Research on Real-Time Tracking and Standardized Evaluation Algorithms for Swimming Posture Integrating Computer Vision

Li Shi[1]
[1]School of Physical Education,, Guangzhou College of Commerce, Guangzhou 511363, Guangdong, China
Corresponding author: Li Shi (e-mail: 13826698077@163.com)

ABSTRACT In order to overcome the problem of high-precision, low-latency capture and quantitative assessment of human posture in swimming in dynamic underwater conditions, the paper suggests a computer vision-based framework of real-time tracking and standardized assessment of swimming postures. To focus on the identification of key points on the torso and limbs of swimmers, first, an underwater target detection network that relies on the multi-scale feature extraction and attention mechanisms will be built. Second, a spatiotemporal context fusion and 3D motion reconstruction module is developed that integrates temporal smoothing methods and camera projection geometry constraints to recreate the 3D skeleton sequence of swimming motions. Lastly, a universal scoring system founded on kinematic parameters and motion template matching is developed to determine the quality of motion quantitatively based on the joint angles, limb symmetry and stroke frequency. The experiments are confirmed using a self-constructed underwater swimming video dataset that contains the samples of four strokes of swimming in varying conditions of light and water quality. Findings indicate that the developed technique is more efficient than other comparative techniques in key point detection performance, processing speed in real-time, and consistency in evaluation. This approach is effective in providing technical assistance in training and injury prevention in swimming that is intelligent. The pose-estimation framework can be integrated with wearable motion sensors for underwater or poolside training monitoring.

INDEX TERMS swimming posture tracking,computer vision,3D motion reconstruction,standardized evaluation,wearable motion sensing

I. INTRODUCTION

BEING a cyclical sport requiring extremely high technical accuracy, swimming directly predetermines the level of athletic activity and the danger of sports injuries by the quality of movements. The conventional method of swimming training depends mainly on the ability of the coach to see and the experience of the coach. This method is not only extremely subjective and incapable of measuring delicate variations in high-velocity underwater motion, but also incapable of giving a quantifiable kinematic parameter, thereby restricting scientific training. Over the past few years, as the computer vision and deep learning technologies advanced at a breakneck pace, the label-free human posture tracking video-based technique of motion analysis has found its way into a researched hotspot within the domain of motion analysis. Nevertheless, the swimming setting poses exceptional challenges, including complicated backgrounds, extreme changes of light and harsh water refraction and occlusion, which lead to the general-purpose human posture estimation algorithms to suffer a substantial loss of accuracy in the water conditions. Thus, the theoretical importance and practical opportunities of research of real-time posture tracking and standardized evaluation algorithms that are specifically oriented at the swimming case are of high importance.

The first paper is focused on overcoming three fundamental problems in the underwater dynamic environment, including stable recognition of the main human points, precise re-creation of the 3D dynamic motion, and objective motion quality quantification. The principal contributions of the paper compared to the previous work are as follows: First, it designs an end-to-end real time tracking and evaluation system of swimming postures, which organically combines the functions of target detection, 3D reconstruction and scoring modules; Second, it presents a strong key point detection network of underwater optical distortion and The proposed approach demonstrates better performance in a variety of evaluation measures, which has been validated through extensive experimentation, which is a viable technical direction in the intelligent transformation of swimming training.

II. RELATED WORK

The key technical challenge of human motion analysis lies in how to achieve high-precision, low-latency, and non-invasive three-dimensional motion capture and dynamic parameter extraction in an uncontrolled environment. Mettes et al. [1] reviewed the current development status, core methods, application areas and future challenges of hyperbolic deep learning in computer vision. Ait-Bennacer et al. [2] used deep learning pose estimation algorithms to extract two-dimensional skeletal key points from each viewpoint, and then reconstructed the three-dimensional human skeleton through triangulation and multi-view fusion. Lugrís et al. [3] used labelless optical motion capture technology to obtain the motion trajectory of the human body surface. Brunner et al. [4] evaluated the accuracy of a commercial labelless motion capture system in ergonomic risk analysis in typical occupational tasks. Nail-Ulloa et al. [5] established a lumbar fatigue failure assessment framework based on inertial motion capture and demonstrated its application through case studies.

Lapresa et al. [6] used a high-precision industrial robot as a motion reference to achieve rigorous verification of the accuracy of MIMU upper limb motion analysis. Lau et al. [7] developed a vision-based integrated method for human motion analysis and fall detection, distinguishing different types of falls (forward, backward, and lateral) and identifying risky actions that lead to falls. Bräunig et al. [8] achieved continuous, non-contact tracking of human motion by simultaneously achieving high-resolution imaging and velocity measurement of each pixel in a single scan. Bini et al. [9] showed that the joint angle measurement error of label-free motion analysis in a controlled environment is close to the clinically acceptable range, but the accuracy decreases significantly in outdoor, competition scenarios, or when wearing loose sportswear. Geisen et al. [10] developed a pure geometric motion parameter extraction method that does not require training data, while evaluating its accuracy and real-time potential. Current research focuses on two-dimensional posture detection or three-dimensional reconstruction in laboratory environments, lacking an end-to-end tracking and evaluation framework designed specifically for swimming, integrating underwater optical distortion compensation and real-time requirements, and has not yet established a standardized, quantifiable motion scoring system for different swimming strokes. This paper proposes a real-time tracking and standardized evaluation algorithm for swimming posture that integrates computer vision, focusing on solving three core problems in underwater dynamic environments: robust key point detection, label-free 3D motion reconstruction, and automatic scoring based on motion biomechanics.

III. METHODS

A. UNDERWATER TARGET DETECTION AND KEY POINT LOCALIZATION MODULE

The first step in achieving swimming posture tracking is to stably detect the swimmer's body and locate the two-dimensional coordinates of key joints in underwater video frames. Underwater environments have problems such as light absorption and scattering, low contrast, bubble interference, and deformation caused by water refraction, which severely degrade the performance of general target detection and key point estimation algorithms when applied directly. To address these challenges, a deep learning network with multi-scale feature fusion and attention enhancement was constructed. The network takes a single frame of underwater image as input and outputs a confidence heatmap of twenty-one human key points, including the head, neck, shoulder center, left and right shoulders, left and right elbows, left and right wrists, hip center, left and right hips, left and right knees, left and right ankles, and six distal points for hands and feet to ensure consistency throughout the tracking and reconstruction stages [11,12].

The backbone of the network employs an improved residual network structure, embedding a channel attention mechanism within the standard residual blocks. In particular, the output of every convolutional layer as a feature map is first reduced to a channel description vector by means of global average pooling. Then, two connected layers, with full connections, learn the coefficient of weight in each channel. Lastly, the weights are re-combined in the original feature map in an adaptive manner to boost the response to discriminative features of swimming movements and to reduce invalid channels introduced by underwater noise. Since swimmers have a large variation in posture changes in the water with limbs that may be far apart as compared to the torso a single-scale convolutional kernel cannot capture both local and global structure simultaneously. Thus, the multi-scale features are added at various layers of the backbone network which represent high-resolution texture details at shallow level and semantic details at deep level. These multi-scale representations are merged by using upsampling and element-wise addition to form a composite feature representation with a combination of localization accuracy and semantic robustness.

The keypoint head then applies transposed convolution to progressively recover the feature map resolution back to a quarter of the size of the input image and each resolution level produces a keypoint heatmap at that scale. To manage the non-rigid deformation due to underwater optical refraction, the data augmentation strategy of random elastic deformation is presented in the network training step. The network is trained to learn deformation-invariant features representations by using a random displacement field on the image to induce refraction distortion. Mean square error is utilized in the loss function to determine the difference between the predicted heatmap and the Gaussian distribution around the ground keypoint. Meanwhile, interjoint connectivity constraints are also proposed, i.e. the poses of the adjacent joints predicted should have a reasonable range of limb length in space. Additional penalty terms will be applied to predictions that are out of the reasonable range. The training dataset of 12,000 images was manually annotated by a team of three researchers specialized in sports biomechanics using the LabelMe software. To ensure data quality, a cross-verification protocol was implemented where 10% of the images were randomly selected for secondary review by a senior expert; any annotation with a pixel deviation greater than 3% was discarded and relabeled [13,14].

B. SPATIOTEMPORAL CONTEXT FUSION AND 3D MOTION RECONSTRUCTION MODULE

Two-dimensional keypoint detection provides the projected coordinates of the swimmer in each frame, but monocular video cannot directly obtain depth information, and due to occlusion and detection noise, the keypoint positions between adjacent frames may exhibit unnatural jitter. To obtain a smooth and physically consistent three-dimensional motion sequence, a spatiotemporal context fusion and 3D motion reconstruction module was designed. This module first uses temporal information from multiple consecutive frames to filter the two-dimensional detection results of the current frame, and then combines the camera imaging model with prior knowledge of human skeleton kinematics to elevate the two-dimensional keypoints to three-dimensional space.

The temporal processing employs an adaptive Kalman filter framework. The state variables are the two-dimensional coordinates and velocity of each keypoint to build a linear motion model. In contrast to the regular Kalman filter the covariance matrix of observation noise is dynamic and depends on the maximum confidence of the detection heatmap in the current frame. Once the heatmap peaks are clear and the response is strong, the detection result is deemed valid, a smaller observational noise variance is assigned; otherwise, it is due to ambiguity in the detection, and a larger observational noise variance is assigned which causes the filter to be more reliant on the prediction outcomes of the motion model. In such a manner, the high-speed response capability of the motion can be kept, but, at the same time, isolated jumps (due to a sharp change in the underwater illumination or obscuration) can be successfully suppressed.

Spatial dimension of 3D reconstruction depends on the parameters of calibration of the camera and the inflexible properties of the human skeleton. The underwater camera that records the swimming videos is first calibrated to get its intrinsic parameter matrix, distortion coefficients and extrinsic parameters in the pool coordinate system. There is a difference between underwater and air calibration because of the light refraction at the water-air interface. It involves using a underwater calibration board which is specialized, and it is captured at the real depth of the water. The same camera parameters are obtained through solving the projection equation following refraction compensation. Given the coordinates of the 2D keypoints per frame, the 3D reconstruction problem can be reduced to an optimization over a set of 3D points that lie on the camera ray and which satisfy the skeleton length constraint. To be more precise, the root node is the hip center, which is presumed to have a reasonable depth relative to the bottom of the pool. This procedure is repeated by optimizing the depth of the root node by minimizing the reprojection error. Once the root node depth is established, the process recursively goes to the attached limbs. The child nodes find the 3D position, which reduces the projection error, within a predetermined range of limb lengths in the direction of the ray position defined by the root node. In the case of symmetrical limbs a symmetry constraint is added, that is, the distance of the symmetrical joints, between the left and right shoulders and hips, to the axis of the body, are required to be similar. When the quality of detecting information on one side is unsatisfactory, the other side can be used to provide information to be corrected.

To further improve the temporal stability of the 3D reconstruction, the results of the frame-by-frame solution are input into a sliding window smoother. To accomplish this, in each fifteen-frame window, on the assumption that the human motion lies within a uniform acceleration model on the small time window, cubic splines are employed to approximate the 3D path of each joint point to eliminate high-frequency noise elements. The 3D skeleton sequence is then smoothed and back-projected onto the image plane and compared with the original 2D detection results. When the deviation is large enough to exceed a predefined threshold, then it is concluded that there is extreme occlusion or detection error at that point. Here, the original detection coordinates are substituted with the smoothed projected coordinates and the 3D reconstruction is re-run. The end result of this iterative optimization process is the 3D position of each of the twenty-one joint points in the world coordinate system, the corresponding joint angles, angular velocities, and other derived kinematic information.

C. ALGORITHM FOR EXTRACTING AND STANDARDIZING SWIMMING ATTITUDE PARAMETERS

Once the three dimensional skeleton sequence of the swimmer is obtained then there is need to extract biomagnificent kinematic parameters and come up with a mapping relationship between these parameters and the standardization of movement thus attaining a standardized assessment of the swimming posture. Various swimming strokes differ greatly in the movement pattern, and so must be represented using different feature extraction and scoring strategies, yet they are defined by the same underlying skeleton representation. This module initially determines the type of stroke used by the swimmer automatically and subsequently breaks down the stroke into phases based on the movement cycle of the respective stroke and lastly calculates important indicators in each phase and compares them with a template library.

The time-frequency characteristics of the skeleton sequence are used to recognize the swimming stroke. On a continuous 3D skeleton data, the movement path of the center of the shoulder in the horizontal plane, the periodicity of the left and right elbow joint angle, and the tilt angle of the torso towards the forward direction are obtained. These three features are combined and fed to a lightweight classification network composed of two convolutional layers and one fully connected layer and the result is the probability distributions of four swimming strokes. Experiments conducted on a dedicated validation subset of 1,200 video segments (300 per stroke type) indicate that with the skeleton data correct, the accuracy of swimming stroke recognition is 98.4%. Specifically, the recognition rates for breaststroke and backstroke were 99.2% and 98.7% respectively, while a slight confusion was observed between freestyle and butterfly (97.8%) due to similar torso undulating frequencies in high-speed segments. This high accuracy provides a robust basis for further specialized analysis. Upon the completion of the swimming stroke recognition stage the system loads the matching motion template and evaluation rules. Using freestyle as an example, the regularized motion cycle is typically broken down into five stages: entry, catch, pull, exit, and arm recovery. Every stage has an optimal range of joint angles and limb space position. In order to automatically recognize such phases, a Hidden Markov Model (HMM) based on three-dimensional joint angles is employed. The continuous shoulder adduction/abduction angles, elbow flexion/extension angles, and wrist posture are the observation variables used to train a five-state HMM, and each state represents a motion phase. Exercise physiology is used to initialize the state transition probability matrix, i.e. transitions can only happen in a sequential order based on the phase order without omitting it. The new skeleton sequence is decoded using the Viterbi algorithm to get the motion phase label of every frame.

In each movement stage, a series of quantitative evaluation indicators are computed. To illustrate, in the pull phase, the lowest angle of the elbow flexion is noted, this is too small an angle means that the arm is overbending, and is not as efficient in propulsion. The angle of attack of the palm relative to the direction of travel is determined during the catch phase; angles other than the optimal angle of attack cause a certain loss in propulsive force. Moreover, the indicators of symmetry at the whole-body level are obtained, comparing the differences in the duration of the pull phases on the right and left sides, and the differences in the scope of shoulder rotation. Periodic stability indicators are measured by computing the coefficient of variation of the time that three pull cycles take, the coefficient of variation is excessively high, indicating unstable rhythm, which is counterproductive to energy conservation. These indicators are specific values that are compared with the normal reference ranges that were retrieved in a database compiled from 50 provincial-level elite swimmers (Master of Sports level) whose movements were captured in standard competition environments. The normal reference ranges were derived using the 10th and 90th percentiles to define the boundaries of 'standard' technique.

The standardized evaluation algorithm finally gives out an overall score between 0 and 100 and also a breakdown of scores concerning each sub-dimension. The overall score is obtained by weighted summation where the weight coefficients are obtained through an analytic hierarchy process (AHP). Five swimming coaches and two sports biomechanics specialists were invited to form a judgment matrix. The consistency of the pairwise comparison matrices was rigorously tested; the calculated Consistency Ratio (CR) for the primary category matrix was 0.042 ($CR < 0.1$), indicating that the expert judgments were sufficiently consistent and the resulting weights are statistically valid. To be more precise, movement standardization (40% of the weight), including the corresponding degree of joint angles to the template at each step; symmetry (30%), which is the consistency of the movements of the left and the right limbs; rhythmic stability (20%), which is based on how regularly the stroke cycle is; and efficiency (10%), which is a combination of the ratio between the frequency of the stroke and the speed of the The raw scores of each sub-dimension are transformed to the range of 0-100 through the piecewise linear transformation. The mapping function threshold is established using the percentile of data of top athletes; an example is that all elbow angles below the 10th-percentile are zero marks and all above the 90th-percentile are full marks with linear interpolations in between.The final comprehensive score is presented in an intuitive numerical form, and the system also generates a text report indicating specific movement aspects requiring improvement.

IV. RESULTS AND DISCUSSION

A. EVALUATION OF KEY POINT DETECTION ACCURACY AND REAL-TIME PERFORMANCE

A quantitative assessment of the proposed underwater keypoint detection module was carried out on a self-constructed test set to ensure the efficiency of the suggested module. The test set consists of two thousand underwater swimming video clips, each of about five seconds length, with a resolution of 1920x1080 pixels. The underwater camera was positioned at a fixed distance of 5.0 meters from the center of the swimming lane. Based on the camera's intrinsic parameters and the 5-meter object distance, the horizontal imaging scale was determined to be approximately 2.6 mm/pixel. The performance indicators were the mean error in detection (measured as the normalized Euclidean distance between the estimated and ground truth keypoints in millimeters on the image plane) and the percentage of correct keypoints detected. A point was considered correctly detected when the distance between the predicted and ground truth keypoints was less than 20 millimeters (corresponding to approximately 7.7 pixels at the 5-meter reference distance). This paper compares our proposed method with three mainstream pose estimation algorithms: OpenPose, HRNet, and DeepLabCut, a version specifically fine-tuned for underwater scenarios. Table 1 presents the results of the keypoint detection of the various methods on the various test subsets. The table data reveals our proposed method has a mean error of detection at 12.7 mm on average across all four swimming strokes, much lower than OpenPose (34.2 mm) and HRNet (28.5 mm), as well as DeepLabCut (18.3 mm) (after underwater tuning). When it comes to the appropriate detection rate, our approach attains 94.1% whereas the highest amongst the compared methods is 85.6% of DeepLabCut. It is interesting to note that in turbid water conditions, the error of OpenPose increases drastically to 51.3 mm primarily due to the fact that suspended particles in the water interfere with the edge information that it is based on. Conversely, our suggested approach, as the channel attention mechanism and multi-scale feature fusion are introduced, is more resilient to this form of noise, where the error only increases by 10.8 mm in clear settings, and by 16.2 mm in strong background situations, which shows that the proposed random elastic deformation data augmentation approach is more effective in enhancing the network to adapt to extreme lighting scenarios.

In the real-time performance evaluation, the single-frame processing time was tested on a high-performance workstation configured with an Intel Core i9-12900K CPU (3.2GHz), an NVIDIA RTX 4090 GPU (24GB VRAM), and 64GB of DDR5 memory running on Ubuntu 22.04 LTS. This specific hardware configuration was chosen to establish a performance upper bound; the efficiency on embedded or mobile edge devices may vary and remains a subject for future optimization. The processing flow includes the entire process of image reading, keypoint detection, 3D reconstruction, and score calculation. The average single-frame processing time of the proposed method is 28 milliseconds, corresponding to a processing speed of approximately 35.7 frames per second. This includes 18 milliseconds for the keypoint detection network, 7 milliseconds for the 3D reconstruction module, and 3 milliseconds for score calculation. This speed exceeds the requirements for real-time processing and can support online analysis of video streams at 30 frames per second. In comparison, OpenPose requires approximately 120 milliseconds to process one frame on the same hardware, which cannot meet the needs of real-time tracking. While HRNet has higher accuracy than OpenPose, it has a higher computational cost, with a single-frame processing time exceeding 150 milliseconds. DeepLabCut, after lightweight modifications, can achieve approximately 45 milliseconds per frame, but this is still higher than the 28 milliseconds of the proposed method. The efficiency advantage of the method presented in this paper mainly comes from the careful design of the network structure, which controls the number of model parameters and computational complexity while maintaining high accuracy.

B. ANALYSIS OF ACCURACY AND ROBUSTNESS OF 3D MOTION RECONSTRUCTION

The accuracy assessment of 3D motion reconstruction requires ground truth data independent of 2D detection. Since it is difficult to obtain simultaneously labeled motion capture ground truth in a real underwater environment, an indirect verification strategy is adopted. Four cameras with different perspectives are deployed in the pool to simultaneously capture data and reconstruct the same swimming motion from multiple perspectives. For any given frame, the 3D coordinates are obtained through triangulation using the 2D detection results from two perspectives, and these are used as reference ground truths. Then, the predicted coordinates are obtained by combining the 2D detection results from a third perspective with the monocular 3D reconstruction method presented in this paper, and the deviation between the two is calculated. While this multi-view triangulation serves as a practical reference, it is recognized that this approach possesses a ‘self-referential’ limitation; systematic biases in the 2D detector across different views could propagate into the reference ground truth. To mitigate this, manual correction of the 2D coordinates was performed for the reference frames to ensure the ‘ground truth’ reflects absolute physical positions as closely as possible. Future studies will involve synchronized integration with optoelectronic motion capture systems to provide independent validation. Table 2 lists the average 3D reconstruction errors of major joints in the four swimming strokes. Joints in the center of the torso, such as the hip and shoulder centers, have the smallest reconstruction errors, at 11.2 mm and 12.4 mm respectively. This is because these joints are relatively stable during swimming and are not easily obscured. The errors in the distal joints are relatively large, with the wrist joint at 18.7 mm and the ankle joint at 19.35 mm. This is related to the high speed of hand and foot movements underwater, the susceptibility to motion blur, and the difficulty of detection. The left knee joint error in breaststroke reaches 20.1 mm. Analysis suggests that the extreme knee flexion and rapid eversion and adduction movements during the breaststroke kick result in drastic changes in the knee region’s shape in the image, making the detection network prone to ambiguity. The average wrist joint error in butterfly stroke is 18.1 mm because the relatively symmetrical hand trajectories during the synchronized arm movements in butterfly stroke facilitate correction using symmetry constraints.

The impact of different degrees of occlusion on reconstruction error was further analyzed. Frames from the test video where different parts of the swimmer’s body were occluded by their own limbs or water reflections were selected and categorized into three groups based on the proportion of occlusion area: mild occlusion, moderate occlusion, and severe occlusion. Under mild occlusion, the reconstruction error of the proposed method was 14.2 mm, slightly higher than the 12.8 mm under no occlusion; under moderate occlusion, the error rose to 19.7 mm; and under

TABLE 1. Performance comparison of different methods in underwater swimming video key point detection

Method	Average error in freestyle swimming (mm)	Average error in breaststroke (mm)	Average error in backstroke (mm)	Average error in butterfly stroke (mm)	Clear water error (mm)	Turbidity error (mm)	Error due to strong backlighting (mm)	Correct detection rate (%)
OpenPose	32.7	35.1	33.8	35.2	28.4	51.3	36.5	72.3
HRNet	27.3	29.2	27.9	29.5	24.1	38.7	31.2	78.9
DeepLabCut (fine-tuning)	17.5	18.8	17.2	19.6	15.3	23.4	19.8	85.6
This article's method	12.1	13.2	12.5	13.0	10.8	16.2	13.5	94.1

TABLE 2. Average error of 3D reconstruction of various joints under different swimming strokes (unit: mm)

Key points	Freestyle error	Breaststroke error	Backstroke error	Butterfly stroke error	Average error
head	13.5	14.2	13.8	13.1	13.7
shoulder center	12.1	13.0	12.4	11.9	12.4
left shoulder	14.3	15.1	14.6	14.0	14.5
right shoulder	14.5	15.3	14.7	14.2	14.7
left elbow	16.2	17.0	16.5	15.8	16.4
right elbow	16.4	17.2	16.7	16.0	16.6
left wrist	18.5	19.2	18.8	18.0	18.6
right wrist	18.9	19.4	19.0	18.2	18.9
Hip center	10.8	11.8	11.2	10.9	11.2
Left hip	13.9	14.8	14.3	13.7	14.2
Right hip	14.1	15.0	14.5	13.9	14.4
left knee	17.3	20.1	17.6	17.0	18.0
right knee	17.5	20.3	17.8	17.2	18.2
left ankle	19.1	19.8	19.3	18.5	19.2
Right ankle	19.4	20.0	19.6	18.8	19.5
Average of all joints	15.5	16.5	15.7	15.0	15.6

severe occlusion, the error reached 31.5 mm. Even under severe occlusion, the proposed method could still output a reasonable skeleton estimate without completely diverging. This is attributed to the adaptive Kalman filter in the spatiotemporal fusion module. When the detection confidence of a frame is extremely low, the filter mainly relies on the prediction of the motion model, avoiding drastic jumps. In contrast, frame-by-frame reconstruction methods without temporal smoothing resulted in errors exceeding 60 mm under severe occlusion and often exhibited unreasonable postures with abnormal limb lengths.

C. CONSISTENCY BETWEEN STANDARDIZED EVALUATION RESULTS AND COACH SCORES

To verify the effectiveness of the standardized evaluation algorithm, fifteen amateur swimmers were recruited. Each swimmer completed a 50-meter test in each of four swimming strokes. The tests were recorded on video and scored by both the algorithm presented in this paper and three experienced swimming coaches. Coaches independently reviewed the videos and provided a composite score on a 100-point scale; the average of these scores was taken as the Expert Reference Value (ERV). It should be noted that while ERV serves as a benchmark, it reflects expert-perceived quality rather than absolute physical ground truth. Thus, the results primarily demonstrate the algorithm's high degree of agreement with expert human judgment, facilitating 'expert-like' automated coaching..

Table 3 summarizes the correlation analysis results between the algorithm scores and the average coach scores. For freestyle, the correlation coefficient reached 0.94, indicating a high degree of consistency between the algorithm scores and human scores. The correlation coefficient for breaststroke was 0.91, slightly lower than that for freestyle. This is because the leg movements in breaststroke are more complex than those in freestyle, and the angle changes of the knee and ankle joints have greater individual variability. The current algorithm's template library is mainly based on data from elite athletes, and its inclusiveness of some unconventional movement patterns of amateur swimmers is insufficient. The correlation coefficient for backstroke was 0.92, and for butterfly it was 0.90. The overall correlation coefficient for all swimming strokes was 0.93, and the within-group correlation coefficient was 0.91, achieving an excellent level of interevaluator consistency. Further analysis of samples with large scoring deviations revealed that these samples often had some extreme characteristics, such as swimmers with extremely unstable rhythms (stretch cycle variation exceeding 30%), or obviously asymmetrical movements with a difference of more than 20 degrees in the rotation amplitude of the left and right shoulders. In these extreme cases, coaches may give lower scores than the algorithm because, in addition to observing specific kinematic deviations, coaches will also make a comprehensive evaluation based on the athlete's overall coordination and fluidity, which are high-level features that have not yet been fully modeled in the current algorithm.

In addition to the correlation of the overall scores, the consistency of scores across different sub-dimensions was also analyzed. The correlation coefficient was 0.89 for movement standardization, 0.94 for symmetry, 0.88 for rhythm stability, and 0.85 for efficiency. The correlation

TABLE 3. Correlation analysis results between algorithm scores and coach scores in this paper.

Swimming strokes	Sample size	Algorithm average score	Coaches' average score	Pearson correlation coefficient	Within-group correlation coefficient
Freestyle	15	78.4	79.1	0.94	0.92
breaststroke	15	72.6	74.2	0.91	0.89
Backstroke	15	76.2	77.0	0.92	0.90
Butterfly stroke	15	70.3	71.5	0.90	0.88
all	60	74.4	75.5	0.93	0.91

coefficient for efficiency was relatively low because efficiency involves the relationship between stroke frequency and swimming speed, and accurate speed measurement relies on additional speed sensors or precise timing information. This paper only estimated speed through inter-frame displacement in the video, which introduces some error. Future work could consider fusing inertial sensor or radar velocity information to improve the accuracy of efficiency evaluation.

V. CONCLUSION

This paper addresses the challenges of posture tracking and evaluation in swimming by proposing a real-time algorithm framework integrating computer vision, comprising three core modules: underwater keypoint detection, 3D motion reconstruction, and standardized evaluation. By designing a multi-scale attention network and a spatiotemporal smoothing strategy, the algorithm effectively overcomes optical interference and occlusion issues in the underwater environment, achieving stable extraction of the swimmer's 3D skeleton sequence. Based on this, a scoring model based on kinematic parameter and motion template matching is established, outputting a quantitative motion quality evaluation, thus validating the method's effectiveness and real-time performance. Limitations of this study include the fact that the test environment is a standard indoor swimming pool, and its performance in outdoor natural water areas needs further validation. Additionally, the algorithm currently focuses primarily on upper limb and trunk movements, with room for improvement in the detailed analysis of leg kicking movements. Future work will extend to open water scenarios and explore the integration of multimodal sensor information to further improve the robustness of motion capture and the comprehensiveness of evaluation.

AUTHOR CONTRIBUTIONS

Li Shi designed, collected and analyzed the data, and drafted the manuscript. Li Shi conducted the study, critically revised the manuscript for important intellectual content, and gave final approval of the version to be published. Li Shi participated fully in the work, take public responsibility for appropriate portions of the content, and agreed to be accountable for all aspects of the work in ensuring that questions related to the accuracy or integrity of any part of the work are appropriately investigated and resolved.

CONFLICTS OF INTEREST

The author declares no conflict of interest.

FUNDING

This research received no external funding.

ACKNOWLEDGEMENTS

Not applicable.

REFERENCES

[1] P. Mettes, M. Ghadimi Atigh, M. Keller-Ressel, et al., "Hyperbolic deep learning in computer vision: A survey," International Journal of Computer Vision, vol. 132, no. 9, pp. 3484-3508, 2024, doi: 10.1007/s11263-024-02043-5.

[2] F. E. Ait-Bennacer, A. Aaroud, K. Akodadi, et al., "Applying Deep Learning and Computer Vision Techniques for an e-Sport and Smart Coaching System Using a Multiview Dataset: Case of Shotokan Karate," Int. J. Online Biomed. Eng, vol. 18, no. 12, pp. 35-53, 2022, doi: 10.3991/ijoe.v18i12.30893.

[3] U. Lugrís, M. Pérez-Soto, F. Michaud, et al., "Human motion capture, reconstruction, and musculoskeletal analysis in real time," Multibody System Dynamics, vol. 60, no. 1, pp. 3-25, 2024, doi: 10.1007/s11044-023-09938-0.

[4] O. Brunner, A. Mertens, V. Nitsch, et al., "Accuracy of a markerless motion capture system for postural ergonomic risk assessment in occupational practice," International Journal of Occupational Safety and Ergonomics, vol. 28, no. 3, pp. 1865-1873, 2022, doi: 10.1080/10803548.2021.1954791.

[5] I. Nail-Ulloa, M. Zabala, N. Pool, et al., "A fatigue failure framework for the assessment of highly variable low back loading using inertial motion capture–a case study," Ergonomics, vol. 69, no. 2, pp. 292-308, 2026, doi: 10.1080/00140139.2025.2460695.

[6] M. Lapresa, C. Tamantini, F. S. Di Luzio, et al., "Validation of magneto-inertial measurement units for upper-limb motion analysis through an anthropomorphic robot," IEEE Sensors Journal, vol. 22, no. 17, pp. 16920-16928, 2022, doi: 10.1109/JSEN.2022.3193313.

[7] X. L. Lau, T. Connie, M. K. O. Goh, et al., "Fall detection and motion analysis using visual approaches," International Journal of Technology, vol. 13, no. 6, pp. 1173-1182, 2022, doi: 10.14716/ijtech.v13i6.5840.

[8] J. Bräunig, S. Heinrich, B. Coppers, et al., "A radar-based concept for simultaneous high-resolution imaging and pixel-wise velocity analysis for Tracking human motion," IEEE Journal of Microwaves, vol. 4, no. 4, pp. 639-652, 2024, doi: 10.1109/JMW.2024.3453570.

[9] R. R. Bini, F. A. Moura, P. R. P. Santiago, et al., "Special issue themes: Markerless motion analysis in sport and exercise," Journal of Sports Sciences, vol. 42, no. 1, pp. 1-2, 2024, doi: 10.1080/02640414.2024.2317652.

[10] M. Geisen, F. Seifriz, F. Fasold, et al., "A Novel Approach to Sensor-Based Motion Analysis for Sports: Piloting the Kabsch Algorithm in Volleyball and Handball," IEEE Sensors Journal, vol. 24, no. 21, pp. 35654-35663, 2024, doi: 10.1109/JSEN.2024.3455173.

[11] J. S. Kanwal, B. S. Sanghera, R. Dabbi, et al., "Pose analysis in free-swimming adult zebrafish, Danio rerio:"fishy" origins of movement design," Brain Behavior and Evolution, vol. 100, no. 2, pp. 93-111, 2025, doi: 10.1159/000543081.

[12] A. Pisaniello, "The game changer: How artificial intelligence is transforming sports performance and strategy," Geopolitical, Social Security and Freedom Journal, vol. 7, no. 1, pp. 75-84, 2024, doi: 10.2478/gssfj-2024-0006.

[13] S. Barbon Junior, A. Pinto, J. V. Barroso, et al., "Sport action mining: Dribbling recognition in soccer," Multimedia Tools and Applications, vol. 81, no. 3, pp. -4364, 2022, doi: 10.1007/s11042-021-11784-1.

[14] G. Paulin and M. Ivasic-Kos, "Review and analysis of synthetic dataset generation methods and techniques for application in computer vision," Artificial intelligence review, vol. 56, no. 9, pp. 9221-9265, 2023, doi: 10.1007/s10462-022-10358-3.