Proceedings of the ASME 2024
International Symposium on Flexible Automation
ISFA2024
July 21–24, 2024, Seattle, WA

ISFA2024-140665

REPETITIVE ACTION COUNTING THROUGH JOINT ANGLE ANALYSIS AND VIDEO TRANSFORMER TECHNIQUES

Haodong Chen[1,*], Niloofar Zendehdel[1], Ming C. Leu[1], Md Moniruzzaman[2], Zhaozheng Yin[2], Solmaz Hajmohammadi[3]

[1]Department of Mechanical and Aerospace Engineering, Missouri University of Science and Technology, Rolla, MO
[2]Department of Biomedical Informatics and Department of Computer Science, Stony Brook University, Stony Brook, NY
[3]Peloton Interactive, New York, NY

ABSTRACT

The quantification of repetitive movements, known as repetitive action counting, is critical in various applications, such as fitness tracking, rehabilitation, and manufacturing operation monitoring. Traditional methods predominantly relied on the estimation of red-green-and-blue (RGB) frames and body pose landmarks to identify the number of action repetitions. However, these methods suffer from several issues, such as instability under varying camera viewpoints, propensity for over-counting or under-counting, challenges in differentiating sub-actions, and inaccuracies in recognizing salient action poses, etc. Our method integrates joint angles with body pose landmarks to address these issues, thereby surpassing the performance benchmarks of existing state-of-the-art repetitive action counting methodologies. The efficacy of our approach is underscored by a Mean Absolute Error (MAE) of 0.211 and an Off-By-One Accuracy (OBOA) of 0.599 on a public repetitive action counting data set, RepCount [1]. Comprehensive experimental results demonstrate the effectiveness and robustness of our method.

Keywords: Repetitive action counting, Pose estimation, Transformer, Skeleton, Pose landmarks

1. INTRODUCTION

Repetitive actions are a fundamental component of a myriad of activities spanning from physical exercise to precision-oriented tasks such as experimental operations and assembly processes [2]. In the realm of physical activity and exercise, the accurate quantification of repetitive movements can greatly enhance the effectiveness of training regimens and rehabilitation programs by ensuring that exercises are performed correctly and consistently [3, 4]. Similarly, in scientific experimentation, the precision and repeatability of actions are paramount, as they directly influence the validity and reliability of experimental results. Assembly operations, whether in manufacturing or delicate tasks such as machine assembly, also rely heavily on the meticulous repetition of actions to maintain quality and efficiency. Consequently, the ability to analyze and quantify repetitive actions becomes a crucial tool for ensuring correctness, efficiency, and quality across diverse fields. This underscores the importance of developing robust methods for repetitive action counting analysis that are adaptable and accurate across varying contexts and applications [5–10].

1.1 Related Works

Existing repetitive action counting methods, such as the method detailed by [1], predominantly utilized inputs from red-green-and-blue (RGB) frame inputs. This preference for RGB inputs stems from their simplicity and directness [11, 12]. While it provided a solution for repetitive action counting but could not independently isolate and recognize periodic movement [13, 14]. Alternative strategies, as explored by [15] et al. and [16] et al., have achieved better performance by using contextual information in repetitive actions. Furthermore, [17] proposed a pose saliency Transformer for repetitive action counting, setting a new benchmark in counting accuracy, which introduces a new mechanism called Pose Saliency Representation (PSR). This mechanism uses the two most salient poses to represent the action, providing a more streamlined and efficient representation than the RGB frame-based representation. Unlike conventional methods that rely on intricate computations to extract high-level semantic information from the spatial and temporal dimensions within frames, PSR simplifies this process. Based on the PSR, a Pose Saliency Transformer for repetitive action counting is proposed in [17]. This framework consists of three components: i) Pose extraction, where pose information is identified and isolated from each frame. ii) A video-Transformer-based model that maps each extracted pose to an action category; and iii) A lightweight action trigger mechanism designed for video-level repetitive action counting, which quantifies the occurrences of specific actions throughout the video sequence. In this model, the repetitive action counting predominantly hinges on human pose estimation,

*Corresponding author: h.chen@mst.edu

which uses landmark detection to identify the joint positions of human skeletons. Despite its innovative approach, this method often encounters various challenges. These include instability under varying camera viewpoints, propensity for over-counting or under-counting repetitions, challenges in differentiating sub-actions, difficulties in accurately identifying salient poses, etc. Such limitations underscore the need for enhanced techniques that can reliably address these issues, thereby improving the robustness and accuracy of repetitive action counting.

In our work, we refine the action counting method by integrating joint angles with the pose landmark to address the repetitive action counting problem using a Transformer network. By leveraging this integrated model, our system achieves a more comprehensive understanding of repetitive actions, leading to significant improvements in the accuracy and robustness of counting mechanisms. Our method addresses several existing issues, such as instability under varying camera viewpoints, propensity for over-counting or under-counting, challenges in differentiating sub-actions, inaccuracies in recognizing salient action poses, etc. Our model significantly improves the accuracy and robustness of repetitive action counting, as evidenced by the *RepCount* data set introduced by [1].

1.2 Contribution

The contribution of this paper is as follows:

- We analyze different combinations of joint angles and body pose landmarks in effectively solving the repetitive action counting problem.
- We improve the repetitive action counting performance in addressing issues such as instability under varying camera viewpoints, propensity for over-counting or under-counting, challenges in differentiating sub-actions, and inaccuracies in recognizing salient action poses, etc.
- Our experimental results obtain better performance than the state-of-the-art results on the well-established public data set.

The structure of the rest contents is as follows: we first introduce our approach in Section 2. Next, we provide experimental results in Section 3. Finally, we give conclusions in Section 4.

2. POSE AND JOINT ANGLE ANNOTATION

As shown in Fig. 1, the extraction of 33 pose landmarks is carried out by the Google Mediapipe BlazePose model. This approach achieves an accuracy of 84.50% accuracy on the public Yoga dataset, which features varying backgrounds, outperforming other methods by 1.1% [18]. By using pose landmarks, the backgrounds are eliminated and only pose landmarks are used in repetitive action counting. Five joint angles are extracted based on the landmarks shown in Fig. 1, which are elbow, shoulder, hip, knee, and ankle angles, thereby providing a comprehensive framework for the evaluation of human posture and movement dynamics.

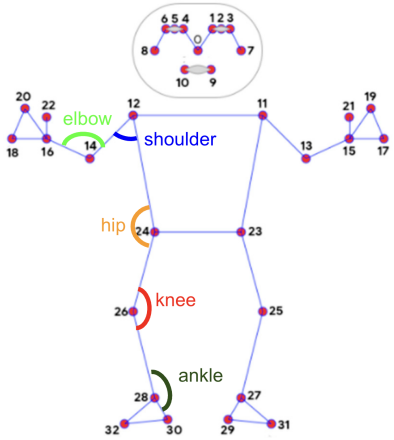


FIGURE 1: BlazePose landmarks and five joint angles

2.1 Salient Pose Annotation

As illustrated in Fig. 2, salient pose annotation [17] leverages just two salient poses to denote the features of each action, thereby establishing a unique mapping between salient poses and repetitive action counting. The presence of two adjacent different salient poses indicates a valid repetitive action. For example, the action *Jump Jack* has two salient poses. The salient pose I is an expansion pose, where hands meet above the head while legs are spread wide, engaging upper body and core muscles in a peak kinetic stance. The salient pose II would typically contrast this by returning limbs to a neutral position, and repositioning of limbs towards the body's midline.

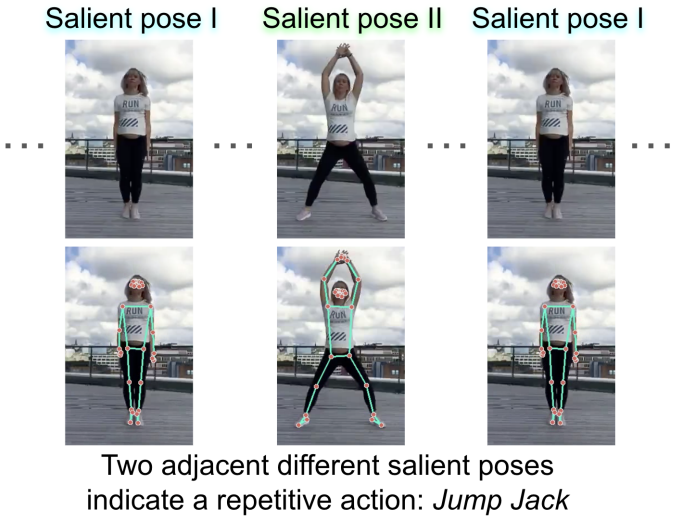


FIGURE 2: Salient pose annotation

2.2 Annotation Correction of the Public Data Set

After analyzing the original data set *RepCount* [1], we noticed that the ground truth count of the action data stu4_5.mp4 in the test data set was incorrectly labeled as 51, whereas the correct ground truth label should be 5 after a thorough review, and we corrected the label annotation. This discovery ensures

the integrity and reliability of the data set for subsequent analyses and applications.

2.3 Architecture of the Repetitive Action Counting Model

The pose mapping is input into the Transformer [19], which leveraging the attention mechanisms, processes only the skeletal and joint angle data to recognize the temporal patterns inherent to the repetitive actions. Figure 3 illustrates this process. The shifted window mechanism is carried out to realize cross-window connections while maintaining the computational efficiency of the non-overlapping window-based self-attention approach.

After pose mapping using the Transformer in [19], we can obtain the density score for each frame and generate a density map from the obtained scores, as shown in Fig. 4. Higher values indicate a higher similarity to the salient pose I, while lower values indicate a higher match to the salient pose II. The action-trigger mechanism is used to compute the time at which two salient poses appear in sequence in an action category, where a specific upper and lower limit is set to differentiate the scores of the two salient poses, thus clustering the non-salient poses in the middle and easily categorizing the salient poses at both ends [20–22].

3. EXPERIMENTS AND RESULTS

3.1 Experiment Setup

The experimental platform is a workstation with an Ubuntu 16.04 system equipped with an Intel Xeon Gold 6226R CPU, an NVIDIA GeForce RTX 3090 graphics card, and 64343M RAM. The *RepCount* dataset, having over 700 videos, was partitioned into training, validation, and testing subsets following a 60:20:20 ratio.

3.2 Evaluation of Different Scenarios Using Joint Angles

We evaluate different scenarios using joint angles in solving the repetitive action counting problem, including:

- Using the 33 landmarks alongside five specific left joint angles: left elbow ($A_{e,\text{left}}$), left shoulder ($A_{s,\text{left}}$), left hip ($A_{h,\text{left}}$), left knee ($A_{k,\text{left}}$), and left ankle ($A_{a,\text{left}}$).
- Using the 33 landmarks alongside five specific right joint angles: right elbow ($A_{e,\text{left}}$), right shoulder ($A_{s,\text{right}}$), right hip ($A_{h,\text{right}}$), right knee ($A_{k,\text{right}}$), and right ankle ($A_{a,\text{right}}$).
- Using the 33 landmarks in conjunction with both left and right joint angles for five key joints: left elbow ($A_{e,\text{left}}$), right elbow ($A_{e,\text{right}}$), left shoulder ($A_{s,\text{left}}$), right shoulder ($A_{s,\text{right}}$), left hip ($A_{h,\text{left}}$), right hip ($A_{h,\text{right}}$), left knee ($A_{k,\text{left}}$), right knee ($A_{k,\text{right}}$), left ankle ($A_{a,\text{left}}$), and right ankle ($A_{a,\text{right}}$).
- Using the 33 landmarks alongside the average values of the five joint angles for both the left and right sides, i.e., denoted as: average elbow angle ($\bar{A}_e$), average shoulder angle ($\bar{A}_s$), average hip angle ($\bar{A}_h$), average knee angle ($\bar{A}_k$), and average ankle angle ($\bar{A}_a$).

We consider the single-side joint angles because all actions in the *RepCount* data set are symmetric. Each scenario is designed to assess the impact of joint angle configurations on the accuracy of repetitive action counts, thereby contributing to a comprehensive understanding of the problem space.

In this paper, we adopt the main evaluation metrics used in previous work [1, 15–17], i.e., mean absolute error (MAE) and off-by-one-accuracy (OBOA) counting accuracy. MAE represents the normalized absolute error between ground truth and prediction, while OBOA measures the repetitive count rate of the entire data set. They can be defined as:

$$MAE = \frac{1}{N}\sum_{i=1}^{N}\frac{|\widetilde{c_i} - c_i|}{\widetilde{c_i}} \tag{1}$$

$$OBOA = \frac{1}{N}\sum_{i=1}^{N}[|\widetilde{c_i} - c_i| \leq 1] \tag{2}$$

where $\widetilde{c_i}$ is the ground truth count, c_i is the prediction count, and N is the number of videos.

The comparison results for the different scenarios are shown in Table 1. We find that the case using landmarks and joint angles performed better than the case using only landmarks. In the cases using landmarks and joint angles, the best performance is obtained by using landmarks and the average values of the five left and right joint angles, in which the MAE $\approx$ 0.211 and OBOA $\approx$ 0.599. Table 2 shows that on the *RepCount* data set, our method consistently outperforms previous methods across both evaluation metrics. Specifically, the MAE achieved by our method is 0.211, which is lower than the 0.236 reported in [17]. Furthermore, the OBOA metric attained by our model is 0.599, surpassing the 0.559 obtained in [17].

3.3 Visualization Comparison of Models: Landmark-Only vs. Landmarks + Joint Angles

To validate the effectiveness of our proposed method, we visually analyze the output density maps obtained using two models: the model using only the landmarks and the model integrating the 5 average joint angles with the landmarks. The density map represents the density of a particular human pose in an input video sample. Higher values (close to 1.00) indicate a higher similarity to salient pose I, while lower values (close to 0.00) indicate a higher match to salient pose II. The density map provides insight into the distribution of the two salient poses throughout the video. Our experiments focus on visualizing and comparing the following issues in repetitive action counting: inability to stably deal with instability under varying camera viewpoints, over-counting, under-counting, difficulty in distinguishing sub-actions, inaccuracy in recognizing salient poses, etc.

Inability to stably deal with varying camera viewpoints: As shown in Fig. 5, we observe that the density map integrating the 5 average joint angles with the landmarks exhibits a more accurate capture of salient poses when the camera viewpoint changes. Specifically, the density values for the salient pose I remain consistently high (close to 1.00) after varying the camera viewpoint. In contrast, the density map using only the landmarks shows a significant drop in density values after the camera viewpoint changes from the front view to the side view, with density values $\approx$ 0.00 in the 440-800 frame range. This difference suggests that compared to using only the landmarks, using both

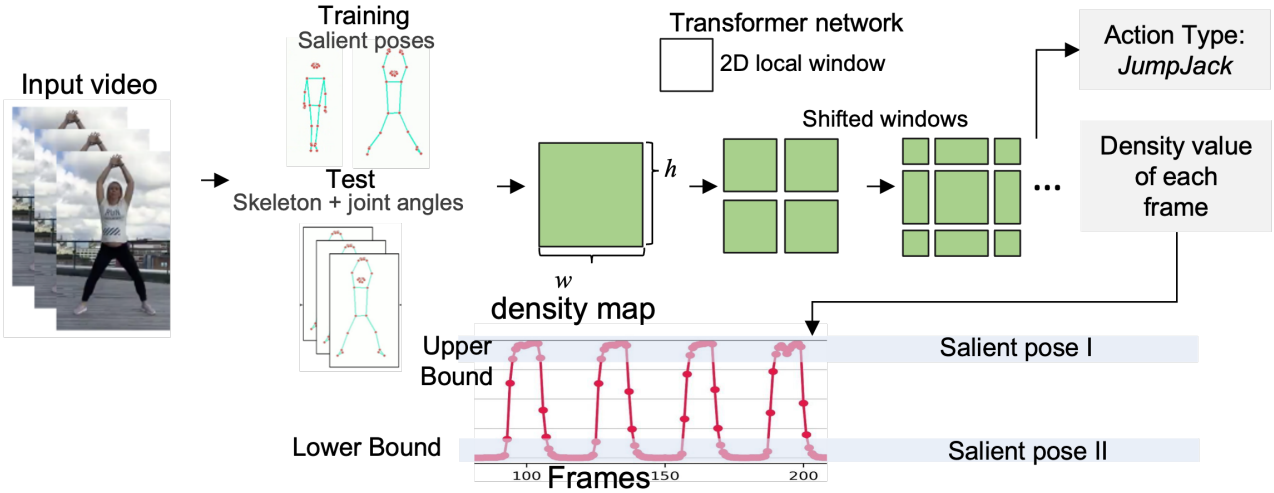


FIGURE 3: Architecture of the repetitive action counting model

TABLE 1: Comparison of different cases using landmarks and joint angles.

Different Cases	MAE↓	OBOA↑
Only landmarks	0.236	0.559
Landmarks + five left joint angles	0.227	0.571
Landmarks + five right joint angles	0.226	0.573
Landmarks + five left and right joint angles	0.213	0.587
Landmarks + average values of the five left and right joint angles	0.211	0.599

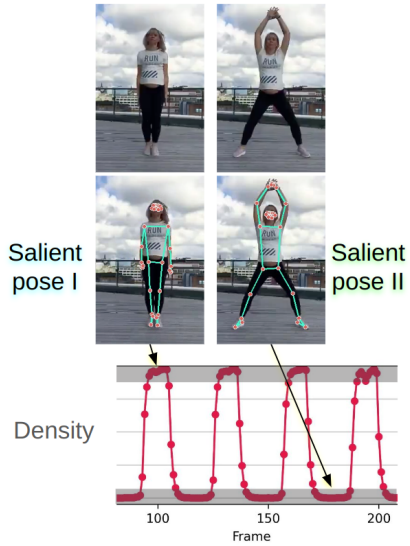


FIGURE 4: Action-trigger mechanism using density values

TABLE 2: Performance comparison on repetitive action counting

Input Type	Method	MAE↓	OBOA↑
Video-level	Zhang et al. [16]	0.879	0.155
	Huang et al. [23]	0.526	0.160
	Hu et al. [1]	0.443	0.291
Pose-level	Yao et al. [17]	0.236	0.559
	Our model	0.211	0.599

the landmarks and 5 average joint angles enables the model not to miss certain salient poses during camera viewpoint changes, thereby maintaining the accuracy of repetitive action counting.

Over-counting: As shown in Fig. 6, the density map generated from the landmarks alone tends to produce over-counts when a subject attempts to perform the *Pull Up* action but fails to complete it due to fatigue. The over counts are generated due to the subject's slight movement around the salient pose where the subject's arms are extended. In such cases, a landmark-only model may misinterpret these attempts as valid transitions between two salient poses, resulting in overcounts in repetitive action counting. However, incorporating the 5 average joint angles allows the model to recognize salient poses more accurately. As shown in Fig. 6, when the subject attempts to perform a *Pull Up* action but fails to complete it, this phenomenon is reflected in the density map obtained by integrating the 5 average joint angles with the landmarks, where the density value is lower than that for the salient pose I, i.e., the subject fully completes the *Pull Up* action with the arms bent. Integrating the 5 average joint angles helps the model recognize when a subject's attempts are unsuccessful, thus avoiding over-counting these partial or incomplete attempts.

Under-counting: As shown in Fig. 7, when a subject endeavors to perform a *Side Raise* action continuously, the density map derived exclusively from the landmarks fluctuates irregularly between 0.00 and 1.00, and shows multiple peaks between 0.50 and 0.75, rather than regularly fluctuating between 0.00 (salient posture II) and 1.00 (salient posture I). However, the density map

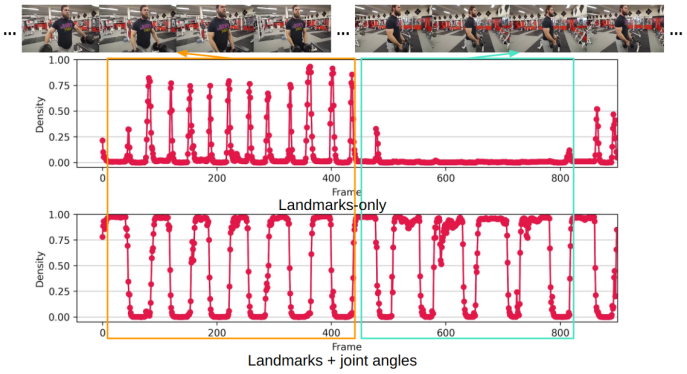


FIGURE 5: Density map comparison: landmarks-only vs. landmarks + joint angles - addressing the inability issue to stably deal with varying camera viewpoints

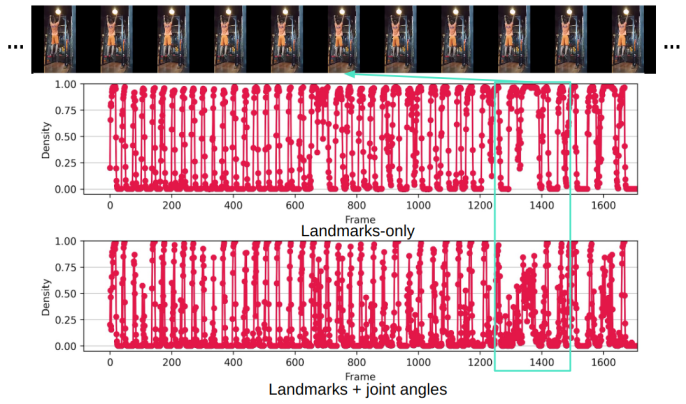


FIGURE 6: Density map comparison: landmarks-only vs. landmarks + joint angles - addressing the over-counting issue

integrating the 5 average joint angles with the landmarks shows density values regularly fluctuating between 0.00 and 1.00, representing the subject's movements between the two salient poses. Also, compared to the density map using only the landmarks, which cannot accurately identify the salient pose I in the 200-400 and 610-800 frame ranges, the density map obtained using both the landmarks and 5 average joint angles can accurately identify the salient pose I and provide the correct density values ≈ 1.00 for the salient pose I. This sample suggests that integrating the joint angles with the landmarks makes the model more sensitive to salient poses than using only the landmarks, thus improving performance when dealing with the under-counting issue.

Difficulty in distinguishing sub-actions: As shown in Fig. 8, an additional *Jump* action is present at the end of the *Pommel Horse* action, which has a lower limb feature similar to the regular *Pommel Horse* action and is a sub-action of the *Pommel Horse*. The density map obtained using only the landmarks incorrectly treats this *Jump* action as a valid count for *Pommel Horse*, with a peak close to 1.00 at around frame 340 in Fig. 8, which suggests that an additional count is computed using the landmark-only model, and illustrates the difficulty in distinguishing sub-actions in action counting using the landmark-only model. However, the density map integrating the 5 average joint angles with the landmarks effectively identifies the *Jump* action

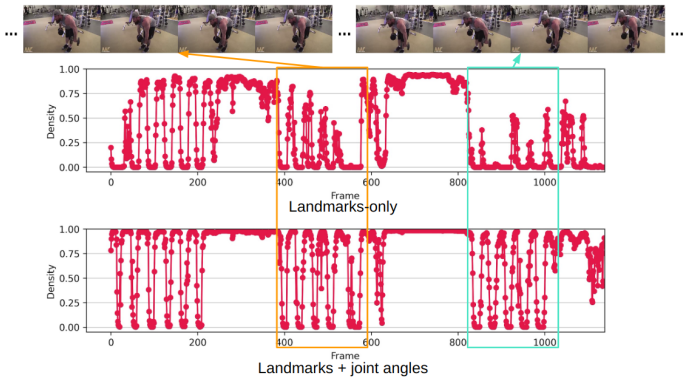


FIGURE 7: Density map comparison: landmarks-only vs. landmarks + joint angles - addressing the under-counting issue

is not a valid count for *Pommel Horse*, resulting in a lower density value ≈ 0.00 for this irrelevant sub-action, which in turn provides a more accurate repetitive action counting.

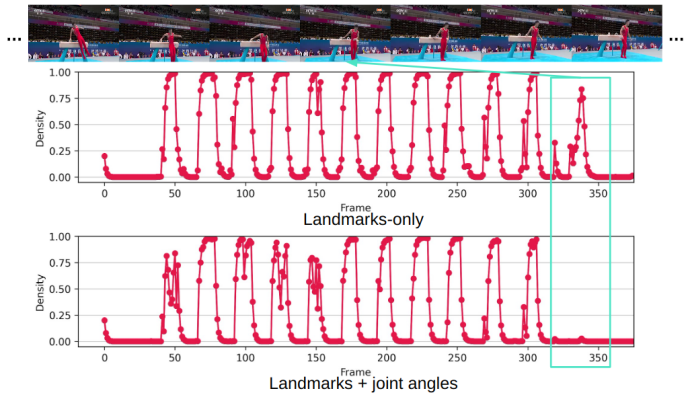
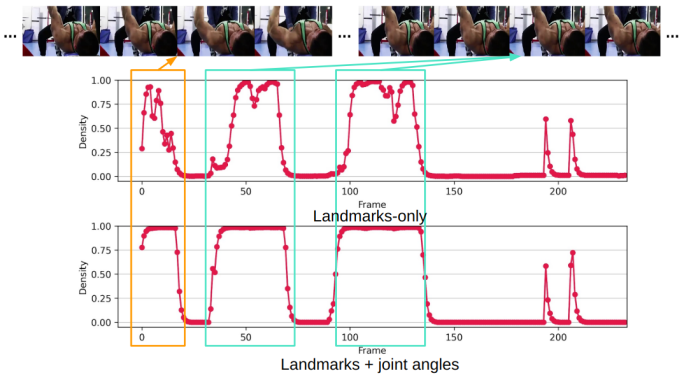


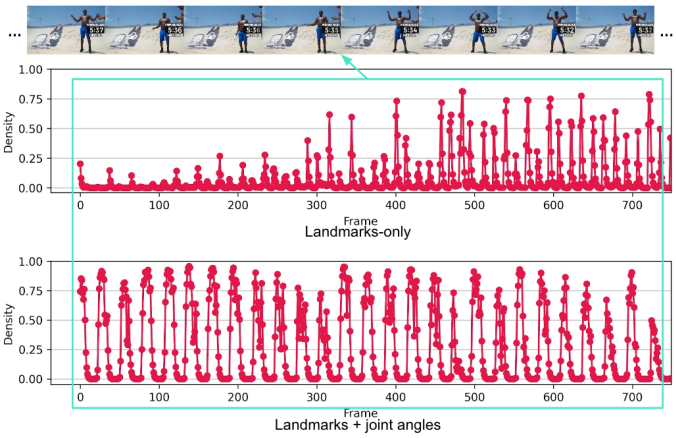
FIGURE 8: Density map comparison: landmarks-only vs. landmarks + joint angles - addressing the difficulty in distinguishing sub-actions issue

Inaccuracy in recognizing salient poses: As shown in Fig. 9a, it is difficult for a model using only the landmarks to provide consistent and reliable density values for the salient pose I (straightened arms) in the counting of the action *Bench Press*. Instability occurs when the subject is hindered by factors such as fatigue or slight movements during the execution of the salient pose I, resulting in drops in the density values of the continuous salient pose I, which should be close to 1.00. However, by integrating the joint angles with the landmarks, the model consistently and accurately identifies the salient pose I (straightened arms) with a stable density value of 1.00. A similar performance is shown in Fig. 9b, where the subject performs the *Jump Jack* action on the right side of the camera view. The density map obtained using only the landmarks does not accurately differentiate between the two salient poses, resulting in small-scale fluctuations between 0.00 and 0.25 in the beginning half of the density map, while the correct density value for the salient pose I should be close to 1.00. However, the density map obtained using both the landmarks and 5 average joint angles provides a clear counting of the *Jump Jack* action, with the density values fluctuating

uniformly between 0.00 and approximately 1.00.



(a) Addressing the inaccuracy in recognizing salient poses in *Bench Press*



(b) Addressing the Inaccuracy in recognizing salient poses in *Jump Jack*

FIGURE 9: Density map comparison: landmarks-only vs. landmarks + joint angles - addressing the inaccuracy issue in recognizing salient poses

These experimental results confirm the advantages of our proposed approach integrating the 5 average joint angles with the landmarks in solving the following issues in repetitive action counting: inability to stably deal with varying camera viewpoints, over-counting, under-counting, difficulty in distinguishing sub-actions, and inaccuracy in recognizing salient poses, making our method a robust and effective approach for the repetitive action counting task.

In addition to the improvements demonstrated above in the regular repetitive action counting cases, we have observed that integrating the 5 average joint angles with the landmarks can lead to effective and robust counting in video samples with various video effects, such as instantaneous brightness changes, zoom shifts, etc. Fig. 10 illustrates that the density map using both the landmarks and 5 average joint angles provides more accurate results in repetitive action counting than only using the landmarks.

4. CONCLUSION

In summary, this paper integrates the 5 average joint angles and body landmarks in solving the repetitive action counting

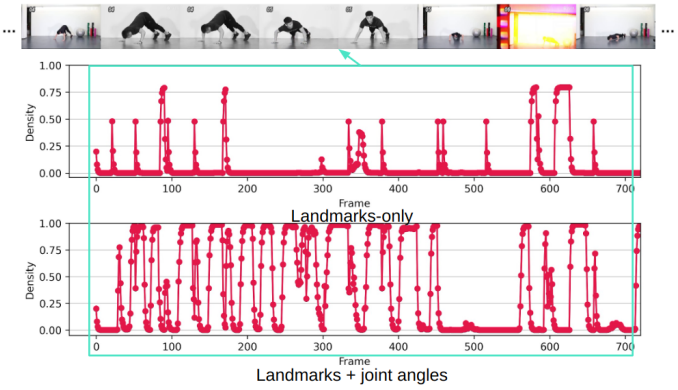


FIGURE 10: Density map comparison: landmarks-only vs. landmarks + joint angles - addressing the video effect issue

problem. Our method significantly improves the performance of repetitive action counting and provides the following improvements: i) Accurate performance in handling camera viewpoint variations. ii) Solving the over-counting and under-counting problems. iii) Improving the recognition of sub-actions. iv) Performing more accurate salient pose recognition. Our method obtains a mean absolute error (MAE) of 0.211 and an off-by-one accuracy (OBOA) counting accuracy of 0.599. Comprehensive experimental results demonstrate the effectiveness and robustness of our proposed method. This innovative method not only enhances overall performance but also effectively addresses the previously outlined challenges. Overall, our results outperform previous state-of-the-art methods and point the way to future research in the repetitive action counting problem area.

ACKNOWLEDGMENTS

This research work was financially supported by the National Science Foundation grant CMMI-1954548 and also by the Intelligent Systems Center at Missouri University of Science and Technology.

REFERENCES

[1] Hu, Huazhang, Dong, Sixun, Zhao, Yiqun, Lian, Dongze, Li, Zhengxin and Gao, Shenghua. "Transrac: Encoding multi-scale temporal correlation with transformers for repetitive action counting." *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*: pp. 19013–19022. 2022. URL https://doi.org/10.48550/arXiv.2204.01018.

[2] Chen, Haodong, Leu, Ming C, Moniruzzaman, Md, Yin, Zhaozheng, Hajmohammadi, Solmaz and Chang, Zhuoqing. "Advancements in Repetitive Action Counting: Joint-Based PoseRAC Model With Improved Performance." *arXiv preprint arXiv:2308.08632* (2023)URL https://doi.org/10.48550/arXiv.2308.08632.

[3] Zhao, Ping, Guan, Haiwei, Zhang, Yating, Chen, Yuwen, Deng, Xueting and Chen, Haodong. "Design of Single-DOF Immersive Upper Limb Rehabilitation System via Kinematic Mapping and Virtual Reality." *International Design Engineering Technical Conferences and Computers*

and Information in Engineering Conference, Vol. 83990: p. V010T10A038. 2020. American Society of Mechanical Engineers. URL https://doi.org/10.1115/DETC2020-22168.

[4] Chen, Hao-dong, Zhu, Hongbo, Teng, Zhiqiang and Zhao, Ping. "Design of a robotic rehabilitation system for mild cognitive impairment based on computer vision." *Journal of Engineering and Science in Medical Diagnostics and Therapy* Vol. 3 No. 2 (2020): p. 021108. URL https://doi.org/10.1115/1.4046396.

[5] Brickwood, Katie-Jane, Watson, Greig, O'Brien, Jane, Williams, Andrew D et al. "Consumer-based wearable activity trackers increase physical activity participation: systematic review and meta-analysis." *JMIR mHealth and uHealth* Vol. 7 No. 4 (2019): p. e11819. DOI 10.2196/11819.

[6] Chen, Haodong, Leu, Ming C and Yin, Zhaozheng. "Real-time multi-modal human–robot collaboration using gestures and speech." *Journal of Manufacturing Science and Engineering* Vol. 144 No. 10 (2022): p. 101007. URL https://doi.org/10.1115/1.4054297.

[7] Hao, Weixing, Parasch, Andrew, Williams, Stephen, Li, Jiayu, Ma, Hongyan, Burken, Joel and Wang, Yang. "Filtration performances of non-medical materials as candidates for manufacturing facemasks and respirators." *International journal of hygiene and environmental health* Vol. 229 (2020): p. 113582. URL https://doi.org/10.1016/j.ijheh.2020.113582.

[8] Briassouli, Alexia and Ahuja, Narendra. "Extraction and analysis of multiple periodic motions in video sequences." *IEEE transactions on pattern analysis and machine intelligence* Vol. 29 No. 7 (2007): pp. 1244–1261. DOI 10.1109/TPAMI.2007.1042.

[9] Hao, Weixing, Xu, Guang and Wang, Yang. "Factors influencing the filtration performance of homemade face masks." *Journal of Occupational and Environmental Hygiene* Vol. 18 No. 3 (2021): pp. 128–138. DOI 10.1080/15459624.2020.1868482.

[10] Chen, Haodong, Zendehdel, Niloofar, Leu, Ming C and Yin, Zhaozheng. "Real-time human-computer interaction using eye gazes." *Manufacturing Letters* Vol. 35 (2023): pp. 883–894. URL https://doi.org/10.1016/j.mfglet.2023.07.024.

[11] Chen, Haodong, Leu, Ming C, Tao, Wenjin and Yin, Zhaozheng. "Design of a real-time human-robot collaboration system using dynamic gestures." *ASME International Mechanical Engineering Congress and Exposition*, Vol. 84492: p. V02BT02A051. 2020. American Society of Mechanical Engineers. URL https://doi.org/10.1115/IMECE2020-23650.

[12] Chen, Haodong, Tao, Wenjin, Leu, Ming C and Yin, Zhaozheng. "Dynamic gesture design and recognition for human-robot collaboration with convolutional neural networks." *International Symposium on Flexible Automation*, Vol. 83617: p. V001T09A001. 2020. American Society of Mechanical Engineers. URL https://doi.org/10.1115/ISFA2020-9609.

[13] Zhao, Ping, Zhang, Yating, Guan, Haiwei, Deng, Xueting and Chen, Haodong. "Design of a single-degree-of-freedom immersive rehabilitation device for clustered upper-limb motion." *Journal of Mechanisms and Robotics* Vol. 13 No. 3 (2021): p. 031006. URL https://doi.org/10.1115/1.4050150.

[14] Tao, Wenjin, Chen, Haodong, Moniruzzaman, Md, Leu, Ming C, Yi, Zhaozheng and Qin, Ruwen. "Attention-based sensor fusion for human activity recognition using IMU signals." *arXiv preprint arXiv:2112.11224* (2021)URL https://doi.org/10.48550/arXiv.2112.11224.

[15] Dwibedi, Debidatta, Aytar, Yusuf, Tompson, Jonathan, Sermanet, Pierre and Zisserman, Andrew. "Counting out time: Class agnostic video repetition counting in the wild." *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*: pp. 10387–10396. 2020. URL https://doi.org/10.48550/arXiv.2006.15418.

[16] Zhang, Huaidong, Xu, Xuemiao, Han, Guoqiang and He, Shengfeng. "Context-aware and scale-insensitive temporal repetition counting." *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*: pp. 670–678. 2020. URL https://doi.org/10.48550/arXiv.2005.08465.

[17] Yao, Ziyu, Cheng, Xuxin and Zou, Yuexian. "Poserac: Pose saliency transformer for repetitive action counting." *arXiv preprint arXiv:2303.08450* (2023)URL https://doi.org/10.48550/arXiv.2303.08450.

[18] Bazarevsky, Valentin, Grishchenko, Ivan, Raveendran, Karthik, Zhu, Tyler, Zhang, Fan and Grundmann, Matthias. "Blazepose: On-device real-time body pose tracking." *arXiv preprint arXiv:2006.10204* (2020)URL https://doi.org/10.48550/arXiv.2006.10204.

[19] Liu, Ze, Ning, Jia, Cao, Yue, Wei, Yixuan, Zhang, Zheng, Lin, Stephen and Hu, Han. "Video swin transformer." *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*: pp. 3202–3211. 2022. URL https://doi.org/10.48550/arXiv.2106.13230.

[20] Chen, Haodong, Zendehdel, Niloofar, Leu, Ming C and Yin, Zhaozheng. "Fine-grained activity classification in assembly based on multi-visual modalities." *Journal of Intelligent Manufacturing* (2023): pp. 1–19URL https://doi.org/10.1007/s10845-023-02152-x.

[21] Onoro-Rubio, Daniel and López-Sastre, Roberto J. "Towards perspective-free object counting with deep learning." *Computer Vision–ECCV 2016: 14th European Conference, Amsterdam, The Netherlands, October 11–14, 2016, Proceedings, Part VII 14*: pp. 615–629. 2016. Springer. URL https://doi.org/10.1007/978-3-319-46478-7_38.

[22] Sreenu, G and Durai, Saleem. "Intelligent video surveillance: a review through deep learning techniques for crowd analysis." *Journal of Big Data* Vol. 6 No. 1 (2019): pp. 1–27. URL https://doi.org/10.1186/s40537-019-0212-5.

[23] Huang, Yifei, Sugano, Yusuke and Sato, Yoichi. "Improving action segmentation via graph-based temporal reasoning.": pp. 14024–14034. 2020. URL https://api.semanticscholar.org/CorpusID:219615799.