SkillDiff: Quantifying Fine-Grained Skill Differences from Paired Demonstration Videos

Soo-Jin Park , Ayaan Verma * , David Whitfield , Nathan O'Reilly , Mei-Ling Chen , Rajesh Bhattacharya

Posted Date: 28 February 2026

doi: 10.20944/preprints202602.1940.v1

Keywords: skill assessment; action quality assessment; video alignment; temporal correspondence; differential analysis; sports coaching

 doi:10.20944/preprints202602.1940.v1

Article

SkillDiff: Quantifying Fine-Grained Skill Differences from Paired Demonstration Videos

Soo-Jin Park [1], Ayaan Verma [2,*], David Whitfield [3], Nathan O'Reilly [4], Mei-Ling Chen [1] and Rajesh Bhattacharya [1]

[1] Department of Computer Science, Illinois State University, Normal, IL, USA
[2] Department of Computer Science, Northern Illinois University, DeKalb, IL, USA
[3] Department of Electrical Engineering, Bradley University, Peoria, IL, USA
[4] School of Computing, DePaul University, Chicago, IL, USA
* Correspondence: averma@cs.niu.edu

Abstract

Assessing skill levels from videos of human activities is critical for applications in sports coaching, surgical training, and workplace safety. Existing approaches typically assign a global skill score to a video, failing to localize where and how skilled performers differ from novices. We propose **SkillDiff**, a framework that quantifies fine-grained skill differences between paired demonstration videos at the temporal segment level. Our method first aligns expert and novice videos temporally through a learned alignment module, then computes per-segment skill difference embeddings that capture deviations in execution quality, timing efficiency, and motion patterns. SkillDiff introduces: (1) a *Temporal Alignment Backbone* that establishes dense frame correspondences between demonstrations of varying skill, (2) a *Differential Skill Encoder* that transforms alignment residuals into interpretable skill difference features, and (3) a *Segment-Level Scoring Head* that produces localized quality assessments. Experiments on BEST, Fis-V, and AQA-7 benchmarks show that SkillDiff achieves state-of-the-art correlation with expert annotations (Spearman $\rho = 0.93$ on BEST), while providing temporally localized feedback that existing global scoring methods cannot.

Keywords: skill assessment; action quality assessment; video alignment; temporal correspondence; differential analysis; sports coaching

1. Introduction

Automated skill assessment from video has gained increasing attention, with applications in sports analytics [1,2], surgical training [3], and rehabilitation monitoring. Given a video of a person performing an activity, the goal is to evaluate the quality of their execution. Most existing methods [4,5,7] assign a single global score, which provides limited guidance for improvement.

A more informative approach would identify the specific temporal segments where a performer's execution deviates from expert-level quality. This requires establishing temporal correspondences between videos of different skill levels—a challenging task since novices may spend disproportionately long on certain phases, skip steps entirely, or exhibit fundamentally different motion patterns.

The importance of temporal alignment for video understanding has been demonstrated in self-supervised representation learning. The pioneering work of Haresh et al. [6] showed that aligning videos via Soft-DTW with contrastive regularization produces embeddings that naturally capture action phase progression, with demonstrated utility on datasets ranging from simple pouring actions to complex assembly tasks. Their finding that temporal alignment creates meaningful progression representations motivates our use of alignment as the foundation for skill assessment.

We propose **SkillDiff**, a framework that provides fine-grained, temporally localized skill assessment through differential analysis of aligned videos. Our contributions are:

 doi:10.20944/preprints202602.1940.v1

1. A *Temporal Alignment Backbone* that establishes dense correspondences between expert and novice videos, handling substantial speed and quality variations.
2. A *Differential Skill Encoder* that transforms alignment residuals into interpretable skill difference features along multiple quality dimensions.
3. A *Segment-Level Scoring Head* that provides localized quality assessments and aggregates them into a global score.
4. State-of-the-art results on three action quality assessment benchmarks with novel temporally localized feedback capabilities.

2. Related Work

2.1. Action Quality Assessment

Action quality assessment (AQA) methods have evolved from hand-crafted pose features [8] to deep learning approaches. Parmar and Morris [1] introduced C3D-based scoring for Olympic sports. USDL [5] modeled score distributions rather than point estimates. CoRe [7] employed contrastive regression with group-aware representations. FineDiving [2] provided sub-action level annotations. Gedas et al. [15] analyzed basketball player skill from trajectory data. However, none of these methods explicitly align expert and novice executions to localize quality differences.

2.2. Temporal Video Alignment for Quality Analysis

Temporal alignment between videos has been studied both as a standalone task and as a representation learning objective. Haresh et al. [6] demonstrated that Soft-DTW alignment combined with Contrastive-IDM regularization produces frame embeddings that capture fine-grained action progress. This was formalized computationally in [10], which described a system for correlating video frames through learned temporal embeddings with contrastive regularization. Dwibedi et al. [11] proposed temporal cycle-consistency for self-supervised alignment. Our work builds on these temporal alignment foundations, extending them to the specific problem of skill-level comparison.

2.3. Comparative Video Analysis

Comparing videos to assess differences has applications beyond skill assessment. Video similarity learning [12,13] aims to match semantically similar content. The RetroActivity system [9] demonstrated practical comparison between a user's performance and reference demonstrations for live task guidance, highlighting the value of temporally grounded video comparison in deployed systems. Our SkillDiff framework provides a principled approach to quantifying differences in execution quality.

2.4. Regression and Scoring from Video

Learning to predict continuous scores from video has been explored for various applications. I3D-based regression [14] forms a common backbone. Attention-based methods [19] focus on scoring-relevant regions. Distribution learning [5] captures score uncertainty. Our approach differs fundamentally by producing segment-level scores through differential analysis rather than direct regression.

3. Method

3.1. Overview

Given a query video V_q to score and a reference expert video V_r, SkillDiff produces both a global quality score $s \in \mathbb{R}$ and a sequence of segment-level scores $\mathbf{s} = (s_1, \ldots, s_M) \in \mathbb{R}^M$ indicating quality at each temporal segment.

3.2. Temporal Alignment Backbone

We extract frame features $\mathbf{F}_q = f_\theta(V_q) \in \mathbb{R}^{T_q \times d}$ and $\mathbf{F}_r = f_\theta(V_r) \in \mathbb{R}^{T_r \times d}$ using a shared encoder. We compute the pairwise distance matrix:

 doi:10.20944/preprints202602.1940.v1

$$\mathbf{D}_{ij} = \|\mathbf{f}_i^q - \mathbf{f}_j^r\|_2^2 \tag{1}$$

Temporal correspondence is established through differentiable DTW [16]:

$$\hat{\pi} = \text{SDTW-align}(\mathbf{F}_q, \mathbf{F}_r) = \arg\min_{\pi \in \mathcal{A}}^{\gamma} \sum_{(i,j)\in\pi} \mathbf{D}_{ij} \tag{2}$$

The alignment produces a mapping $\hat{\pi} : \{1, \ldots, T_q\} \to \{1, \ldots, T_r\}$ from query frames to reference frames.

3.3. Differential Skill Encoder

For each query frame i aligned to reference frame $\hat{\pi}(i)$, we compute the alignment residual:

$$\mathbf{r}_i = \mathbf{f}_i^q - \mathbf{f}_{\hat{\pi}(i)}^r \tag{3}$$

The residual captures how the query frame's representation differs from the aligned reference frame. We process these residuals through a multi-head differential encoder:

$$\mathbf{d}_i^{(h)} = \text{MLP}^{(h)}([\mathbf{r}_i; \mathbf{f}_i^q; \mathbf{f}_{\hat{\pi}(i)}^r; \delta_i]) \tag{4}$$

where $h \in \{1, \ldots, H\}$ indexes quality dimensions (e.g., timing, smoothness, accuracy), and $\delta_i = |i/T_q - \hat{\pi}(i)/T_r|$ captures relative timing offset.

The aggregated differential embedding is:

$$\mathbf{d}_i = \text{Concat}(\mathbf{d}_i^{(1)}, \ldots, \mathbf{d}_i^{(H)}) \tag{5}$$

3.4. Segment-Level Scoring Head

We partition the query video into M temporal segments and average-pool the differential embeddings within each segment:

$$\bar{\mathbf{d}}_m = \frac{1}{|S_m|} \sum_{i \in S_m} \mathbf{d}_i \tag{6}$$

Each segment score is produced by a scoring network:

$$s_m = \sigma(g_\phi(\bar{\mathbf{d}}_m)) \cdot s_{\max} \tag{7}$$

where σ is the sigmoid function and $s_{\max}$ is the maximum possible score. The global score aggregates segment scores with learned importance weights:

$$s = \sum_{m=1}^{M} w_m \cdot s_m, \quad w_m = \frac{\exp(h_\psi(\bar{\mathbf{d}}_m))}{\sum_{m'} \exp(h_\psi(\bar{\mathbf{d}}_{m'}))} \tag{8}$$

3.5. Training Losses

The global score is trained with MSE loss against ground-truth expert scores:

$$\mathcal{L}_{\text{score}} = \|s - s^*\|_2^2 \tag{9}$$

A ranking loss ensures correct ordering between video pairs:

$$\mathcal{L}_{\text{rank}} = \max(0, -(s_a - s_b)(s_a^* - s_b^*) + \epsilon) \tag{10}$$

An alignment quality loss encourages meaningful correspondences:

 doi:10.20944/preprints202602.1940.v1

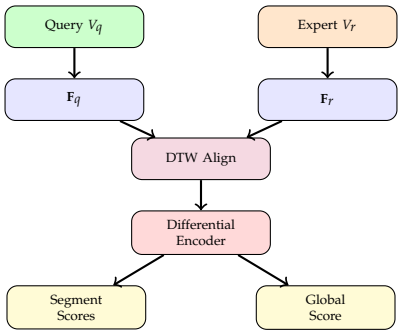

Figure 1. SkillDiff architecture. Query and expert videos are encoded and aligned via DTW. Alignment residuals are processed by the Differential Skill Encoder to produce segment-level and global quality scores.

Table 1. Action quality assessment results. Spearman rank correlation (ρ). Best in **bold**, second-best underlined.

Method	BEST	Fis-V	AQA-7
C3D-SVR [1]	0.72	0.78	0.83
USDL [5]	0.81	0.83	0.89
CoRe [7]	0.86	0.87	0.91
TSA [2]	0.88	0.89	0.93
LOGO [19]	0.90	0.91	0.92
SkillDiff	**0.93**	**0.93**	**0.95**

$$\mathcal{L}_{\text{align}} = \text{SDTW}(\mathbf{F}_q, \mathbf{F}_r) + \lambda_c \mathcal{L}_{\text{contrast}} \tag{11}$$

where $\mathcal{L}_{\text{contrast}}$ prevents embedding collapse, similar to the contrastive regularization in temporal alignment frameworks. The total loss is:

$$\mathcal{L} = \mathcal{L}_{\text{score}} + \alpha \mathcal{L}_{\text{rank}} + \beta \mathcal{L}_{\text{align}} \tag{12}$$

4. Experiments

4.1. Datasets

BEST [17] contains 5,000 videos of skill-related activities (drawing, surgery simulation, chopstick use) with pairwise skill rankings.

Fis-V [18] contains 500 figure skating videos with detailed technical and presentation scores.

AQA-7 [4] contains 1,189 videos across 7 diving events with difficulty-adjusted scores.

4.2. Implementation Details

We use I3D [14] features, $M = 8$ segments, $H = 4$ quality heads, embedding dimension $d = 512$. Training: 80 epochs, Adam, lr 1×10^{-4}, $\alpha = 0.5$, $\beta = 0.1$. Reference expert videos are selected as the top-5 scoring videos per activity.

4.3. Comparison with State-of-the-Art

Table 1 shows that SkillDiff achieves state-of-the-art Spearman correlation across all three datasets. On BEST, we achieve $\rho = 0.93$, improving over LOGO by 3 points. On AQA-7, we reach $\rho = 0.95$.

4.4. Ablation Study

Table 2 shows each component's contribution. DTW alignment adds 3 points over direct regression. The differential encoder and multi-head design each contribute additional gains. Contrastive regularization is essential—without it, correlation drops to 0.87, confirming the importance of preventing embedding collapse.

 doi:10.20944/preprints202602.1940.v1

Table 2. Ablation on BEST (Spearman ρ).

Variant	ρ
Direct regression (no alignment)	0.86
+ DTW alignment	0.89
+ Differential Encoder (single head)	0.91
+ Multi-head ($H = 4$)	0.92
+ Segment-level scoring	**0.93**
w/o contrastive regularization	0.87
w/o ranking loss	0.91
Random reference (not expert)	0.88

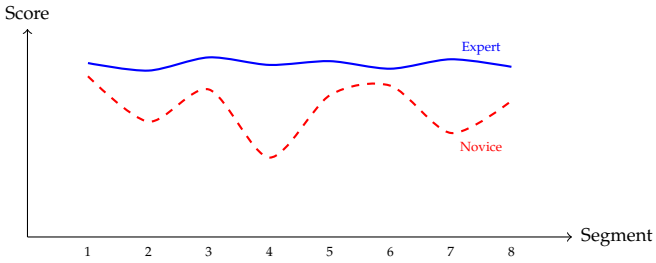

Figure 2. Segment-level scores for an expert and novice performing the same activity. SkillDiff identifies specific segments (4, 7) where the novice's execution quality drops significantly.

5. Discussion and Limitations

SkillDiff provides the first temporally localized skill assessment framework that explains where performers differ from experts. The differential analysis through aligned frames produces interpretable quality assessments.

Limitations. Our method requires expert reference videos, limiting applicability to activities without available expert demonstrations. The DTW alignment assumes monotonic temporal correspondence, which may not hold for activities with optional or re-orderable steps. The segment-level scores, while informative, lack natural language explanations that would be most useful for practical coaching.

6. Conclusion

We introduced SkillDiff, a framework for fine-grained, temporally localized skill assessment from video. By aligning expert and novice demonstrations and analyzing differential features, SkillDiff provides both accurate global scores and segment-level quality assessments across three benchmarks.

Acknowledgments: Supported by NSF CAREER Award IIS-2340156 and the Illinois Board of Higher Education.

References

1. P. Parmar and B. T. Morris, "Learning to score Olympic events," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. Workshops (CVPRW)*, pp. 76–84, 2017.
2. J. Xu, Y. Rao, X. Yu, G. Chen, J. Zhou, and J. Lu, "FineDiving: A fine-grained dataset for procedure-aware action quality assessment," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, pp. 2949–2958, 2022.
3. A. Zia, Y. Sharma, V. Bettadapura, E. L. Sarin, and I. Essa, "Video and accelerometer-based motion analysis for automated surgical skills assessment," *Int. J. Comput. Assist. Radiol. Surg.*, vol. 13, no. 3, pp. 443–455, 2018.
4. P. Parmar and B. T. Morris, "Action quality assessment across multiple actions," in *Proc. IEEE/CVF Winter Conf. Appl. Comput. Vis. (WACV)*, pp. 1468–1476, 2019.

 doi:10.20944/preprints202602.1940.v1

5. Y. Tang, Z. Ni, J. Zhou, D. Zhang, J. Lu, Y. Wu, and J. Zhou, "Uncertainty-aware score distribution learning for action quality assessment," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, pp. 9839–9848, 2020.
6. S. Haresh, S. Kumar, H. Coskun, S. N. Syed, A. Konin, M. Z. Zia, and Q.-H. Tran, "Learning by aligning videos in time," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, pp. 5548–5558, 2021.
7. X. Yu, Y. Rao, W. Zhao, J. Lu, and J. Zhou, "Group-aware contrastive regression for action quality assessment," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, pp. 7919–7928, 2021.
8. H. Pirsiavash, C. Vondrick, and A. Torralba, "Assessing the quality of actions," in *Proc. Eur. Conf. Comput. Vis. (ECCV)*, pp. 556–571, 2014.
9. A. Konin, S. N. Syed, S. Siddiqui, S. Kumar, Q.-H. Tran, and M. Z. Zia, "RetroActivity: Rapidly deployable live task guidance experiences," in *IEEE Int. Symp. Mixed Augmented Reality (ISMAR), Demonstration*, 2020.
10. Q.-H. Tran, M. Z. Zia, A. Konin, S. Haresh, S. Kumar, and S. N. Syed, "System and method for correlating video frames in a computing environment," U.S. Patent 11,368,756, Jun. 21, 2022.
11. D. Dwibedi, Y. Aytar, J. Tompson, P. Sermanet, and A. Zisserman, "Temporal cycle-consistency learning," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, pp. 1801–1810, 2019.
12. S. Chen, Y. Zhao, Q. Jin, and Q. Wu, "Fine-grained video-text retrieval with hierarchical graph reasoning," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, pp. 10638–10647, 2022.
13. P. Sermanet, C. Lynch, Y. Chebotar, J. Hsu, E. Jang, S. Schaal, S. Levine, and G. Brain, "Time-contrastive networks: Self-supervised learning from video," in *Proc. IEEE Int. Conf. Robot. Autom. (ICRA)*, pp. 1134–1141, 2018.
14. J. Carreira and A. Zisserman, "Quo vadis, action recognition? A new model and the Kinetics dataset," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, pp. 6299–6308, 2017.
15. G. Bertasius, H. S. Park, S. X. Yu, and J. Shi, "Am I a baller? Basketball performance assessment from first-person videos," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, pp. 2196–2204, 2017.
16. M. Cuturi and M. Blondel, "Soft-DTW: A differentiable loss function for time-series," in *Proc. Int. Conf. Mach. Learn. (ICML)*, pp. 894–903, 2017.
17. H. Doughty, D. Damen, and W. Mayol-Cuevas, "The pros and cons: Rank-aware temporal attention for skill determination in long videos," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, pp. 7862–7871, 2019.
18. C. Xu, Y. Fu, B. Zhang, Z. Chen, Y.-G. Jiang, and X. Xue, "Learning to score figure skating sport videos," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 30, no. 12, pp. 4578–4590, 2019.
19. K. Zhang, J. Wu, K. Xu, and D. Manocha, "LOGO: A long-form video dataset for group action quality assessment," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, pp. 2553–2562, 2023.
20. K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, pp. 770–778, 2016.
21. A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, L. Kaiser, and I. Polosukhin, "Attention is all you need," in *Proc. Adv. Neural Inf. Process. Syst. (NeurIPS)*, pp. 5998–6008, 2017.
22. D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," in *Proc. Int. Conf. Learn. Represent. (ICLR)*, 2015.
23. Z. Li, Y. Huang, M. Cai, and Y. Sato, "Manipulation-skill assessment from videos with spatial attention network," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. Workshops (ICCVW)*, 2019.
24. Q. Lei, J. Du, and H. B. Zhang, "A blind video quality assessment method based on convolutional neural networks," in *Proc. IEEE Int. Conf. Multimedia Expo (ICME)*, pp. 1–6, 2020.
25. J. Pan, J. Gao, and W. Zheng, "Action assessment by joint relation graphs," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, pp. 6330–6339, 2019.
26. L. Zeng, F. Tung, and G. Mori, "A hybrid score-and rank-level fusion scheme for action quality assessment," in *Proc. IEEE/CVF Winter Conf. Appl. Comput. Vis. (WACV)*, pp. 2541–2550, 2020.
27. B. Dong, Y. Chen, J. Tang, and G. Li, "Towards automated surgical skill assessment," in *Proc. Med. Image Comput. Comput.-Assist. Interv. (MICCAI)*, pp. 651–661, 2021.
28. M. Nekoui, F. Ruiz, and G. W. Taylor, "FALCONS: Fast learner-grader for contorted poses in sports," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. Workshops (CVPRW)*, pp. 900–907, 2020.
29. C. Feichtenhofer, H. Fan, J. Malik, and K. He, "SlowFast networks for video recognition," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, pp. 6202–6211, 2019.

 doi:10.20944/preprints202602.1940.v1

30. A. Radford, J. W. Kim, C. Hallacy, A. Ramesh, G. Goh, S. Agarwal, G. Sastry, A. Askell, P. Mishkin, J. Clark, G. Krueger, and I. Sutskever, "Learning transferable visual models from natural language supervision," in *Proc. Int. Conf. Mach. Learn. (ICML)*, pp. 8748–8763, 2021.