Basketball Shooting Performance Analysis Using Multi-modal Wearable and Mobile Sensing in Semi-Naturalistic Settings

Sixuan Wu[1], Alexander Hoelzemann[2], Marius Bock[2], Kristof Van Laerhoven[2],
Thomas Plötz[1], and Alexander T. Adams[1]
[1]School of Interactive Computing, Georgia Institute of Technology, Atlanta, USA
[2]Ubiquitous Computing, University of Siegen, Siegen, Germany
{swu469, thomas.ploetz, aadams322}@gatech.edu
{alexander.hoelzemann, marius.bock}@uni-siegen.de, kvl@eti.uni-siegen.de

2024 IEEE 20th International Conference on Body Sensor Networks (BSN) | 979-8-3315-3014-3/24/$31.00 ©2024 IEEE | DOI: 10.1109/BSN63547.2024.10780653

***Abstract*—Wearable devices have become efficient tools for sports performance analysis. Professional systems heavily rely on the high-tech setup, which are expensive and privacy-invasive for amateur players. This paper addresses the gap between advanced professional systems and limited consumer options by proposing a low-cost, privacy-preserving approach for basketball shot detection and outcome prediction. We leverage accelerometer data from wrist-worn smartwatches, combined with audio recordings, to develop a system capable of identifying shot movements and predicting shot outcomes. The shot detection was achieved by a 1D CNN model through accelerometer data and outcome classification was achieved by an audio classification model. We evaluated the system on 6 participants, and the macro F1 score for shot outcome classification in data streams are 81.53% and 78.07% on dominant hand and non-dominant hand, respectively. Our system opens up explorations in other domains, including medical or industrial activity recognition, where similar approaches can be applied.**

***Index Terms*—machine learning, human activity recognition, sports, basketball, performance analysis, multi-modality, IMU**

I. Introduction

Professional and semi-professional athletes, across various sports disciplines, now seek to optimize their performance using wearable devices. Top-tier professional clubs, such as Derby County, Liverpool FC, Manchester City or Borussia Dortmund in soccer [1], and arenas in the National Basketball Association (NBA) [3], have integrated comprehensive activity recognition systems into their performance monitoring methodologies. These systems heavily rely on high-tech equipment, including specialized and costly camera setups, to capture the playing area from multiple angles. While these solutions are effective in professional environments like sports stadiums, they are not feasible for amateur sports enthusiasts. A noticeable gap exists between the advanced capabilities available to professional and semi-professional athletes and the more limited options accessible to casual consumers.

Prior works have explored the multiple use-cases of wearable devices in sports training and performance analysis. Wearable devices were widely used to measure running activity and provide feedback to users [4], [5], [15]. Moreover, Khan et al. [11] utilized hierarchical representations by leveraging wrist-wearable devices to evaluate athletics' performance in cricket, and further developed a system for gymnastics and medical training [10]. LAX-score [9] is a score, calculated through physiological and motion signals, proposed by Jung et al. to quantify the team performance in lacrosse.

Inspired by the publication of Hang-Time HAR dataset [7] and its potential use cases for basketball and sports activity recognition in general, we aimed to further develop the idea of focusing on a specific sport and create a system capable of detecting specific shot movements and predicting whether a shot will hit or miss the basket. Even if computer-vision algorithms could predict the shot outcomes, non-vision-based algorithms could be less invasive on the users privacy and easier to set up [13], [14]. Previously, researchers utilized wrist-wearable devices to distinguish the shooting types [12], and classify players' activities in basketball games [7], [8]. The gap still exists for shot movement segmentation in naturalistic setting and predict shot outcomes from non-vision-based data. Our system leverages accelerometer data recorded by a commodity smartwatch, in combination with audio recordings. This approach is easy to deploy on any court, making it accessible for use in various settings. In conclusion, this paper seeks to explore the extent to which a low-cost and low-effort system like ours can effectively detect a complex sports activity, such as shooting a basketball, including specific cases like free throws, 2-point and 3-point shots, and shooting from a standing or moving posture.

II. Methods

A. Dataset

Data was collected from 6 male participants (age: 25-37, weight: 70-88kg, height: 172-190cm) with previous basketball experiences, including one left-handed (P6) and five right-handed individuals. Among them two were authors from the University of Siegen (P1, P2), whose data was collected on an indoor basketball court, while the data of four other individuals (P3-P6) followed the IRB (Protocol H24098) at the Georgia Institute of Technology, recorded on an outdoor basketball court. P2 was self-identified as semi-professional, and others were amateur players with varied skills level.

Participants were instructed to wear Bangle.js smartwatches on both wrists, which were used to record accelerometer data in three axis. Cameras with embedded microphones were positioned at the basketball court's corner, facing to the basket and participants, to record video and audio data. The data collection included seven sessions: two free throw sessions (FT1, FT2), three-point shots sessions with and without defense (3PTD, 3PT), mid-range shots sessions (MRD, MR) with and without defense, and a session for stop jump shots (SJ), there were at least 15 shot instances in each session for each participants. During the collection process, two of the participants were paired off, with one shooting while the other rebounded and played defense, and roles alternated after each session. The data collection process simulated naturalistic shooting drills. Therefore, the dataset captured movements related to shooting, passing, dribbling, rebounding, and defensive actions. This study mainly focused on shooting movements, and the dataset contained over 3 hours of video and audio data, over 12 hours of accelerometer data, and over 850 shooting instances.

Participants were instructed to perform three jumps at the beginning and end of the data collection for devices synchronization. The accelerometer data was resampled to a frequency of 50 Hz by resample function in scikit-learn library[1]. Following synchronization and resampling, we utilized the ELAN tool [16] for labeling. We segmented the shooting movements followed NBA rule [2], which was defined as *"the player has started his shooting motion and continues until the shooting motion ceases and he returns to a normal floor position"*. Furthermore, we segmented the audio following each shot to capture instances when the basketball made contact with the rim or net.

B. Accelerometer-based shot detection

We aimed at detecting when the player took a shot by 3-D accelerometer data. We first used sliding window method to segment the accelerometer data. The window size we set was 4 seconds ($4 \times 50 = 200$ data points), and the step size for the sliding window we used was 2 seconds ($2 \times 50 = 100$ data points). Then, we designed and implemented an 1-D CNN binary classification model on temporal dimension to detect whether each window contained shooting segment. The 1-D CNN contained three blocks, where each block included one 1-D convolution layer, one max pooling, one LeakyReLU and one batch normalization layer. The kernel size for each convolution layers were 7, 5, 3, and the the output channels were 8, 16, 32 respectively. The stride and kernel size set to max pooling layers were 2. One adaptive averaging pooling layer and two fully connected layers were used after the convolution blocks. The dimensions used for fully connected layer was 32, 8 and 2. LeakyReLU with 0.2 negative slope was used as the activation function. Considered that the windows containing no shot segments were much more than the windows including shot segments, we used FocalLoss as the criterion during the model training.

C. Audio-based shot outcomes classification

Our objective was to use audio data to distinguish between successful and unsuccessful shots, as the auditory characteristics of a ball passing through the net are distinct from those of it striking the rim. We processed audio segments of each shot using a window size of 2.24 seconds (48000 Hz, 2 channels) as input for a deep learning model. These windows were initially created by cropping from the complete audio file, with the shot audio segments positioned at the center of each window. To increase the dataset size, we adjusted the window inputs' centers to positions 0.25 and 0.75 within each shot audio segment. Additionally, we incorporated audio segments preceding and following the shot audio segments into the dataset to simulate instances without shot audio.

For shot outcomes binary classification, we adapted the end-to-end deep learning model proposed [17], as it demonstrated promising results in cough detection. Our model utilized two sets of 1-D CNN for feature extraction. The kernel sizes for the first set were set as 4, 8, 16, 32, with corresponding strides of 1, 4, 12, and 48. The number of output channels was set to 32. The kernel sizes for the second set were 2, 4, 8, 16, with strides of 48, 12, 4, and 1, respectively, and the output channel was 56. We applied LeakyReLU and batch normalization after each 1-D convolution operation. Then max pooling layers with kernel size and stride 10 were applied after two sets of convolution operations. After concatenating the extracted feature vectors, the feature map was 224×224. ResNet 18 [6] was used as the backbone to process the feature map, and fully connected layer followed by a Dropout layer (dropout rate = 0.2) was adapted to a 2-dimension output. We utilized data augmentation techniques during the training process, including random shifting, padding, and amplification.

D. Shot outcomes classification in data streams

By combining accelerometer data to detect shots and audio data for classifying shot outcomes, we aimed to explore the feasibility of categorizing shot outcomes within continuous data streams. After synchronizing the audio and accelerometer data, we implemented a sliding window technique on the accelerometer data to identify shot instances within each window as mentioned in section II-B. Consistent with our training data, we set the window size to be 4 seconds with a step size of 2 seconds. Whenever a shooting motion was detected within a window, the following audio data stream underwent shot outcome classification using the designated model in section II-C. For this classification, we utilized a sliding window with a window size of 2.24 seconds and a step size of 1.12 seconds. Three audio windows were considered for each detected shot to determine the shot outcome. If at least one of the audio windows indicated a hit outcome within the three windows analyzed, the shot would be classified as hit. Otherwise, the shot would be categorized as miss.

[1] https://scikit-learn.org/stable/modules/generated/sklearn.utils.resample.html

TABLE I: F1 score of LOPO shot detection results

Participants	**Dominant Hand**		**Non-dominant Hand**	
	Shot	Others	Shot	Others
P1	0.9369	0.9816	0.7120	0.9006
P2	0.8487	0.9539	0.3082	0.8817
P3	0.8806	0.9793	0.7164	0.9462
P4	0.8816	0.9775	0.6897	0.9333
P5	0.8223	0.9758	0.5947	0.9533
P6	0.8330	0.9768	0.0236	0.9397
Overall	0.8672	0.9742	0.5074	0.9258

TABLE II: F1 score of LOSO shot detection results

Sessions	**Dominant Hand**		**Non-dominant Hand**	
	Shot	Others	Shot	Others
FT1	0.9210	0.9804	0.8483	0.9598
FT2	0.8780	0.9723	0.8396	0.9623
3PT	0.8991	0.9785	0.8633	0.9689
3PTD	0.8952	0.9856	0.8497	0.9799
MR	0.9071	0.9722	0.8940	0.9666
MRD	0.9019	0.9878	0.8078	0.9738
SJ	0.9174	0.9918	0.8492	0.9843
Overall	0.9028	0.9812	0.8503	0.9708

III. Results

A. Accelerometer-based shot detection

We assessed our model's performance using two cross-validation methods: leave-one-participant-out (LOPO) and leave-one-session-out (LOSO). In LOPO, the model was trained on data from five participants and tested on the remaining participants to evaluate its generalizability across different individuals. Additionally, we explored personalized approaches to potentially enhance model performance by using LOSO evaluation, where the model was trained on six sessions and validated on a separate session.

In LOPO evaluation (Table I), our model achieved an F1 score of 86.72% on the dominant hand and 50.74% on the non-dominant hand. Conversely, in LOSO evaluation (Table II), the model attained an F1 score of 90.28% on dominant hands and 85.03% on non-dominant hands. We observed that the personalized approach (LOSO) outperformed the LOPO model, particularly on the non-dominant hand. Furthermore, models trained on dominant hands performed better than those trained on non-dominant hands. However, it's worth noting that wearing smartwatches on dominant hands during shooting drills may cause discomfort for participants.

TABLE III: F1 score of shot outcome classification

Sessions	**Indoor**		**Outdoor**		**Both**	
	Hit	Miss	Hit	Miss	Hit	Miss
FT1	0.9245	0.9480	0.7685	0.9247	0.8156	0.9229
FT2	0.8879	0.9491	0.7610	0.9405	0.8231	0.9464
3PT	0.7750	0.9297	0.8217	0.9433	0.7062	0.9201
3PTD	0.8333	0.9767	0.7327	0.9403	0.7254	0.9561
MR	0.8636	0.9542	0.7688	0.9209	0.7571	0.9233
MRD	0.8485	0.9630	0.7296	0.9433	0.7420	0.9467
SJ	0.9565	0.9848	0.7746	0.9576	0.8360	0.9681
Overall	0.8699	0.9579	0.7653	0.9387	0.7722	0.9405

TABLE IV: F1 score of LOSO shot outcome detection in data stream

Sessions	**Dominant Hand**			**Non-dominant Hand**		
	No Shot	Hit	Miss	No Shot	Hit	Miss
FT1	0.9714	0.7511	0.7517	0.9588	0.7459	0.6855
FT2	0.9726	0.7746	0.7962	0.9607	0.7861	0.7311
3PT	0.9764	0.6759	0.7285	0.9691	0.6494	0.7059
3PTD	0.9840	0.5600	0.7956	0.9754	0.5385	0.7326
MR	0.9733	0.6514	0.7376	0.9631	0.6292	0.6267
MRD	0.9860	0.6400	0.7963	0.9731	0.6055	0.7216
SJ	0.9880	0.8070	0.8046	0.9801	0.7018	0.7541
Overall	0.9788	0.6943	0.7729	0.9686	0.6652	0.7082

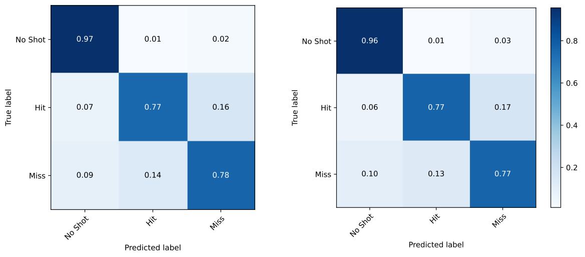

Fig. 1: (left) Normalized confusion matrix evaluated on dominant hand. (right) Normalized confusion matrix evaluated on non-dominant hand.

B. Audio-based shot outcomes classification

We evaluated the model using the LOSO cross-validation method, examining whether the basketball court type influenced the model's performance. The cross-validation was conducted separately on audio data from indoor courts (P1, P2), outdoor courts (P3-P6), and on a combination of both indoor and outdoor court data (P1-P6). The shot outcome classification experiment results were summarized in Table III. For hits, the F1 scores were 86.99%, 76.53%, and 77.22% on indoor, outdoor, and combined court datasets respectively, while for misses, the F1 scores were 95.79%, 93.87%, and 94.05% respectively. It was observed that the model performed better on indoor data compared to outdoor data. This was because of echoes in indoor environments, which could result in clearer recorded audio.

C. Shot outcomes classification in data streams

We investigated whether our algorithm could detect hit and miss shots in the data stream. To accomplish this, we used personalized shot detection models trained using the LOSO method and an audio-based binary shot outcome classification model trained on both indoor and outdoor courts.

The results were summarized in Table IV. For accelerometer data from dominant hands, the F1 scores for hit and miss detection were 69.43% and 77.29%, respectively. Similarly, for non-dominant hands, the F1 scores for hit and miss detection were 66.52% and 70.82%, respectively. The overall macro F1 scores were 81.53% and 78.07% when using the shot detection model trained on dominant hands and non-dominant hands, respectively. The 3PTD session had the lowest hit

F1 score among all sessions. This could be because of the challenging nature of hitting the basket under defensive conditions, resulting in fewer instances of successful hits compared to other sessions. Consequently, the reduced number of hit instances likely contributed to the lower F1 score observed in this particular session. Furthermore, Figure 1 displayed the confusion matrix across all sessions. The visualization revealed that the majority of shot outcome detection was predicted accurately, with no significant difference observed between using dominant and non-dominant hands.

IV. Discussion and Conclusion

Our project aims to improve the analysis of basketball performance in shooting drills by leveraging wearable and mobile devices. While our current approach effectively identifies most shots, there are instances of misclassification. We used accelerometer data to segment shots and audio data to classify outcomes. However, inaccuracies can occur when shot segments are incorrectly identified, leading to inaccurate outcome predictions. Therefore, developing an end-to-end model that incorporates features from both accelerometer and audio data can potentially enhance performance analysis.

Additionally, some shot outcomes are misclassified even when the shot segments are accurately located, particularly in the cases where shots bounce in after hitting the rim. One future work of this project could explore finer granularity such as swishes, shots that bounce in or out, and airballs, rather than only classifying shots as hits or misses. This finer granularity can also provide valuable insights for analyzing performance in basketball shooting drills. Furthermore, the microphones we used were integrated with cameras placed at the corner of the basketball court for labeling and recording. To potentially improve the quality of audio recordings, we could consider placing microphones under the rim.

For practical applications, one limitation is background noise. To address this, de-noising algorithms and data augmentation techniques, such as adding noise to the dataset, can be applied to enhance the models' performance. Moreover, our approach can be easily adapted to an accelerometer-vision system by capturing only rim videos in noisy environments. We also observed that shot detection performance using LOPO for the non-dominant hand was significantly lower compared to LOSO detection. Future work should focus on determining the amount of data needed for accurate shot detection, and developing an application for personalized calibration at home by . Although we evaluated our approach on both indoor and outdoor basketball courts, our data included only male players with basketball experience. Future studies should include a more diverse population.

In conclusion, we present our multi-modal basketball shooting performance analysis system in this paper. To the best of our knowledge, our system is the first non-vision-based basketball shooting outcome prediction system on non-dominant hands. Our system was evaluated on 6 participants on two basketball courts, and achieved 81.53% and 78.07% macro F1 score on dominant hand and non-dominant hand, respectively. Future work should aim to recruit more diverse participants and develop noise-robust, end-to-end models to enhance the granularity of shooting outcome predictions. We believe our multi-modal sensing approach can also be applied to other sports and high-activity clinical applications.

References

[1] How is Data used in the Premier League? — AnalyiSport — analyisport.com. https://analyisport.com/insights/how-is-data-used-in-the-premier-league/. Accessed 09-10-2024.

[2] Shooting movement definition in nba rule. https://official.nba.com/rule-no-4-definitions/. Accessed: 09-10-2024.

[3] Nba and sony's hawk-eye innovations launch strategic partnership powering next generation tracking technology. https://pr.nba.com/nba-sony-hawk-eye-innovations-partnership/, Mar 2023. Accessed 09-10-2024.

[4] L. M. Baker, A. Yawar, D. E. Lieberman, and C. J. Walsh. Predicting overstriding with wearable imus during treadmill and overground running. *Scientific Reports*, 14(1):6347, 2024.

[5] Y. Chen, C.-C. Chen, L.-C. Tang, and W.-H. Chieng. Enhancing running exercise with iot, blockchain, and heart rate adaptive running music. *IEEE Access*, 2024.

[6] K. He, X. Zhang, S. Ren, and J. Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 770–778, 2016.

[7] A. Hoelzemann, J. L. Romero, M. Bock, K. V. Laerhoven, and Q. Lv. Hang-time har: A benchmark dataset for basketball activity recognition using wrist-worn inertial sensors. *Sensors*, 23(13):5879, 2023.

[8] A. Hölzemann and K. Van Laerhoven. Using wrist-worn activity recognition for basketball game analysis. In *Proceedings of the 5th international Workshop on Sensor-based Activity Recognition and Interaction*, pp. 1–6, 2018.

[9] W. Jung, A. Watson, S. Kuehn, E. Korem, K. Koltermann, M. Sun, S. Wang, Z. Liu, and G. Zhou. Lax-score: Quantifying team performance in lacrosse and exploring imu features towards performance enhancement. *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies*, 5(3):1–28, 2021.

[10] A. Khan, S. Mellor, R. King, B. Janko, W. Harwin, R. S. Sherratt, I. Craddock, and T. Plötz. Generalized and efficient skill assessment from imu data with applications in gymnastics and medical training. *ACM Transactions on Computing for Healthcare*, 2(1):1–21, 2020.

[11] A. Khan, J. Nicholson, and T. Plötz. Activity recognition for quality assessment of batting shots in cricket using a hierarchical representation. *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies*, 1(3):1–31, 2017.

[12] C. Lian, R. Ma, X. Wang, Y. Zhao, H. Peng, T. Yang, M. Zhang, W. Zhang, X. Sha, and W. J. Li. Ann-enhanced iot wristband for recognition of player identity and shot types based on basketball shooting motion analysis. *IEEE Sensors Journal*, 22(2):1404–1413, 2021.

[13] A. Maksai, X. Wang, and P. Fua. What players do with the ball: A physically constrained interaction modeling. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 972–981, 2016.

[14] V. Ramanathan, J. Huang, S. Abu-El-Haija, A. Gorban, K. Murphy, and L. Fei-Fei. Detecting events and key actors in multi-person videos. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 3043–3053, 2016.

[15] M. Seuter, A. Pollock, G. Bauer, and C. Kray. Recognizing running movement changes with quaternions on a sports watch. *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies*, 4(4):1–18, 2020.

[16] P. Wittenburg, H. Brugman, A. Russel, A. Klassmann, and H. Sloetjes. ELAN: A Professional Framework for Multimodality Research. In *5th International Conference on Language Resources and Evaluation*, 2006.

[17] X. Xu, E. Nemati, K. Vatanparvar, V. Nathan, T. Ahmed, M. M. Rahman, D. McCaffrey, J. Kuang, and J. A. Gao. Listen2cough: Leveraging end-to-end deep learning cough detection model to enhance lung health assessment using passively sensed audio. *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies*, 5(1):1–22, 2021.