to predict the median judge's score with a rank correlation of 96%, suggesting that there is still significant room for improvement.[2]

4.3 Limitations

While our system is able to predict the quality of actions with some success, it has many limitations. One of the major bottlenecks is the pose estimation. Fig.2 shows a few examples of the successes and failures of the pose estimation. Pose estimation in our datasets is very challenging since the performers contort their body in many unusual configurations with significant variation in appearance. The frequent occlusion by clothing for figure skating noticeably harms the pose estimation performance. When the pose estimation is poor, the quality score is strongly affected, suggesting that advances in pose estimation or using depth sensors for pose can improve our system. Future work in action quality can be made robust against these types of failures as well by accounting for the uncertainty in the pose estimation.

[2] Olympic diving competitions have two scores: the technical difficulty and the score. The final quality of the action is then the product of these two quantities. Judges are told the technical difficulty apriori, which gives them a slight competitive edge over our algorithms. We did not model the technical difficulty in the interest of building a general system.

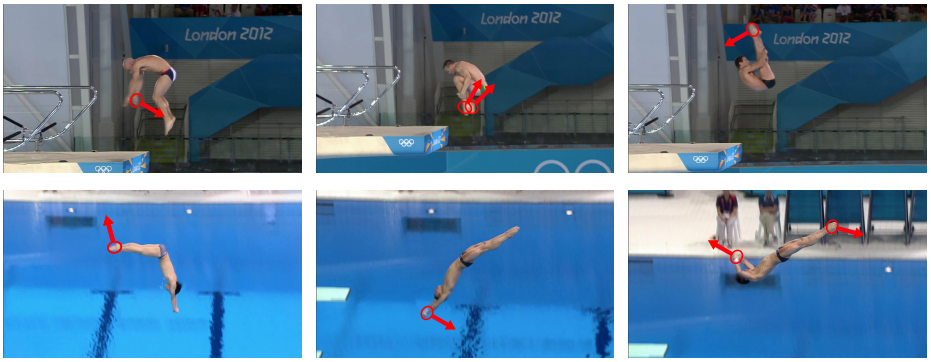


Fig. 7. Diving Feedback Proposals: We show feedback for some of the divers. The red vectors are instructing the divers to move their body in the direction of the arrow. In general, the feedback instructs divers to tuck their legs more and straighten their body before entering the pool.

Our system is designed to work for one human performer only, and does not model coordination between multiple people, which is often important for many types of sports and activities. We believe that future work in explicitly modeling team activities and interactions can significantly advance action quality assessment. Moreover, we do not model objects used during actions (such as sports balls or tools), and we do not consider physical outcomes (such as splashes in diving), which may be important features for some activities. Finally, while our representation captures the movements of human joint locations, we do not explicitly model their synchronization (e.g., keeping legs together) or repetitions (e.g., waving hands back and forth). We suspect a stronger quality assessment model will factor in these visual elements.

4.4 Feedback for Improvement

In addition to quality assessment, we evaluate the feedback vectors that our method provides. Fig.7 and Fig.8 show qualitatively a sample of the feedback that our algorithm suggests. In general, the feedback is reasonable, often making modifications to the extremities of the performer.

In order to quantitatively evaluate our feedback method, we needed to acquire ground truth annotations. We consulted with the MIT diving team coach who watched a subset of the videos in our dataset (27 in total) and provided suggestions on how to improve the dive. The diving coach gave us specific feedback (such as "move left foot down") as well as high-level feedback (e.g., "legs should be straight here" or "tuck arms more"). We translated each feedback from the coach into one of three classes, referring to whether the diver should adjust his upper body, his lower body, or maintain the same pose on each frame. Due to the subjective nature of the task, the diving coach was not able to provide more

Fig. 8. Figure Skating Feedback Proposals: We show feedback for some of the figure skaters where the red vectors are instructions for the figure skaters.

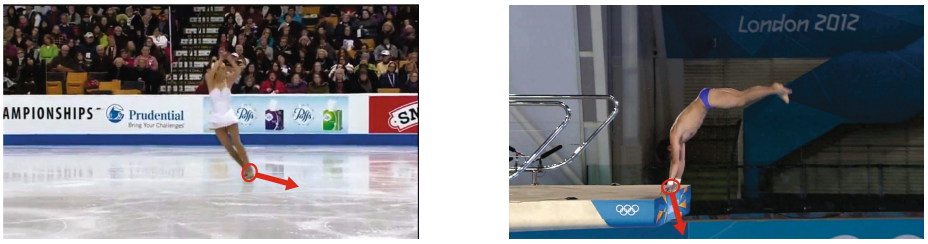


Fig. 9. Feedback Limitations: The feedback we generate is not perfect. If the figure skater or diver were to rely completely on the feedback above, they may fall over. Our model does not factor in physical laws, motivating work in support inference [37, 38].

detailed feedback annotations. Hence, the feedback is coarsely mapped into these three classes.

We then evaluate our feedback as a detection problem. We consider a feedback proposal from our algorithm as correct if it suggests to move a body part within a one second range of the coach making the same suggestion. We use the magnitude of the feedback gradient as the importance of the feedback proposal. We use a leave-one-out approach where we predict feedback on a video heldout from training. Our feedback proposals obtain 53.18% AP overall for diving, compared to 27% AP chance level. We compute chance by randomly generating feedback that uniformly chooses between the upper body and lower body.

Since our action quality assessment model is not aware of physical laws, the feedback suggestions can be physically implausible. Fig.9 shows a few cases where if the performer listened to our feedback, they might fall over. Our method's lack of physical models motivates work in support inference [37, 38].

Interestingly, by averaging the feedback across all divers in our dataset, we can find the most common feedback produced by our model. Fig.10 shows the magnitude of feedback for each frame and each joint averaged over all divers. For visualization proposes, we warp all videos to have the same length. Most of the feedback suggests correcting the feet and hands, and the most important frames

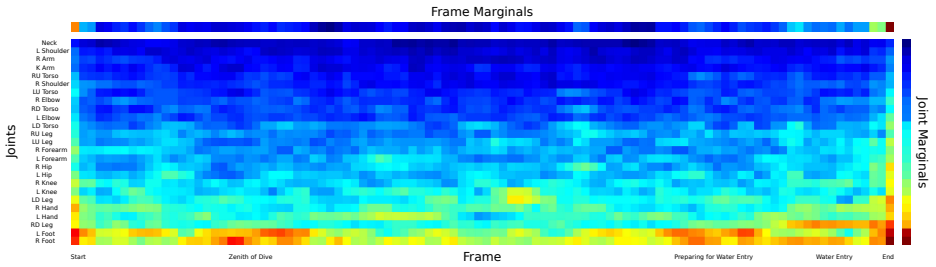


Fig. 10. Visualizing Common Feedback: We visualize the average feedback magnitude across the entire diving dataset for each joint and frame. Red means high feedback and blue means low feedback. The top and right edges show marginals over frames and joints respectively. R and L stand for right and left respectively, and U and D stand for upper and lower body, respectively. Feet are the most common area for feedback on Olympic divers, and that the beginning and end of the dive are the most important time points.

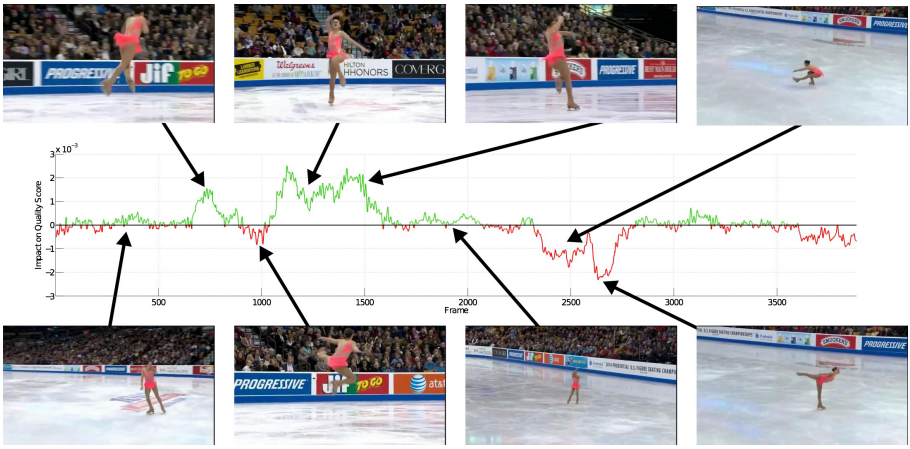
turn out to be the initial jump off the diving board, the zenith of the dive, and the moment right before the diver enters the water.

4.5 Highlighting Impact

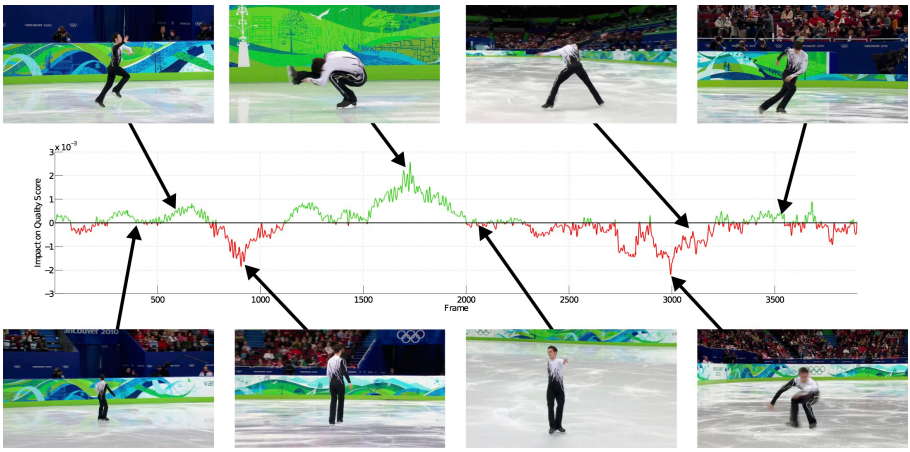
We qualitatively analyze the video highlights produced by finding the segments that contributed the most to the final quality score. We believe that this measure can be useful for video summarization since it reveals, out of a long video, which clips are the most important for the action quality. We computed impact on a routine from the figure skating dataset in Fig.11. Notice when the impact is near zero, the figure skater is in a standard, up-right position, or in-between maneuvers. The points of maximum impact correspond to jumps and twists of the figure skater, which contributes positively to the score if the skater performs it correctly, and negatively otherwise.

4.6 Discussion

If quality assessment is a subjective task, is it reasonable for a machine to still obtain reasonable results? Remarkably, the independent Olympic judges agree with each other 96% of the time, which suggests that there is some underlying structure in the data. One hypothesis to explain this correlation is that the judges are following a complex system of rules to gauge the score. If so, then the job of a machine quality assessment system is to extract these rules. While the approach in this paper attempts to learn these rules, we are still a long way from high performance on this task.



(a)



(b)

Fig. 11. Video Highlights: By calculating the impact each frame has on the score of the video, we can summarize long videos with the segments that have the largest impact on the quality score. Notice how, above, when the impact is close to zero, the skater is usually in an upright standard position, and when the impact is large, the skater is performing a maneuver.

5 Conclusions

Assessing the quality of actions is an important problem with many real-world applications in health care, sports and search. To enable these applications, we have introduced a general learning-based framework to automatically assess an action's quality from videos as well as to provide feedback for how the performer can improve. We evaluated our system on a dataset of Olympic divers and figure skaters, and we show that our approach is significantly better at assessing an

action's quality than a non-expert human. Although the quality of an action is a subjective measure, the independent Olympic judges have a large correlation. This implies that there is a well defined underlying rule that a computer vision system should be able to learn from data. Our hope is that this paper will motivate more work in this relatively unexplored area.

Acknowledgments. We thank Zoya Bylinkskii and Sudeep Pillai for comments and the MIT diving team for their helpful feedback. Funding was provided by a NSF GRFP to CV and a Google research award and ONR MURI N000141010933 to AT.

References

1. Gordon, A.S.: Automated video assessment of human performance. In: AI-ED. (1995)
2. Jug, M., Perš, J., Dežman, B., Kovačič, S.: Trajectory based assessment of coordinated human activity. Springer (2003)
3. Perše, M., Kristan, M., Perš, J., Kovacic, S.: Automatic Evaluation of Organized Basketball Activity using Bayesian Networks. Computer Vision Winter Workshop (2007)
4. Pirsiavash, H., Ramanan, D.: Detecting activities of daily living in first-person camera views. In: CVPR. (2012)
5. Ke, Y., Tang, X., Jing, F.: The design of high-level features for photo quality assessment. In: CVPR. (2006)
6. Gygli, M., Grabner, H., Riemenschneider, H., Nater, F., Van Gool, L.: The interestingness of images. (2013)
7. Datta, R., Joshi, D., Li, J., Wang, J.Z.: Studying aesthetics in photographic images using a computational approach. In: ECCV. (2006)
8. Dhar, S., Ordonez, V., Berg, T.L.: High level describable attributes for predicting aesthetics and interestingness. In: CVPR. (2011)
9. Gupta, A., Kembhavi, A., Davis, L.S.: Observing human-object interactions: Using spatial and functional compatibility for recognition. PAMI (2009)
10. Yao, B., Fei-Fei, L.: Action recognition with exemplar based 2.5d graph matching. In: ECCV. (2012)
11. Yang, W., Wang, Y., Mori, G.: Recognizing human actions from still images with latent poses. In: CVPR. (2010)
12. Maji, S., Bourdev, L., Malik, J.: Action recognition from a distributed representation of pose and appearance. In: CVPR. (2011)
13. Delaitre, V., Sivic, J., Laptev, I., et al.: Learning person-object interactions for action recognition in still images. In: NIPS. (2011)
14. Laptev, I., Perez, P.: Retrieving actions in movies. In: ICCV. (2007)
15. Sadanand, S., Corso, J.J.: Action bank: A high-level representation of activity in video. In: CVPR. (2012)
16. Rodriguez, M., Ahmed, J., Shah, M.: Action mach a spatio-temporal maximum average correlation height filter for action recognition. In: CVPR. (2008) 1–8
17. Efros, A., Berg, A., Mori, G., Malik, J.: Recognizing action at a distance. In: CVPR. (2003)
18. Shechtman, E., Irani, M.: Space-time behavior based correlation. In: PAMI. (2007)

19. Poppe, R.: A survey on vision-based human action recognition. Image and Vision Computing **28**(6) (2010) 976–990
20. Aggarwal, J.K., Ryoo, M.S.: Human activity analysis: A review. ACM Comput. Surv. 16
21. Wang, H., Ullah, M.M., Klaser, A., Laptev, I., Schmid, C.: Evaluation of local spatio-temporal features for action recognition. In: BMVC. (2009)
22. Niebles, J., Chen, C., Fei-Fei, L.: Modeling temporal structure of decomposable motion segments for activity classification. ECCV (2010)
23. Laptev, I.: On space-time interest points. ICCV (2005)
24. Le, Q.V., Zou, W.Y., Yeung, S.Y., Ng, A.Y.: Learning hierarchical invariant spatio-temporal features for action recognition with independent subspace analysis. In: CVPR. (2011)
25. Wang, J., Liu, Z., Wu, Y., Yuan, J.: Mining actionlet ensemble for action recognition with depth cameras. In: CVPR. (2012)
26. Ekin, A., Tekalp, A.M., Mehrotra, R.: Automatic soccer video analysis and summarization. Transactions on Image Processing (2003)
27. Khosla, A., Hamid, R., Lin, C.J., Sundaresan, N.: Large-scale video summarization using web-image priors. In: CVPR. (2013)
28. Gong, Y., Liu, X.: Video summarization using singular value decomposition. In: CVPR. (2000)
29. Rav-Acha, A., Pritch, Y., Peleg, S.: Making a long video short: Dynamic video synopsis. In: CVPR. (2006)
30. Ngo, C.W., Ma, Y.F., Zhang, H.J.: Video summarization and scene detection by graph modeling. Circuits and Systems for Video Technology (2005)
31. Jiang, R.M., Sadka, A.H., Crookes, D.: Hierarchical video summarization in reference subspace. Consumer Electronics, IEEE Transactions on (2009)
32. Marszalek, M., Laptev, I., Schmid, C.: Actions in context. In: CVPR. (2009)
33. Yang, Y., Ramanan, D.: Articulated pose estimation with flexible mixtures-of-parts. In: CVPR. (2011)
34. Park, D., Ramanan, D.: N-best maximal decoders for part models. In: ICCV. (2011)
35. Drucker, H., Burges, C.J., Kaufman, L., Smola, A., Vapnik, V.: Support vector regression machines. NIPS (1997)
36. Chang, C.C., Lin, C.J.: Libsvm: a library for support vector machines. ACM Transactions on Intelligent Systems and Technology (TIST) (2011)
37. Silberman, N., Hoiem, D., Kohli, P., Fergus, R.: Indoor segmentation and support inference from rgbd images. In: ECCV. (2012)
38. Zheng, B., Zhao, Y., Yu, J.C., Ikeuchi, K., Zhu, S.C.: Detecting potential falling objects by inferring human action and natural disturbance. In: IEEE Int. Conf. on Robotics and Automation (ICRA). (2014)