Sports Engineering (2023) 26:45
https://doi.org/10.1007/s12283-023-00437-1

ORIGINAL ARTICLE

Time-weighted motion history image for human activity classification in sports

Hideto Komori[1] · Mariko Isogawa[2] · Dan Mikami[1] · Takasuke Nagai[2] · Yoshimitsu Aoki[1]

Accepted: 12 September 2023 / Published online: 24 October 2023

Abstract

Vision-based human activity classification has remarkable potential for various applications in the sports context (e.g., motion analysis for performance enhancement, active sensing for athletes, etc.). Recently, learning-based human activity classifications have been widely researched. However, in sports scenes in which more detailed and player-specific classifications are required, this is a quite challenging task; in many cases, only a limited number of datasets are available, unlike daily movements such as walking or climbing stairs. Therefore, this paper proposes a time-weighted motion history image, an effective image sequence representation for learning-based human activity classification. Unlike conventional MHI based on the assumption that "the newer frame is more important," our method generates importance-aware representation so that the predictor can "see" the frames that contribute to analyzing the specific human activity. Experimental results have shown the superiority of our method.

Keywords Temporal template · Motion history image (MHI) · Human activity classification

1 Introduction

The ability to analyze an athlete's motion is essential for understanding sports scenes. It is crucial to analyze the motion of specific individual athletes since their form often differs significantly from one athlete to another. Thus personalized analysis enables us to identify their weaknesses and provide feedback to improve their performance. Various human activity analysis methods have been proposed [1, 2]. Specifically, vision-based approach that uses video has been preferred because it does not require the athlete to wear a sensor for each match and can be installed in real-world scenarios [3, 4].

Many recent vision-based human activity classification approaches are learning-based [5–7], which require a large amount of training data. Unlike the tasks easily collect a lot of data that classify daily activity such as walking or climbing stairs, data collection for sports activity analysis has a severe issue. The sports context often demands a more nuanced level of prediction. Examples include predicting whether a player is going to kick a ball to the left or right or whether they'll throw a straight or curved ball. Furthermore, these predictions often need to be athlete-specific (because every athlete has their own unique mannerism.) This issue causes a lack of available data due to the difficulty of collecting enough data since it takes a long time and effort from the athletes. Thus, methods in which classification models can be trained with only a small dataset are required.

There have been many efforts to train network models with a small dataset. One well-known data-level approach is data augmentation, which lies at the heart of all successful applications of deep learning [8, 9]. However, due to the high degree of freedom of video sequence data and its individual-specific manner of sports scene data, it is difficult to cover the variants of data space exhaustively

✉ Hideto Komori
hide1998@keio.jp

Mariko Isogawa
mariko.isogawa@ieee.org

Dan Mikami
mikami.dan@ieee.org

Takasuke Nagai
takasuke.nagai@ntt.com

Yoshimitsu Aoki
aoki@elec.keio.ac.jp

[1] Keio University, Hiyoshi, Yokohama, Kanagawa 223-8522, Japan

[2] NTT, 1-1 Hikarino-oka, Yokosuka, Kanagawa 2390847, Japan

using data augmentation. Another possible way is the parameter-level approach, which is known for meta-learning [10–12]. Generally, it consists of two different models (i.e., a teacher model and a student model) for teaching one network how to learn which features are important in the task. The teacher model learns how to encapsulate the parameter space, while the student model learns how to recognize and classify the actual items in the dataset. That is, it still requires a large dataset for a teacher model. However, as far as we know, there is no publicly available large dataset that contains different motions included in the same sports action, e.g., motions with two different pitch types (straight, curve) included in the pitching action.

Unlike the above approaches, this paper explores the potential for a feature extraction approach. Previously, temporal templates [13–16] have attracted great attention due to its use in vision-based scene analysis. If we use temporal templates as input, we can use learning-based approaches with small amount of data because temporal templates are not video but image [17]. In particular, motion history image (MHI), in which the motion information extracted from continuous frames are combined into one grayscale image, is widely used for human activity analysis [18, 19]. There are various features that represent motion, such as optical flow [20, 21], for human activity recognition. However, in our case, with the small amount of data available, we consider that MHI is particularly effective because it can be generated stably for all samples without training. Therefore, MHI, a simple yet effective method, helps learning-based methods to be trained only on a small amount of data.

However, conventional MHI has a limitation that it assumes "the newer frame is more important," which is not always true [22]. Figure 1 explains the case in which this assumption is not supported. Suppose the tennis player in Fig. 1 has the mannerism of keeping an open stance by spreading and bracing his feet just before the cross shot, while he takes a semi-open stance before the straight shot. Also, suppose we want to classify his tennis strokes. In this case, the latter half of motion is almost identical, and the important part that distinguishes the stroke is the former of the motion. Now we can see, even though the beginning of the motion is completely different, the appearance of generated MHI based on the "newer-is-more-important" assumption is almost the same; the information to be contributed to classification is missing.

To overcome this limitation of conventional MHI, we propose a time-weighted motion history image (TW-MHI) that introduces a frame level of importance into MHI. Figure 1 (top) shows the example of TW-MHI. Compared to conventional MHI, the appearance of generated TW-MHI is largely different for a straight and cross shot, retaining the mannerisms of the tennis player. We also describe a way of generating the temporal importance function that reflects a change of frame level of importance into TW-MHI. TW-MHI is effective for the activity classification of specific athletes since it focuses on the mannerisms that appear in the form. At the same time, it keeps the advantages of MHI that enable us to train our method on a small dataset. To summarize, our contributions are as follows: (1) We propose TW-MHI, a novel feature representation that introduces a frame level of importance into MHI. (2) We also propose a way of generating a temporal importance function that

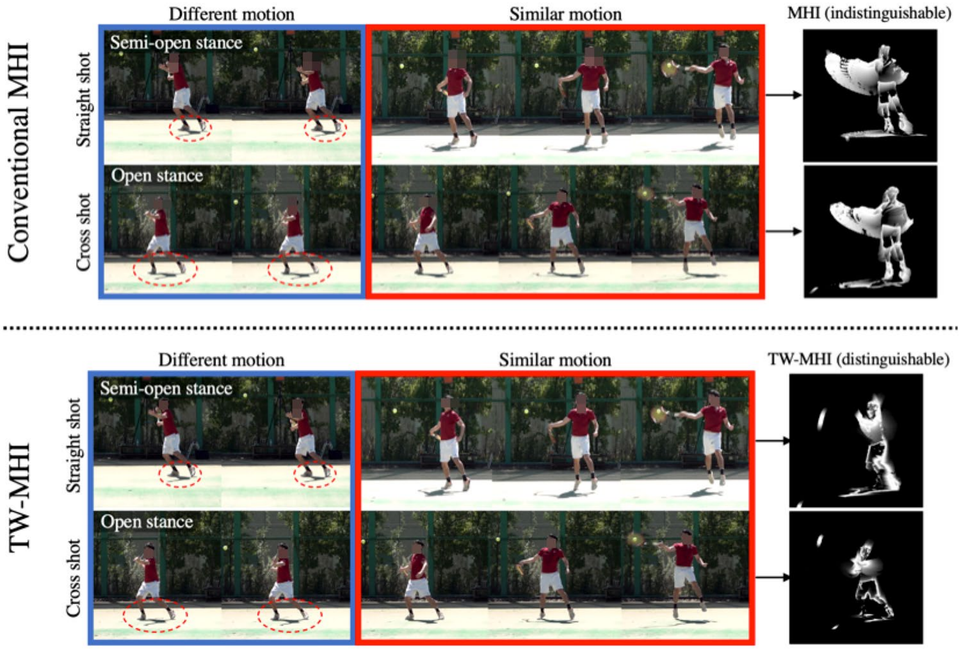

Fig. 1 An example of conventional MHI (top) lacks clues for the classification of two different shots due to the "newer-is-important" assumption. The player has the mannerism of keeping an semi-open stance just before the straight shot, while he takes a open stance before the cross shot. Even though the player's motions are different at the beginning of the sequences, the generated conventional MHIs have almost no differences. On the other side, TW-MHI (bottom) can represent the difference between a straight and a cross shot because it can emphasize important frames, such as the beginning of shot motion

defines the frame level of importance for TW-MHI. (3) We provide extensive experimentation and show that our method outperforms other baseline methods.

2 Method

This paper proposes TW-MHI, an enhanced MHI with the frame level of importance. We first briefly introduce a conventional MHI in Sect. 2.1 and next describe our proposed TW-MHI in Sect. 2.2. Finally, we describe a method of designing a temporal importance function that defines the frame level of importance for TW-MHI in Sect. 2.3.

2.1 Conventional MHI

Conventional MHI [13] combines spatio-temporal motion information that exists in several continuous frames into one static image template. The pixel intensity of MHI represents a function of the motion history at that coordinate, where brighter values correspond to a more recent motion. That is, the silhouette sequence of input video frames is condensed into one grayscale image, where dominant motion information is preserved so that we can see the motion flow. Conventional MHI $\hat{H}_\tau$ is computed as follows:

$$\hat{H}_\tau(x, y, t) = \begin{cases} \tau & \text{if} \quad D(x, y, t) = 1 \\ \max(0, \hat{H}_\tau(x, y, t-1) - \delta) & \text{otherwise.} \end{cases} \tag{1}$$

Here, (x, y) and t show the position and time, and $D(x, y, t)$ represents the object's presence (or motion) in the current frame. The duration τ determines the temporal extent of the movement and δ is the decay parameter. Figure 2 shows an example of conventional MHI.

Since MHI compactly represents motion sequences and it is less susceptible to image noise (i.e., holes, shadows), it helps effective training even though the dataset is small. However, it has a limitation that it assumes "the newer frame is more important," which is not always true. In the following section, we propose an enhanced MHI to overcome this limitation.

2.2 Proposed method: time-weighted motion history image (TW-MHI)

Our goal is to develop a TW-MHI capable of reflecting the frame level of importance to MHI. To this end, TW-MHI introduces a weighted parameter as a function that determines the importance of each frame. We refer to this function as the "temporal importance function." TW-MHI denoted as H is generated as follows:

$$H(x, y, t) = \underset{t=0\cdots\tau}{max}(M(x, y, t)) \tag{2}$$

where (x, y) and t show the position and time. Being the same as in the conventional MHI, the duration τ determines the temporal extent of the movement. $M(x, y, t)$ denotes the weighted $D(x, y, t)$ that represents the object's presence (or motion) in the current frame as below:

$$M(x, y, t) = D(x, y, t) \cdot k(t). \tag{3}$$

Here, $k(t)$ represents the temporal importance function.

By designing $k(t)$ as we describe in the following subsection, TW-MHI can be dedicated to each classification task so that it effectively works. Figure 3 shows an example of TW-MHIs corresponded to two different temporal importance functions. Unlike the conventional MHI that linearly preserves most recent motions, TW-MHI represents weighted motions as designed.

2.3 Temporal importance function

This section describes two different ways of designing temporal importance functions. First, we show the way of heuristically determining the temporal importance function. Then, we show the way of the method for auto-generated temporal importance function.

2.3.1 Heuristically determined temporal importance function

An advantage of our approach is its easiness to reflect the heuristic knowledge of human experts, which is an important characteristic in the context of learning with a small data of individuals.

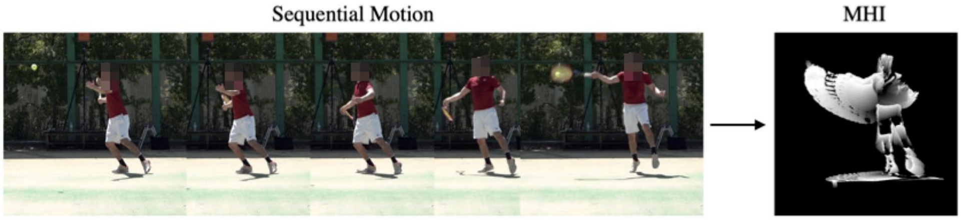

Fig. 2 An example of conventional MHI generated from tennis shot sequence, where brighter values correspond to a more recent motion (same image sequence as the straight shot in Fig. 1)

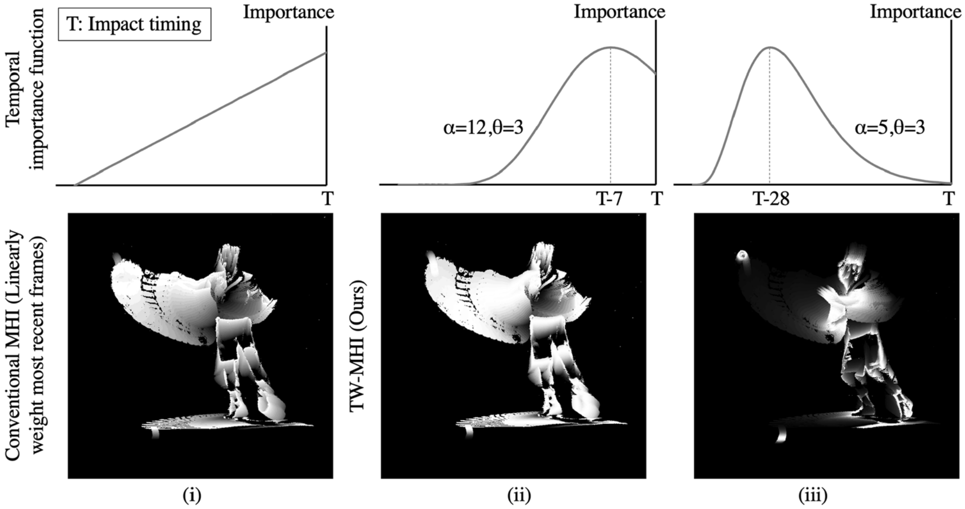

Fig. 3 Generated TW-MHIs based on various temporal importance functions. (same image sequence as the straight shot in Fig. 1.)

To take advantage of this property, we introduce a gamma distribution as a flexible fitting model to formulate temporal importance function $k(t)$ as below:

$$k(t) = t^{\alpha-1}\frac{e^{-t/\theta}}{\Gamma(\alpha)\theta^{\alpha}}, \tag{4}$$

where $\Gamma(\cdot)$ is the gamma function. α and θ are the hyperparameters to determine the shape and scale of the distribution. These parameters can be set to represent the best fit of the human expert's knowledge. The specific parameter settings for our task will be described in Sect. 3.1.2.

2.3.2 Auto-determined temporal importance function

The knowledge of human experts is not always available. Also, the heuristic settings typically take a great deal of users' working time and require considerable input. To reduce the human labor required, we also propose a method for temporal importance function generation so that the function can be automatically determined.

Based on the assumption that "if the classification model trained with the MHI that has the specific weighted peak frame is more accurate, that frame is more important to solve the task," our temporal importance function generation flow consists of the following five steps:

1. Generate initial temporal importance functions $k(t)$ for every step of d frames based on Eq. 4.
2. Generate TW-MHIs corresponding to each $k(t)$.
3. Train human activity classification models with each generated TW-MHIs.
4. Compute the estimation accuracy $A(t)$ for each network models.
5. Generate a temporal importance function from a graph plotting $A(t)$. The values between steps are interpolated by a linear function.

In this paper, we used VGG16 as the human activity classification models in step 3 and 4, and used a fivefold cross-validation scheme to train and test VGG16. We show detailed description of classification in *Network and training* part in Sect. 3.1.2. Figure 4 illustrates this flow. The TW-MHI generated by this auto-generated process is shown on the right in Fig. 4. TW-MHI requires multiple times of $A(t)$ computations while conventional MHI requires training once. However, note that it is only during the initial step where $A(t)$ computation is required, and the computational cost of the TW-MHI for the prediction step is the same as that of the conventional MHI. Furthermore, the prediction accuracy is greatly improved compared to MHI.

3 Experiments and results

We conducted experiments to verify the effectiveness of the proposed TW-MHI. In the experiment, we solve prediction of the shot direction in tennis as a classification task of straight or cross stroke, given a video up to the moment the racket hit the ball as input. We first describe the dataset used in the experiment and then the implementation details and results.

3.1 Experimental settings

3.1.1 Dataset

As far as we know, datasets that include specific movements of specific athletes are not publicly available. Therefore, we prepared the original tennis and baseball dataset

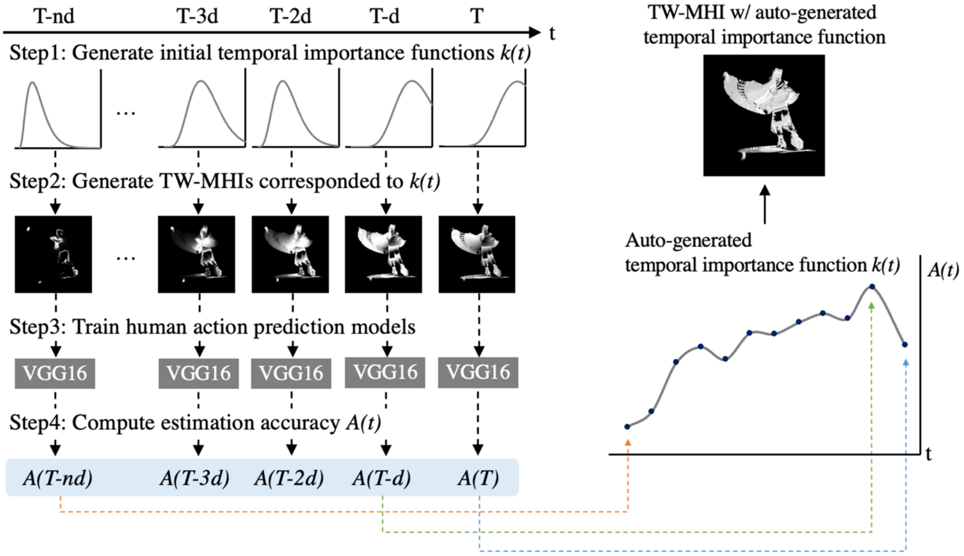
T-nd
T-3d
T-2d
T-d
T
t
Step1: Generate initial temporal importance functions k(t)
…
Step2: Generate TW-MHIs corresponded to k(t)
…
Step3: Train human action prediction models
VGG16
VGG16
VGG16
VGG16
VGG16
Step4: Compute estimation accuracy A(t)
A(T-nd)
A(T-3d)
A(T-2d)
A(T-d)
A(T)
TW-MHI w/ auto-generated temporal importance function
Auto-generated temporal importance function k(t)
A(t)
t

Fig. 4 The flow to determine the temporal importance function and generate TW-MHI without any heuristic settings. (same image sequence as the straight shot in Fig. 1.)

for the activity classification task with two classes: straight or cross shot in tennis and straight or curve pitches in baseball. In the tennis dataset, four amateur players hit the ball with their forehands assuming stroke play in tennis. In the baseball dataset, one amateur player pitched the ball. All players in the tennis dataset are males aged around 20 with more than 5 years of experience. This is a retrospective study that used image sequences of amateur tennis players as part of their practice. Thus, obtaining prior ethics approval was not possible. However, all of the participants agreed that all of their data can be used in this research. Also, the task is their usual practice, and this experiment does not cause any changes in their behavior, this study is clearly conforming to the principles of the Declaration of Helsinki. In addition, we declare that this study is in line with the ethical guidelines of the Host Institution. A camera fixed in front of the player was used to capture the video. The spatial resolution and the frame rate of the camera were 1920 × 1080 pixels and 120 fps, respectively. We captured 569 (310 straight, 259 cross) video sequences in the tennis dataset, and 150 (75 straight, 75 curve) video sequences in the baseball dataset. Each video was spatially cropped to 800 × 800 pixels so that it included the whole body parts of the player. Further, we temporally trimmed each video from the impact and release timing (i.e., the timing of the racket hitting the ball in tennis and releasing the ball in baseball) to the 40 frames before the impact. We make this dataset publicly available. https://github.com/hide1095/TWMHI_dataset.

3.1.2 Implementation details

Heuristically determined TW-MHI. We conducted preliminary experiment to determine parameters for the temporal importance function model that we described in Eq. 4.

We asked three human participants to perform an activity classification task. The participants were given 20 videos of tennis strokes and were asked to predict whether each stroke was a straight shot or a cross shot. 20 sample videos consists of ten straight shots and ten cross shots. At the same time, the participants were asked to answer the time frames for each video that they thought important to distinguish each motion. Figure 5 shows the results of the histogram of the frames that the participants focused on, while Table 1 shows the results of estimation accuracy, i.e., the rate of correctly answered by participants. Figure 5 describes the timing that determines the peak of the temporal importance function. Note that the prediction accuracy of participant 2 was clearly inferior to that of other participants, so we produced the histogram without participant 2. As shown in

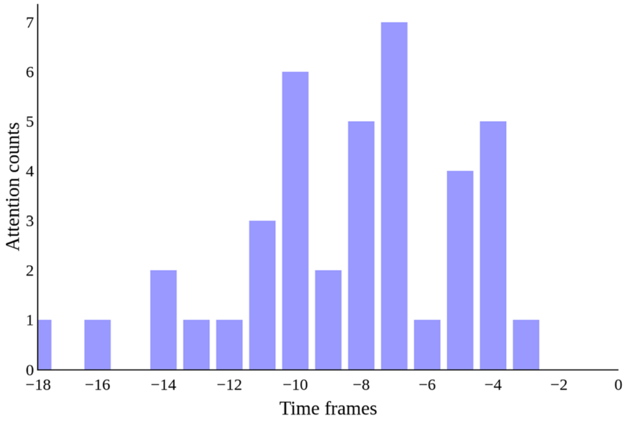

Fig. 5 Histogram of the frames that the participant 1 and 3 focused on. Horizontal axis origin, i.e., 0 is the impact timing of the tennis shot. The minus values represent the former frames prior to the impact

Table 1 Estimation accuracy (the rate of correctly answered) by human participants

	Accuracy (%)		
	Straight	Cross	Total
Participants1	81.8	77.8	80.0
Participants2	54.5	55.6	55.0
Participants3	63.6	77.8	70.0
Mean	66.7	70.3	68.3

The participants were asked to classify sample videos of tennis strokes that include two class motions: straight or cross shot

Fig. 5, it is clear that the moment of impact is not the timing that participants focused on when classifying. This is consistent with the concept of TW-MHI. The timing that the participants paid the most attention to when classifying was seven frames (0.06 s) before the impact of the tennis shot. Based on this result, we set the parameters for Eq. 4 with $\alpha = 12$ and $\theta = 3$ so that our temporal importance function has its peak 0.06 s before the impact.

In the baseball dataset, the player has a mannerism of the upper body falling forward early when pitching straight. Therefore, we set the parameter for Eq. 4 with $\alpha = 8$ and $\theta = 3$ so that the peak of the temporal importance function came at the moment when the upper body started to tilt.

Auto-generated TW-MHI. For the network trained on TW-MHI with automatically determined temporal importance function, we computed temporal importance function following 2.3.

We applied $d = 3$ frames as intervals by fixing $\alpha = 3$ and varying θ from 3 to 13 in Eq. 4. To model the temporal importance function, initial estimation accuracy, i.e., $A(t)$, was used. Therefore, the data used for temporal importance function was the same as VGG16 model that computes $A(t)$. Refer to Section 3.1.1 for the details of the training dataset.

Network and training. To implement our human activity classification method, we used VGG16 [23] as base network architecture. The network is trained on TW-MHI extracted from given video frames. For network training, the Adam [24] optimizer was employed at a learning rate of $1e-4$. All of the training stop at 100 epoch. VGG16 is pre-trained on ImageNet [25]. Following the existing work that succeeded in network model training with human behaviour estimation with a small human dataset [26, 27], we used cross-validation with 80–20 train-test data split for both proposed and baseline methods. These splits share the same random seed in all training and testing. We experimented with all the combinations of train and test data in cross-validation, in this case five combinations which is a fold number. We utilized the mean of five times estimation accuracy as the final accuracy. Also, assuming a real-world scenario to analyze a specific single player, we use a data sequence with the same players for both training and testing. We then compute the mean of estimation accuracy with all individual players for quantitative comparison.

3.1.3 Baseline methods

For quantitative experiment, we compare our methods against the following baselines:

- VGG16+MHI: A method that uses the same base network with our method, i.e., VGG16 [23], trained on conventional MHI [13] to investigate the effect of our TW-MHI.
- Ullah et al. [28]: A method uses long short-term memory (LSTM) trained on extracted feature by VGG16 from RGB images for a comparison against one of the state-of-the-art networks for human activity recognition. VGG16 is pre-trained on ImageNet [25].
- Tran et al. [29]: A method that uses C3D trained on RGB images for a comparison against one of the current state-of-the-art methods for human activity recognition. C3D is pre-trained on Sports1M [30], according to paper [29].
- Anurag et al., [7]: A method that uses video vision transformer (ViViT) trained on RGB images for a comparison between state-of-the-art video classification networks. ViViT is pre-trained on Kinetics400 [31], according to paper [7].

3.2 Qualitative experiment: temporal importance function generation

This subsection qualitatively investigates whether the auto-generated temporal importance function explained in Sect. 2.3 effectively represents the frame importance. Figure 6 shows the input video sequences of both straight and cross shots in tennis. From the direction of the player's racket that faces the direction to be hitted, we humans can distinguish these two shots with the four to ten frames before the impact timing. As shown in Fig. 6, our auto-generated temporal importance function has higher importance weights for those frames, which is qualitatively consistent with human perception.

Figure 7 shows the input video sequences of both straight and curve pitching in baseball. We can distinguish these two pitches to pay attention to 16–22 frames before the release timing because the player has the mannerism of the upper body falling forward early when pitching straight. Figure 7 shows the auto-generated temporal importance function has high importance weights between 16 and 22 frames before the release timing. This result is qualitatively consistent with human perception.

3.3 Quantitative experiment: human activity classification

This subsection investigates the efficacy of our human activity classification trained with the proposed TW-MHI. The results of this experiment are shown in Table 2. As shown in Table 2, both proposed methods (heuristic and auto-generated) achieved higher accuracy than VGG16+MHI for both dataset, indicating that introducing temporal importance function to MHI effectively works for human activity classification tasks.

We also compared our methods against three state-of-the-art human activity classification networks. Estimation accuracies with these three network models were relatively low (almost chance rate) due to the small amount of dataset. In addition, standard deviations of these three networks are larger than these of the proposed method. The results show that even the latest vision approaches cannot learn without

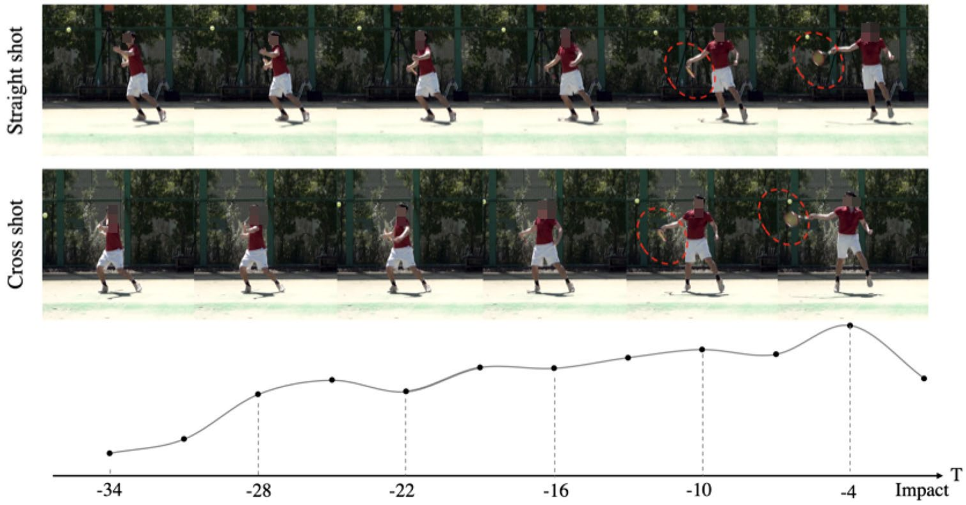

Fig. 6 Auto-generated temporal importance function (bottom) given input video sequences of straight (top) and cross (middle) shots in tennis. The generated temporal importance function precisely has higher importance weights at four to ten frames before the impact timing, which is important to distinguish these two shots (see the red circles)

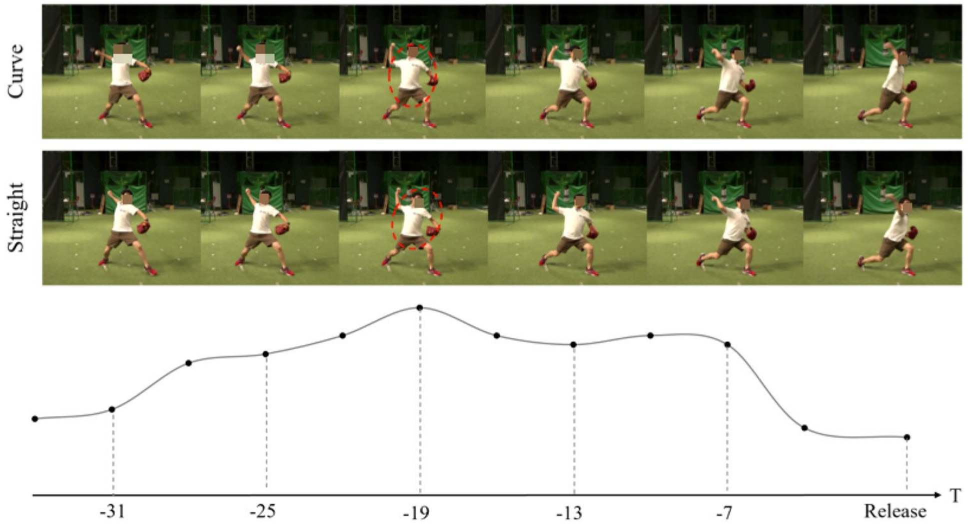

Fig. 7 Auto-generated temporal importance function (bottom) given input video sequences of curve (top) and straight (middle) pitches in baseball. The temporal importance function has higher importance weights at 16–22 frames before the release timing, which is important to distinguish these two pitches (see the red circles)

Table 2 The results of quantitative comparison with other baselines

Method		VGG16 + MHI	Ullah et al. [28]	Tran et al. [29]	Anurag et al. [7]	Ours (w/heuristic TW-MHI)	Ours (w/auto-generated TW-MHI)
Accuracy (%)	Tennis	89.78 ± 1.92	62.70 ± 3.61	62.63 ± 4.84	59.46 ± 3.41	91.37 ± 2.33	90.49 ± 2.95
	Baseball	94.85 ± 1.93	58.00 ± 1.82	59.33 ± 23.14	51.33 ± 7.68	96.67 ± 3.34	97.27 ± 1.75

large amounts of data. In contrast, our methods are successful in classification., which indicates our method's efficacy. In addition to the above, we emphasize that even though the temporal importance function is auto-generated, it still has higher estimation accuracy than other baselines.

Here, we do not intend to assert that our the accuracy of the proposed method outperforms that of humans, comparing Tables 1 and 2. We conducted a t-test for the estimation accuracy between our proposed method and manual classification by a human and found no significant differences between them $(t(2) = 3.17, p = 0.043)$.

4 Conclusion

In this paper, we described TW-MHI, a feature representation that reflects the frame level of importance to MHI. This is achieved by introducing the temporal importance function to design TW-MHI, and the dedicated TW-MHI allows the method to be effectively trained even though only a small dataset is available.

To investigate our method's efficacy, we conducted both qualitative and quantitative experiments using tennis and baseball datasets, and the results showed that our auto-generated temporal importance function is consistent with human perception, and our classification network trained with TW-MHI outperforms other baseline methods.

Supplementary Information The online version contains supplementary material available at https://doi.org/10.1007/s12283-023-00437-1.

Author Contributions HK, DM, and YA were involved in collecting data. HK, DM, and MI were involved in method design. All authors critically revised the report, commented on drafts of the manuscript, and approved the final report.

Funding No funding was received.

Availability of data and materials The datasets generated during and/or analyzed during the current study are not publicly available due to privacy of participants, but are available from the corresponding author on reasonable request.

Code availability Closed source.

Declarations

Conflict of interest The authors declare that they have no conflict of interest.

Ethical statement Prior to the test, the players were informed of the study's purpose, possibility of datasets being made publicly available and the right to withdraw at any time. All players agreed to participate, and the study was conducted in accordance with the Declaration of Helsinki.

Consent to participate All participants have consensus about participating.

Consent for publication All participants have consensus about publication.

References

1. Lloret J, Garcia M, Catala A et al (2016) A group-based wireless body sensors network using energy harvesting for soccer team monitoring. Int J Sens Netw 21(4):208–225. https://doi.org/10.1504/IJSNET.2016.079172
2. Rana M, Mittal V (2021) Wearable sensors for real-time kinematics analysis in sports: a review. In: IEEE Sens J, pp 1187–1207, https://doi.org/10.1109/JSEN.2020.3019016
3. Ceseracciu E, Sawacha Z, Fantozzi S et al (2011) Markerless analysis of front crawl swimming. J Biomech 44(12):2236–2242. https://doi.org/10.1016/j.jbiomech.2011.06.003
4. Ibraheem OW, Irwansyah A, Hagemeyer J, et al (2017) Reconfigurable vision processing system for player tracking in indoor sports. In: Conference on Design and Architectures for Signal and Image Processing, pp 1–6, https://doi.org/10.1109/DASIP.2017.8122114
5. Hochreiter S, Schmidhuber J (1997) Long short-term memory. In: Neural Comput, pp 1735–1780
6. Ji S, Xu W, Yang M, et al (2013) 3d convolutional neural networks for human action recognition. In: IEEE Trans. Pattern Anal. Mach. Intell., pp 221–231, https://doi.org/10.1109/TPAMI.2012.59
7. Anurag A, Dehghani M, Heigold G, et al (2021) Vivit: A video vision transformer. In: Proceedings of the IEEE/CVF International Conference on Computer Vision, pp 6836–6846
8. Zhang H, Cissé M, Dauphin YN, et al (2017) mixup: beyond empirical risk minimization. arXiv: 1710.09412
9. Yun S, Han D, Oh SJ, et al (2019) Cutmix: regularization strategy to train strong classifiers with localizable features. arXiv: 1905.04899
10. Cao K, Ji J, Cao Z, et al (2020) Few-shot video classification via temporal alignment. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition
11. Yang S, Liu L, Xu M (2021) Free lunch for few-shot learning: distribution calibration. In: International Conference on Learning Representations
12. Snell J, Swersky K, Zemel RS (2017) Prototypical networks for few-shot learning. In: Guyon I, von Luxburg U, Bengio S, et al (eds) Advances in neural information processing systems, pp 4077–4087, https://proceedings.neurips.cc/paper/2017/hash/cb8da6767461f2812ae4290eac7cbc42-Abstract.html
13. Davis JW, Bobick AF (1997) The representation and recognition of human movement using temporal templates. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. IEEE Computer Society, Los Alamitos, CA, p 928, https://doi.org/10.1109/CVPR.1997.609439

14. Valstar M, Pantic M, Patras I (2004) Motion history for facial action detection in video. In: IEEE International Conference on Systems, Man and Cybernetics, pp 635–640 vol.1, https://doi.org/10.1109/ICSMC.2004.1398371
15. Tsai DM, Chiu WY, Lee MH (2015) Optical flow-motion history image (of-mhi) for action recognition. In: Signal Image Video Process., pp 1897–1906, https://doi.org/10.1007/s11760-014-0677-9
16. Du Y, Fu Y, Wang L (2015) Skeleton based action recognition with convolutional neural network. In: Asian Conference on Pattern Recognition, pp 579–583, https://doi.org/10.1109/ACPR.2015.7486569
17. Shabanian M, Wenzel M, DeVincenzo JP (2021) Infant brain age classification: 2d cnn outperforms 3d cnn in small dataset. arXiv preprint arXiv:2112.13811
18. Sincan OM, Keles HY (2021) Using motion history images with 3d convolutional networks in isolated sign language recognition. arXiv:org/abs/2110.12396
19. Chun Q, Zhang E (2017) Human action recognition based on improved motion history image and deep convolutional neural networks. In: International Congress on Image and Signal Processing, BioMedical Engineering and Informatics, pp 1–5, https://doi.org/10.1109/CISP-BMEI.2017.8302061
20. Beauchemin SS, Barron JL (1995) The computation of optical flow. ACM computing surveys (CSUR) 27(3):433–466
21. Ilg E, Mayer N, Saikia T, et al (2017) Flownet 2.0: Evolution of optical flow estimation with deep networks. In: Proceedings of the IEEE conference on computer vision and pattern recognition, pp 2462–2470
22. Zhi Y, Tong Z, Wang L, et al (2021) Mgsampler: An explainable sampling strategy for video action recognition. In: Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), pp 1513–1522
23. Simonyan K, Zisserman A (2015) Very deep convolutional networks for large-scale image recognition. In: arXiv, arXiv:org/abs/1409.1556
24. Kingma DP, Ba J (2015) Adam: a method for stochastic optimization. In: International Conference on Learning Representations, arXiv:org/abs/1412.6980
25. Deng J, Dong W, Socher R, et al (2009) Imagenet: A large-scale hierarchical image database. In: 2009 IEEE conference on computer vision and pattern recognition, Ieee, pp 248–255
26. Yuan Y, Kitani K (2019) Ego-pose estimation and forecasting as real-time pd control. In: Proceedings of the IEEE/CVF International Conference on Computer Vision, pp 10,082–10,092
27. Isogawa M, Yuan Y, O'Toole M, et al (2020) Optical non-line-of-sight physics-based 3d human pose estimation. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp 7013–7022
28. Ullah A, Ahmad J, Muhammad K et al (2018) Action recognition in video sequences using deep bi-directional lstm with cnn features. IEEE Access 6:1155–1166. https://doi.org/10.1109/ACCESS.2017.2778011
29. Tran D, Bourdev L, Fergus R, et al (2015) Learning spatiotemporal features with 3d convolutional networks. In: International Conference on Computer Vision, pp 4489–4497, https://doi.org/10.1109/ICCV.2015.510
30. Karpathy A, Toderici G, Shetty S, et al (2014) Large-scale video classification with convolutional neural networks. In: CVPR
31. Kay W, Carreira J, Simonyan K, et al (2017) The kinetics human action video dataset. arXiv preprint arXiv:1705.06950

Publisher's Note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.