Medium Scale Benchmark for Cricket Excited Actions Understanding

Altaf Hussain Noman Khan Muhammad Munsif Min Je Kim Sung Wook Baik*

Sejong University, Republic of Korea

Abstract

The Sports Action Recognition (SAR) domain is of significant importance in research, with diverse applications, ranging from aiding coaches in strategic decision-making to empowering athletes and contributing to real-time commercial entertainment. Despite the existence of extensive large-scale and small-scale datasets, the direct application of these datasets to specific sports domains, such as cricket, poses challenges. Existing datasets predominantly center around daily life actions, lacking the necessary granularity for in-depth sports analyses. Current Cricket Action Analysis (CAA) datasets have limitations, including their small scale, modality constraints, and their narrow focus on specific aspects, such as cricket batting. Recognizing the need for a more comprehensive benchmark, this article introduces the Cricket Excited Actions (CEA) dataset. Developed in collaboration with professional cricket players, the CEA dataset encompasses challenging multi-person actions within realistic cricket scenarios. The selected activity classes, such as Clean Bowled, Six, Four, and Catches, adhere to official standards and represent pivotal moments in cricket matches. Through precise annotation and empirical studies, utilizing state-of-the-art action recognition model architectures, this study provides a valuable resource for further research and makes significant contributions by offering support essential to advancing CAA within the cricket sports community. The data and code are available at https://github.com/Altaf-hucn/Cricket-Excited-Actions-Benchmark.

1. Introduction

Recently, Sports Action Recognition (SAR) methods have emerged as a vital and dynamic research domain of Computer Vision (CV), attracting considerable attention from both academics and industry [20] [2] [29]. SAR has found applications, particularly in television programs, where it serves the purpose of generating highlights and providing entertainment. It also plays a pivotal role in assisting coaches in making informed decisions and empowering athletes to conduct comprehensive self-analyses of their performance. Despite significant advances in ordinary human Action Recognition (AR), facilitated by the large-scale labeled video datasets [42] [23] [7] [8] [16] [15] [47] and the existence of diverse Deep Learning (DL) models [41] [50] [12] [30] [3] [28], direct use of these for specific sports domains, such as cricket, remains a challenge.

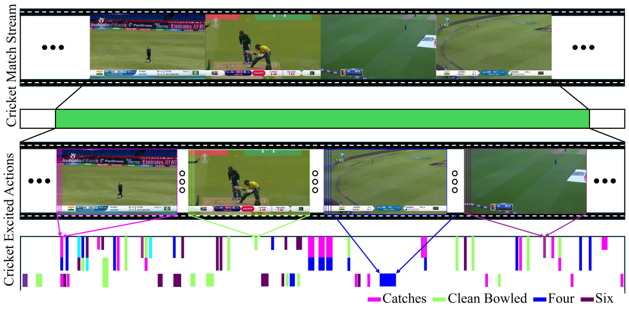


Figure 1. Visual representation of temporal annotations of the Cricket Excited Actions (CEA) datasets for activity recognition.

Prominent datasets, such as ActivityNet [7] and Kenectics-400 [22], primarily focus on the activities of daily life, such as walking, running, and sitting. While some datasets are related to sports action, their labels often lack granularity, making them difficult to directly apply to specific sports analyses, such as cricket. Existing datasets cover a wide array of sports, including squash, basketball, football, tennis, volleyball, hockey, badminton, gymnastics, skating, etc. Some datasets combine different sports categories. Yet, to the best of our knowledge, only three datasets have been proposed for Cricket Action Analysis (CAA) [2] [35]. However, these datasets present certain limitations. E.g. DPC_Images [35] is an image dataset containing a total of 8,646 images, covering only two classes: delivery and play. EXINP [35] is an audio dataset comprising 868 audio clips categorized into Excited, Interval, and Normal Play. CKT [2] is a smaller dataset, encompassing categories such

[*] Corresponding author: Sung Wook Baik (*sbaik@sejong.ac.kr*)

as Pull Shot, Bowled, Reverse Sweep, Defense, and Cover Drive. This dataset comprises a total of 722 video clips, with each category having only 150 videos. The practical application of this dataset in CAA is constrained due to the generality of its classes, restricting its use for practical CAA applications.

Therefore, we hypothesize the imperative necessity for a novel benchmark dataset for CAA with which to boost advances in the field of cricket. Such a benchmark should comprehensively address various realistic challenges, including 1) the benchmark must incorporate scenarios where multiple players engage in distinct actions within the same scene. The motion information of each player should play a pivotal role in discerning and classifying their actions. Additionally, the dataset should encompass variability in playing conditions, capturing the dynamic nature of cricket gameplay. 2) Each action's duration should be thoroughly defined both temporally and semantically in relation to the time. This precision is essential for accurate CAA, contributing to an understanding of the temporal dynamics inherent in the SAR. 3) The benchmark dataset must cover the practical applicability of cricket activities in order to be able to use it for other applications. Their complexity necessitates the incorporation of accurate human pose information, the consideration of potential interactions between players, and addressing occlusion challenges. Furthermore, it should provide a comprehensive and challenging scenario for the evaluation of existing state-of-the-art AR models. These considerations highlight the existing research gaps and limitations in the sports datasets for recognizing cricket-specific actions.

Based on these guidelines and in collaboration with professional cricket players, we introduce a benchmark dataset referred to as Cricket Excited Actions (CEA), aiming to enhance the emerging field of the study of cricket. It is characterized by its large-scale size, high quality, multi-person composition, and inclusion of fine-grained activities with practical applicability in the domain. This dataset contains four distinct classes of activity, depicted in Figure 1: Clean Bowled, Six, Four, and Catches. The rationale behind selecting these specific categories is based on several considerations: First, these categories feature instances of multiple concurrent actions involving spatial and temporal patterns across multiple frames. Secondly, the boundaries and details of the chosen classes are precisely defined by official organizations, such as the International Cricket Council (ICC) [19]. This adherence to established standards enhances the dataset's reliability and relevance to real-world cricket scenarios. Third, recognizing such activities requires learning the long-term spatiotemporal structure of players' interactions with the ball and bat. The dataset thus facilitates the development of sports-based DL models capable of understanding the dynamics inherent in these cricket actions. Finally, and most importantly, these activities are key moments of enjoyment in cricket matches. Additionally, coaches are particularly focused on improving players' performance in these specific areas, making them crucial for comprehensive skill development.

Usually, to develop a benchmark dataset, precision is required, specifically in the annotations. This must involve domain experts and professional players in the video labeling process. Therefore, to avoid potential mistakes in labeling and to ensure accuracy, we employed professionals with a profound understanding of cricket for the video annotation. Our annotation process was structured in two sequential stages: 1) an initial annotation by a team of domain knowledge experts for each category, and 2) subsequent refinement of spatial and temporal aspects by professional players within the quality control team. Furthermore, we performed an empirical study of state-of-the-art AR methods using the CEA dataset and conducted a comparative analysis with a prior benchmark dataset for cricket. The outcomes of our study are detailed in a report on a comprehensive analysis, shedding light on the complex spatiotemporal aspects of cricket. Based on our analyses, we identified several challenges that warrant attention, including the complexity of the spatial patterns in cricket actions, modeling long-term temporal dynamics, taking into account variable playing conditions, managing limited training data, and addressing player occlusion. We anticipate that the introduction of CEA as a challenging benchmark dataset will contribute to the advancement of the cricket community, fostering practical applications. The primary contributions of this paper can be summarized as follows:

- This study introduces a new benchmark dataset for cricket, called CEA, to enhance spatiotemporal CAA using CV methods. It facilitates advances in CAA while addressing the practical needs of the industry. By optimizing player performance in key activities, CEA bridges academia with real-world sports applications.
- We conducted detailed experiments on CEA using baseline methods to analyse their performance. These experiments reveal the primary challenges associated with recognizing complex sports activities, providing valuable insights for the research community to use in future endeavors in this domain.

2. Related Datasets

DL models inherently require substantial amounts of data. Consequently, in the emerging stages of AR research, a predominant focus was placed on action classification. Noteworthy datasets employed for this purpose include KTH [40], Weizmann [6], UCF101 [42], and HMDB51 [23]. More recently, the introduction of large-scale datasets such as YouTube-8M [1], Something-something [15], and Kinetics [8] has been pivotal, and their pre-trained feature rep-

Datasets	Sports	Years	Modalities	#Videos	Avg. Length	# Categories
CVBASE Handball [38]	Handball	2006	RGB	3	10 m	-
CVBASE Squash [38]	Squash	2006	RGB	2	10 m	-
UCF sports [39]	Multiple	2008	RGB	150	6.39s	10
MSR Action3D [26]	Multiple	2010	RGB, depth	567	-	20
Olympic [36]	Multiple	2010	RGB	800	-	16
Hockey Fight [4]	Hockey	2011	RGB	1,000	-	2
ACASVA [10]	Tennis	2011	RGB	6	-	4
THETIS [14]	Tennis	2013	RGB, depth, skeleton	1,980	-	12
Sports 1M [21]	Multiple	2014	RGB	1M	36s	487
Football Action [46]	Football	2017	RGB	3,281	-	5
SpaceJam [13]	Basketball	2018	RGB	15	1.5h	10
Diving48 [27]	Diving	2018	RGB	18,404	-	48
TTStroke-21 [33]	Table Tennis	2018	RGB	129	43 m	21
GolfDB [34]	Golf	2019	RGB	1,400	-	8
FineBasketball [17]	Basketball	2020	RGB	3,399	-	26
Stroke Recognition [24]	Table Tennis	2021	RGB	22,111	-	11
NPUBasketball [32]	Basketball	2021	RGB, depth, skeleton	2,169	-	12
Win-Fail [37]	Multiple	2022	RGB	1,634	3.3	2
Stroke Forecasting [49]	Badminton	2022	RGB	43,191	-	10
FenceNet [51]	Fencing	2022	RGB	652	-	6
Cricket Related Datasets						
CKT [2]	Cricket	2023	RGB	722	3	5
DPC_Images [35]	Cricket	2023	RGB	8,646 (Frames)	-	2
EXINP [35]	Cricket	2023	Audio	868	14	3

Table 1. A list of sports-related datasets. The term ‘multiple’ means that the dataset covers various sports.

resentations have proven beneficial for downstream tasks. However, it is significant that most of these datasets primarily revolve around daily life activities. In contrast, datasets specifically tailored for sport-related actions are relatively limited. This section conducts a comprehensive review of datasets relevant to sports actions, presented in Table 1.

2.1. Football

Soccer-ISSIA [11]: This dataset encompasses 18,000 high-resolution frames captured by 6 static cameras. This dataset is predominantly employed for tasks such as player tracking, detection, and team activity recognition. However, the dataset is relatively limited in scale.

Football Action [46]: This proprietary dataset was recorded using 14 synchronized Full HD cameras, providing annotations for player positions with bounding boxes and comprising five distinct activity categories. Regrettably, this dataset is not accessible to the public. It plays a pivotal role in propelling advances in football AR research. They address diverse research requirements, spanning from player detection to AR, each characterized by distinct scales and levels of annotation detail. However, these datasets are limited in scale and unable to be accessed publicly.

2.2. Basketball

SpaceJam [13]: Collecting 10 categories of basketball actions from NBA and Italian championship videos, SpaceJam includes RGB images and estimated player poses. It is suitable for developing skeleton-based AR models but is limited by its small scale.

FineBasketball [17]: Developed for fine-grained basketball AR, this dataset includes three broad categories and 26 fine-grained categories. It is limited due to its imbalanced data distribution.

NPUBasketball [32]: Comprising 2,169 self-recorded video clips of basketball actions, NPUBasketball provides RGB frames, depth maps, and player skeletons. It is suitable for various AR models but transferring models trained on it to broadcasting videos is difficult due to its self-recorded nature.

These datasets cater to different needs in basketball AR, providing annotations for player positions, ball movements, and various action categories. However, they vary in scale, level of detail of the annotation, and challenges, thereby offering researchers a diverse set of resources for advancing research in basketball analysis.

2.3. Volleyball

HierVolleyball [18]: Developed for team AR, this dataset contains 1,525 annotated frames from 15 YouTube volleyball videos. It includes action labels for individual players as well as group activities, such as setting, spiking, and passing. While these volleyball datasets offer dense annotations, such as player bounding boxes, they are relatively small in scale and feature coarse action categories.

2.4. Hockey

Hockey Fight [4]: Hockey fight is a binary classification dataset that contains fight and non-fight instances in hockey games. This dataset comprises 1,000 video clips from NHL games, each containing 50 frames with corresponding labels. However, this dataset is limited by its small-scale and coarse categories, with 'fight' focusing solely on binary classification, and the player tracking is primarily for player detection.

2.5. Tennis

ACASVA [10]: Developed for tennis AR, ACASVA consists of 6 broadcast videos of tennis games with three categories of action: hit, non-hit, and serve. It annotates the players' positions and temporal boundaries of the actions but only provides the extracted features of the video clips instead of the original videos.

THETIS [14] : Comprising 1,980 self-recorded videos, THETIS includes 12 tennis actions categorized into backhand shots, forehand shots, service shots, and smashes. It provides RGB frames, depth videos, and 2D/3D skeleton videos, allowing the development of multiple types of AR models.

While existing tennis datasets are limited by their small-scale and coarse annotations, they offer multiple modalities, such as RGB frames, textual descriptions, and depth maps, which can benefit research in multimodal learning approaches.

2.6. Table Tennis

TTStroke-21 [33]: Comprising 129 self-recorded videos of 94 hour-long games, TTStroke-21 annotates 1,378 actions into 21 categories, such as serve backhand spin and forehand loop. Despite the fast-paced nature of table tennis, this dataset is not considered to pose difficulties, possibly due to its high frame rate (120 FPS).

Stroke Recognition [24]: Similar to TTStroke-21, but larger in scale, Stroke Recognition includes 22111 trimmed videos with strokes categorized into 11 categories. Despite its size, the use of this dataset is less challenging: high accuracy can be achieved with simple models.

These table tennis datasets offer valuable resources for stroke recognition and analysis, featuring annotations for various aspects of the game, such as stroke categories, ball positions, and player poses. However, limitations, such as the size of the dataset, imbalanced data, and quality of the annotations, need to be considered in using these datasets effectively.

2.7. Badminton

Stroke Forecasting [49]: Consisting of 43191 trimmed video clips, each annotated with one of 10 stroke categories, Stroke Forecasting also enables stroke forecasting tasks, predicting the next stroke in a rally based on previous strokes.

2.8. Diving

Diving48 [27]: Including 18404 video segments covering 48 fine-grained diving categories, Diving48 provides a relatively unbiased dataset for AR tasks. It is diverse in its covered sport, providing annotations for various aspects such as player positions, stroke boundaries, action categories, and quality assessment scores.

2.9. Multiple Types of Sports

UCF Sports [39]: A dataset with 150 video clips at 10 FPS and covering 10 sports categories including diving, golf swing, and running. MSR Action3D [26]: Comprising 576 sequences of depth maps, MSR Action3D enables sports AR tasks such as tennis serve and golf swing.

Olympic [36]: This dataset includes 800 videos covering 16 sports categories, such as long jump, tennis serve, and diving, sourced from YouTube.

Sports 1M [21]: A large-scale dataset with around one million videos sourced from YouTube, covering 487 sports categories and facilitating coarse and fine-grained classification.

2.10. Others

Other datasets are very few in number and have focused on other sports instead of those discussed above. Such datasets include Win–Fail [37], a dataset that focuses on recognizing win or fail outcomes of actions, collecting 817 win–fail video pairs from various domains.

CVBASE Handball [38] & CVBASE Squash [38]: These datasets provide trajectories of players and action categories for handball and squash AR tasks. GolfDB [34]: Facilitating golf swing analysis, GolfDB includes high-quality video segments with action labels and player bounding boxes.

FenceNet [51]: Comprising 652 videos of expert-level fencers performing actions across six categories, FenceNet offers RGB frames, 3D skeleton data, and depth data. All of these datasets are proposed for sports that are not particularly popular, and none of these datasets have specifically been designed to focus on cricket.

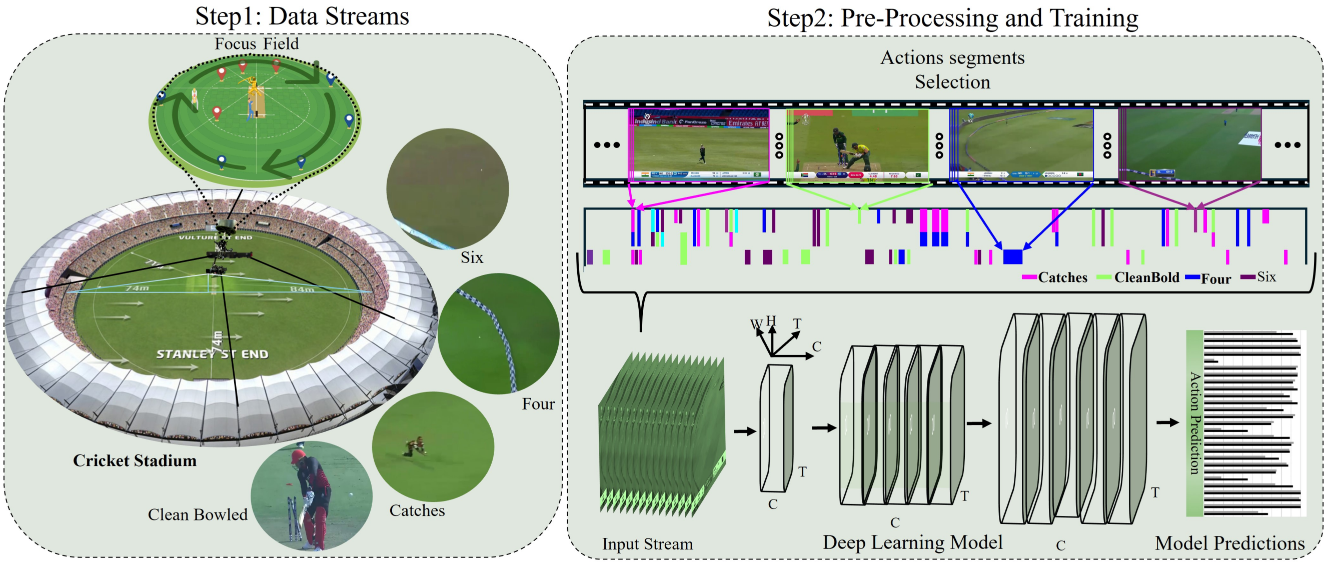


Figure 2. Framework for cricket sport activity recognition. Consists of two modules, where first we collected data for cricket and then pre-processed and trained various state-of-the-art baseline models.

2.11. Cricket Sport Action Recognition

All the sports datasets discussed earlier were proposed for sports other than cricket. Our work specifically concentrates on cricket actions. Currently, only three datasets are available for CAA, namely DPC_Images [35], EXINP [35], and CKT [2]. The details of these datasets are provided below.

DPC_Images[35]: It is a frame-based dataset designed for summarizing cricket videos. Comprising two distinct classes, namely, delivery and play, this dataset is collected from international cricket tournament matches, including the ICC Under-19 Cricket World Cup 2022, India tour of South Africa (2022), and the ICC Men's T20 World Cup 2021. The dataset encompasses a total of 8,646 images, with the delivery class containing 2,304 images and the play class 6,442 images. The delivery class features images corresponding to activities of cricket bowling, characterized by the bowlers swinging their arm above the shoulder and releasing the ball without further straightening of the elbow after reaching shoulder level. In contrast, all activities, or events apart from the delivery, such as various gameplay actions, are included in the play class.

EXINP [35]: This dataset is an audio dataset capturing audio segments from cricket play. Categorized into excited, interval, and normal play, this dataset comprises a total of 868 audio segments, each one second long. The audio segments are collected from the Big Bash League (2017–2018). There are three distinct categories: interval (119), normal play (492), and excited (257). The categorization is accomplished through manual segmentation and labeling, aligning events in each segment with the defined categories. Specifically, the excited category encompasses key events in the cricket domain, featuring high audio information. The normal play category includes audio segments with routine and normal events, characterized by less audio information and minimal significance for the key events. The interval category contains audio segments with substantial audio information but lacking an association with key events in the context of the cricket match.

CKT [2]: This dataset has been curated from sources including YouTube and cricket-info websites. Encompassing a total of 722 videos, this dataset only provides a representation of batting activities, including pull shot, bowled, reverse sweep, defense, and cover drive. All videos hold a frame rate of 30 frames per second and share backgrounds of the grounds, pitches, and spectator accommodations. Each class within the dataset consists of 150 videos. The dimensions of each frame are standardized at 840 x 480 pixels.

Although DPC_Images [35], EXINP [35], and CKT [2] contribute valuable resources for CAA, but they have certain limitations. DPC_Images [35] is an image dataset, designed for summarization tasks, and focuses on only two classes, limiting its coverage of diverse cricket actions and potentially hindering generalization. Similarly, the fact that EXINP[35] is an audio dataset impacts the model's ability to generalize across various audio segments. CKT [2] is a video dataset, and its relatively small scale and the generality of its activity classes limit the development of robust CAA models. Furthermore, the uniformity in the background and setting aids controlled experiments but limits its adaptability to varied environments. Lastly, the shortness of their video clips limits the ability to capture the full temporal dynamics between the player's interaction and the boundary in certain cricket actions. Therefore, efforts to enhance the emerging field of CAA should consider these limitations and propose a more suitable dataset for informed decision-making in the context of CAA tasks.

3. The Proposed Cricket Excited Actions Dataset

The primary objective of our CEA dataset is to establish a novel and demanding benchmark for CAA. An overall pictorial representation of the method of collecting and processing this dataset is presented in Figure 2, which shows how the dataset encompasses scenarios with multiple players engaging in various actions, each possessing distinct temporal durations. The inherent complexity of each class within the dataset is thoroughly designed to align with practical applications to cricket and sports activity. In this section, we present a comprehensive discussion covering the dataset's construction, its distinctive characteristics, and the diverse challenges associated with it.

3.1. Construction of the Dataset

Generating the Action Vocabulary: To generate the action vocabulary in cricket, we selected the categories of Clean Bowled, Sixes, Fours, and Catches, because they involve multiple players, have less ambiguous actions, and have well-defined temporal boundaries. Additionally, these 4 activities are in great demand in several applications, such as TV sports commercials and coaches' training a player for these key actions. Generally, cricket competitions at the international level are classified into three playing formats: 1) Test matches 2) One-Day Internationals, and 3) Twenty20 Internationals. These matches are played under the rules and regulations approved by the ICC, which also provides the match officials [19].

Data preparation: These 4 categories were selected and downloaded from several cricket videos from YouTube-verified channels whose generation authorities have been approved by the ICC website [19]. To ensure accuracy and avoid potential mistakes in labeling, for annotating the videos we engaged professionals with a profound understanding of cricket. The team of domain knowledge experts for each category was engaged and then the annotations were refined in their spatial and temporal aspects by professional players during the quality control step. Each video selected had a high resolution and dimensions of 720 or 1080p. These videos are long videos, and our desired classes arise randomly, therefore, we manually annotated each video to define the temporal boundary of each action. Any content containing other background scenes, such as awards, strategic time, or reviews for the Third Empire, was discarded. In the quality control, the annotations underwent two stages of scrutiny. Initially, domain experts verified each clip, rectifying inaccuracies and adding missing annotations, ensuring all cricket actions were taken into account. Subsequently, each instance was reviewed at a playback of 5 FPS, correcting any temporal discrepancies. These rigorous quality control measures ensured the accuracy and reliability of the cricket dataset, bolstering its suitability for research and analysis.

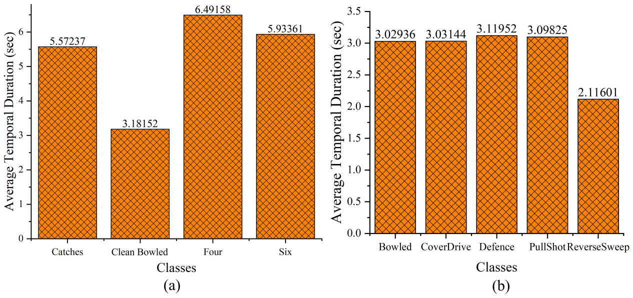
3.2. Characteristics and Statistics of the Dataset

In the CEA dataset, 4 classes contain fine-grained actions collected from multiple cricket matches. Table 1 compares the statistics of the existing datasets about cricket. For instance, EXINP [35] is an audio dataset with 868 audio clips which is not directly used to analyse the cricket activity recognition. Similarly, DPC_Images [35] is an RGB images dataset with 8646 images for two classes, namely, Delivery, with 2304 images, and Play, with 6442. CKT [2] is an RGB video dataset with 5 classes. However, this dataset is limited to only 722 video clips, each only three seconds long. Additionally, their action categories are not directly used for practical applications such as the generation of TV commercials of highlights. In response to these considerations, our CEA is a large benchmark dataset with a total of 2,146 clips. Its actions have different temporal durations, which makes the development and use of AR models more difficult. Furthermore, our action instances are often related to a longer temporal context and intersection with the context, as shown in Figure 3. The categories selected enhance the field of cricket research with the context of their applications. Our CEA has several distinctive characteristics when compared with existing datasets.

Challenging: CEA presents significant complexity compared to other datasets. It features multiple players assuming various positions to execute actions, requiring that models discern the categories of the actions amidst diverse backgrounds. Moreover, the dataset covers a wide range of temporal dynamics with overlapping actions, which are marked by fast movements of the camera and the players, resulting in notable deformation and occlusion of the actions. The similarities of the features of the existing and our proposed datasets are shown in Figure 4.

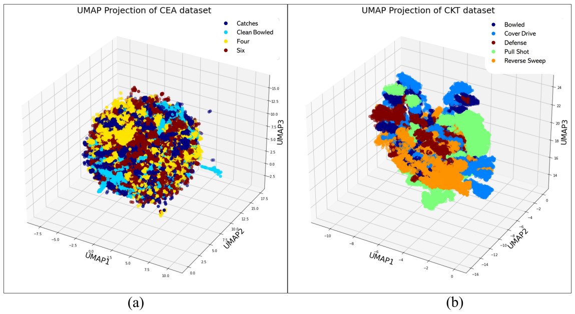
High Quality: The high-resolution data (720p or 1080p) were collected from different sources and diverse tournaments, ensuring detailed preservation of the activities. Additionally, a professional athlete annotated precisely the dataset with temporal boundaries. The annotation process is combined with strict quality control measures, ensuring the consistency and cleanliness of the annotations.

Diversity: The videos in each category were collected from various cricket events across different countries and genders, promoting a less biased and better-balanced dataset suitable for comprehensive sports analysis.



(a) (b)

Figure 3. Visual representation of the average temporal duration for each category. Here, (a) represents our proposed and (b) represents the CKT dataset.



(a) (b)

Figure 4. Classes similarity of existing and our proposed dataset. Here, (a) represents our and (b) represents the CKT dataset.

Application Orientation: The versatility of CEA extends to various applications within the analysis of cricket events, from facilitating the automatic generation of highlights for TV commercials to enabling AI-driven referee systems and technical assistance, the dataset has many uses. Additionally, it serves as a valuable resource for assessing a player's abilities, formulating training plans, formulating game strategies, and facilitating player trades between teams.

4. Experiments and Analysis

In this section, we offer a thorough examination of state-of-the-art baseline AR models using CKT [2] and our proposed CEA dataset which illuminate the manifold challenges present in CEA.

4.1. Experimental Setting

Experiments are conducted using Linux operating systems, the PyTorch DL library, and the NVIDIA GeForce RTX 3090 GPU. We used the MMAction2 [9] benchmarks with a standard split for training and testing AR. Each model was trained for 100 epochs without pre-trained weights, due to the fact that multiple players were engaging in distinct actions within the same scene, thus necessitating a nuanced approach to understanding the task especially when shifting from daily activities to sports actions.

4.2. Spatio-temporal Action Recognition Results

In the literature on sports and AR, numerous methods have been proposed, demonstrating remarkable performance and being effectively used in different applications. In order to contribute to the growing body of knowledge in cricket SAR, we conducted a rigorous evaluation of our newly introduced CEA and existing CKT [2] dataset. With these experiments, this work will serve as a valuable benchmark for the research community in this domain.

4.2.1 Performance of the CNN models

To establish a robust baseline, we employed the 10 state-of-the-art models tabulated in Table 2, encompassing both CNN and transformer-based architectures, due to their proven efficacy in AR tasks. The inclusion of these models is grounded in the rich domain knowledge of CV and DL, recognizing their ability to capture both spatial and temporal information, something which is crucial for understanding complex sports activities.

In the CNN category, our evaluation involved prominent models, such as C3D [43], TSN [48], C2D [50], I3D [8], R2+1D [44], CSN [46], and SlowOnly [12]. These models have demonstrated exceptional performance in various AR scenarios, highlighting their applicability and versatility. The reported achievements for these models underscore their effectiveness in extracting meaningful features from the spatio-temporal dynamics inherent in CEA, achieving a top-1 % accuracy of, respectively, 0.8326, 0.8018, 0.8150, 0.8062, 0.8590, 0.7577, and 0.4141. On the other hand, all of the models exhibit higher performance, as shown in Table 2, on the CKT dataset compared to our proposed dataset, highlighting challenges within our proposed dataset.

4.2.2 Performance of the Transformer Models

Regardless of the CNN models, our evaluation extended to transformer-based architectures such as TimeSformer [5], VideoSwin [31], and UniFormerV2 [25] models. These transformer models have exhibited significant prowess in capturing long-range dependencies and temporal intricacies, making them pertinent choices for spatiotemporal AR in cricket. The reported results for TimeSformer [5], VideoSwin [31], and UniFormerV2 [25] underscore their competitive performance, achieving, 0.7665, 0.3480, 0.6211 top 1 accuracy on the testing data. Among all these models, the CNN models achieved higher performance as compared to the transformer models due to their inductive biases. Analysis of the learning of the intermediate features of the R2+1D [44] is shown in the testing videos as shown

Methods	Year	Res	Top-1 (%)	
			CEA	CKT[2]
C3D [43]	2015	112×112	83.26	81.13
TSN [48]	2016	224×224	80.18	–
C2D [50]	2018	224×224	81.50	84.91
I3D [8]	2018	224×224	80.62	97.17
R2+1D [44]	2018	112×112	85.90	91.51
CSN [45]	2019	224×224	75.77	99.06
SlowOnly [12]	2019	256×256	41.41	99.06
TimeSformer [5]	2021	224×224	76.65	96.23
VideoSwin [31]	2022	224×224	34.80	–
UniFormerV2 [25]	2022	224×224	62.11	–

Table 2. Performance of various state-of-the-art baseline action recognition methods on proposed CEA and existing CKT datasets.

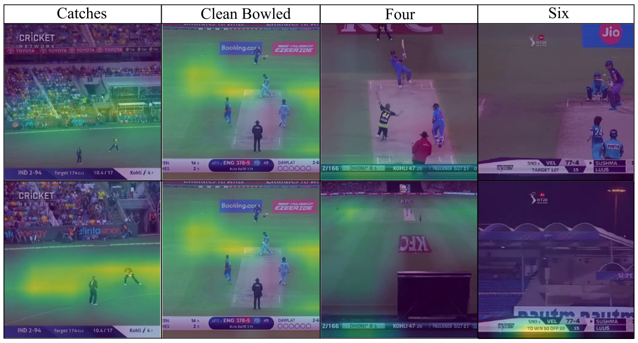


Figure 5. Attention of the R2+1D model on the test set of our proposed dataset.

in Figure 5. Their confusion performance on the testing data is shown in Figure 6.

This comprehensive evaluation of the existing dataset and our CEA dataset not only establishes it as a robust baseline for CAA but also sheds light on the nuanced strengths and capabilities of both CNN and transformer models in the context of spatio-temporal action analysis. These findings contribute valuable insights to the broader discourse on sports analytics, emphasizing the role of advanced DL architectures in understanding and interpreting intricate sports activities.

However, our CEA dataset poses several challenges for CAA. Firstly, it involves complex actions with multiple players, requiring any model to detect subtle motion cues accurately. Secondly, diverse temporal dynamics across action categories complicate the task of recognition. Thirdly, the rapid movements of the camera and the players introduce challenges like occlusions, limiting DL models' effectiveness. Finally, the dataset's focus on practical applications in the industry intensifies the challenge of the development and evaluation of a model in real-world sports scenarios.

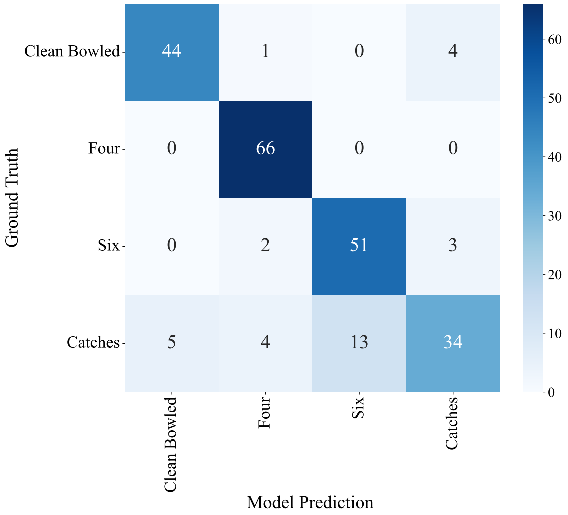


Figure 6. Confusion matrix of the R2+1D model on the testing data.

5. Conclusion

This study has introduced the CEA dataset, which addresses the limitations of existing datasets in the domain of CAA. The proposed dataset is a large-scale, high-quality dataset with multi-person actions, filling a crucial gap in this domain. The selected activity classes, aligned with official standards, focus on key moments in cricket matches. Through empirical studies, we have identified the challenges such data present to the development of AR models, including the complexity of the spatial patterns and the long-term temporal dynamics. The CEA dataset not only can serve as a valuable benchmark for CAA but can also contribute insights for addressing these challenges. We believe CEA will catalyse advances in recognizing cricket-specific actions and enhance the practical applications of CAA in various domains. In the future, it needed to focus on expanding the CEA dataset to include more diverse cricket actions, refining the architectures of the models to address the spatial and temporal challenges, exploring multi-modal approaches for improved recognition, and fostering collaborations for specialized models aligned with cricketing contexts.

Acknowledgement: This work was supported by a grant from the National Research Foundation of Korea (NRF), funded by the Korean government (MSIT), Grant/Award Number: (2023R1A2C1005788).

References

[1] Sami Abu-El-Haija, Nisarg Kothari, Joonseok Lee, Paul Natsev, George Toderici, Balakrishnan Varadarajan, and

Sudheendra Vijayanarasimhan. Youtube-8m: A large-scale video classification benchmark. *arXiv preprint arXiv:1609.08675*, 2016. 2

[2] Waqas Ahmad, Muhammad Munsif, Habib Ullah, Mohib Ullah, Alhanouf Abdulrahman Alsuwailem, Abdul Khader Jilani Saudagar, Khan Muhammad, and Muhammad Sajjad. Optimized deep learning-based cricket activity focused network and medium scale benchmark. *Alexandria Engineering Journal*, 73:771–779, 2023. 1, 3, 5, 6, 7, 8

[3] Anurag Arnab, Mostafa Dehghani, Georg Heigold, Chen Sun, Mario Lučić, and Cordelia Schmid. Vivit: A video vision transformer. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 6836–6846, 2021. 1

[4] Enrique Bermejo Nievas, Oscar Deniz Suarez, Gloria Bueno García, and Rahul Sukthankar. Violence detection in video using computer vision techniques. In *Computer Analysis of Images and Patterns: 14th International Conference, CAIP 2011, Seville, Spain, August 29-31, 2011, Proceedings, Part II 14*, pages 332–339. Springer, 2011. 3, 4

[5] Gedas Bertasius, Heng Wang, and Lorenzo Torresani. Is space-time attention all you need for video understanding? In *ICML*, page 4, 2021. 7, 8

[6] Moshe Blank, Lena Gorelick, Eli Shechtman, Michal Irani, and Ronen Basri. Actions as space-time shapes. In *Tenth IEEE International Conference on Computer Vision (ICCV'05) Volume 1*, pages 1395–1402. IEEE, 2005. 2

[7] Fabian Caba Heilbron, Victor Escorcia, Bernard Ghanem, and Juan Carlos Niebles. Activitynet: A large-scale video benchmark for human activity understanding. In *Proceedings of the ieee conference on computer vision and pattern recognition*, pages 961–970, 2015. 1

[8] Joao Carreira and Andrew Zisserman. Quo vadis, action recognition? a new model and the kinetics dataset. In *proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 6299–6308, 2017. 1, 2, 7, 8

[9] MMAction Contributors. Openmmlab's next generation video understanding toolbox and benchmark. 2020. 7

[10] Teofilo De Campos, Mark Barnard, Krystian Mikolajczyk, Josef Kittler, Fei Yan, William Christmas, and David Windridge. An evaluation of bags-of-words and spatio-temporal shapes for action recognition. In *2011 IEEE Workshop on Applications of Computer Vision (WACV)*, pages 344–351. IEEE, 2011. 3, 4

[11] Tiziana D'Orazio, Marco Leo, Nicola Mosca, Paolo Spagnolo, and Pier Luigi Mazzeo. A semi-automatic system for ground truth generation of soccer video sequences. In *2009 Sixth IEEE International Conference on Advanced Video and Signal Based Surveillance*, pages 559–564. IEEE, 2009. 3

[12] Christoph Feichtenhofer, Haoqi Fan, Jitendra Malik, and Kaiming He. Slowfast networks for video recognition. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 6202–6211, 2019. 1, 7, 8

[13] Simone Francia, Simone Calderara, and Dott Fabio Lanzi. Classificazione di azioni cestistiche mediante tecniche di deep learning. *URL: https://www. researchgate. net/publication/330534530_Classificazione_di_Azioni_ Cestistiche_mediante_Tecniche_di_Deep_Learning*, 2018. 3

[14] Sofia Gourgari, Georgios Goudelis, Konstantinos Karpouzis, and Stefanos Kollias. Thetis: Three dimensional tennis shots a human action dataset. In *Proceedings of the IEEE conference on computer vision and pattern recognition workshops*, pages 676–681, 2013. 3, 4

[15] Raghav Goyal, Samira Ebrahimi Kahou, Vincent Michalski, Joanna Materzynska, Susanne Westphal, Heuna Kim, Valentin Haenel, Ingo Fruend, Peter Yianilos, Moritz Mueller-Freitag, et al. The" something something" video database for learning and evaluating visual common sense. In *Proceedings of the IEEE international conference on computer vision*, pages 5842–5850, 2017. 1, 2

[16] Chunhui Gu, Chen Sun, David A Ross, Carl Vondrick, Caroline Pantofaru, Yeqing Li, Sudheendra Vijayanarasimhan, George Toderici, Susanna Ricco, Rahul Sukthankar, et al. Ava: A video dataset of spatio-temporally localized atomic visual actions. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 6047–6056, 2018. 1

[17] Xiaofan Gu, Xinwei Xue, and Feng Wang. Fine-grained action recognition on a novel basketball dataset. In *ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 2563–2567. IEEE, 2020. 3

[18] Mostafa S Ibrahim, Srikanth Muralidharan, Zhiwei Deng, Arash Vahdat, and Greg Mori. A hierarchical deep temporal model for group activity recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1971–1980, 2016. 4

[19] ICC. Icc playing handbook, 2024. 2, 6

[20] Christian Keilstrup Ingwersen, Christian Møller Mikkelstrup, Janus Nørtoft Jensen, Morten Rieger Hannemose, and Anders Bjorholm Dahl. Sportspose-a dynamic 3d sports pose dataset. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 5218–5227, 2023. 1

[21] Andrej Karpathy, George Toderici, Sanketh Shetty, Thomas Leung, Rahul Sukthankar, and Li Fei-Fei. Large-scale video classification with convolutional neural networks. In *Proceedings of the IEEE conference on Computer Vision and Pattern Recognition*, pages 1725–1732, 2014. 3, 4

[22] Will Kay, Joao Carreira, Karen Simonyan, Brian Zhang, Chloe Hillier, Sudheendra Vijayanarasimhan, Fabio Viola, Tim Green, Trevor Back, Paul Natsev, et al. The kinetics human action video dataset. *arXiv preprint arXiv:1705.06950*, 2017. 1

[23] Hildegard Kuehne, Hueihan Jhuang, Estíbaliz Garrote, Tomaso Poggio, and Thomas Serre. Hmdb: a large video database for human motion recognition. In *2011 International conference on computer vision*, pages 2556–2563. IEEE, 2011. 1, 2

[24] Kaustubh Milind Kulkarni and Sucheth Shenoy. Table tennis stroke recognition using two-dimensional human pose estimation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 4576–4584, 2021. 3, 4

[25] Kunchang Li, Yali Wang, Yinan He, Yizhuo Li, Yi Wang, Limin Wang, and Yu Qiao. Uniformerv2: Spatiotemporal

learning by arming image vits with video uniformer. *arXiv preprint arXiv:2211.09552*, 2022. 7, 8

[26] Wanqing Li, Zhengyou Zhang, and Zicheng Liu. Action recognition based on a bag of 3d points. In *2010 IEEE computer society conference on computer vision and pattern recognition-workshops*, pages 9–14. IEEE, 2010. 3, 4

[27] Yingwei Li, Yi Li, and Nuno Vasconcelos. Resound: Towards action recognition without representation bias. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 513–528, 2018. 3, 4

[28] Yan Li, Bin Ji, Xintian Shi, Jianguo Zhang, Bin Kang, and Limin Wang. Tea: Temporal excitation and aggregation for action recognition. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 909–918, 2020. 1

[29] Yixuan Li, Lei Chen, Runyu He, Zhenzhi Wang, Gangshan Wu, and Limin Wang. Multisports: A multi-person video dataset of spatio-temporally localized sports actions. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 13536–13545, 2021. 1

[30] Ji Lin, Chuang Gan, and Song Han. Tsm: Temporal shift module for efficient video understanding. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 7083–7093, 2019. 1

[31] Ze Liu, Jia Ning, Yue Cao, Yixuan Wei, Zheng Zhang, Stephen Lin, and Han Hu. Video swin transformer. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 3202–3211, 2022. 7, 8

[32] Chunyan Ma, Ji Fan, Jinghao Yao, and Tao Zhang. Npu rgb+ d dataset and a feature-enhanced lstm-dgcn method for action recognition of basketball players. *Applied Sciences*, 11 (10):4426, 2021. 3

[33] Pierre-Etienne Martin, Jenny Benois-Pineau, Renaud Péteri, and Julien Morlier. Sport action recognition with siamese spatio-temporal cnns: Application to table tennis. In *2018 International Conference on Content-Based Multimedia Indexing (CBMI)*, pages 1–6. IEEE, 2018. 3, 4

[34] William McNally, Kanav Vats, Tyler Pinto, Chris Dulhanty, John McPhee, and Alexander Wong. Golfdb: A video database for golf swing sequencing. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition workshops*, pages 0–0, 2019. 3, 4

[35] Pulkit Narwal, Neelam Duhan, and Komal Kumar Bhatia. A novel multi-modal neural network approach for dynamic and generic sports video summarization. *Engineering Applications of Artificial Intelligence*, 126:106964, 2023. 1, 3, 5, 6

[36] Juan Carlos Niebles, Chih-Wei Chen, and Li Fei-Fei. Modeling temporal structure of decomposable motion segments for activity classification. In *Computer Vision–ECCV 2010: 11th European Conference on Computer Vision, Heraklion, Crete, Greece, September 5-11, 2010, Proceedings, Part II 11*, pages 392–405. Springer, 2010. 3, 4

[37] Paritosh Parmar and Brendan Morris. Win-fail action recognition. In *Proceedings of the IEEE/CVF winter conference on applications of computer vision*, pages 161–171, 2022. 3, 4

[38] Janez Pers. Cvbase 06 dataset: a dataset for development and testing of computer vision based methods in sport environments. *SN, Ljubljana*, 2005. 3, 4

[39] Mikel D Rodriguez, Javed Ahmed, and Mubarak Shah. Action mach a spatio-temporal maximum average correlation height filter for action recognition. In *2008 IEEE conference on computer vision and pattern recognition*, pages 1–8. IEEE, 2008. 3, 4

[40] Christian Schuldt, Ivan Laptev, and Barbara Caputo. Recognizing human actions: a local svm approach. In *Proceedings of the 17th International Conference on Pattern Recognition, 2004. ICPR 2004.*, pages 32–36. IEEE, 2004. 2

[41] Karen Simonyan and Andrew Zisserman. Two-stream convolutional networks for action recognition in videos. *Advances in neural information processing systems*, 27, 2014. 1

[42] Khurram Soomro, Amir Roshan Zamir, and Mubarak Shah. Ucf101: A dataset of 101 human actions classes from videos in the wild. *arXiv preprint arXiv:1212.0402*, 2012. 1, 2

[43] Du Tran, Lubomir Bourdev, Rob Fergus, Lorenzo Torresani, and Manohar Paluri. Learning spatiotemporal features with 3d convolutional networks. In *Proceedings of the IEEE international conference on computer vision*, pages 4489–4497, 2015. 7, 8

[44] Du Tran, Heng Wang, Lorenzo Torresani, Jamie Ray, Yann LeCun, and Manohar Paluri. A closer look at spatiotemporal convolutions for action recognition. In *Proceedings of the IEEE conference on Computer Vision and Pattern Recognition*, pages 6450–6459, 2018. 7, 8

[45] Du Tran, Heng Wang, Lorenzo Torresani, and Matt Feiszli. Video classification with channel-separated convolutional networks. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 5552–5561, 2019. 8

[46] Takamasa Tsunoda, Yasuhiro Komori, Masakazu Matsugu, and Tatsuya Harada. Football action recognition using hierarchical lstm. In *Proceedings of the IEEE conference on computer vision and pattern recognition workshops*, pages 99–107, 2017. 3

[47] Athanasios Voulodimos, Dimitrios Kosmopoulos, Georgios Vasileiou, Emmanuel Sardis, Vasileios Anagnostopoulos, Constantinos Lalos, Anastasios Doulamis, and Theodora Varvarigou. A threefold dataset for activity and workflow recognition in complex industrial environments. *IEEE MultiMedia*, 19(03):42–52, 2012. 1

[48] Limin Wang, Yuanjun Xiong, Zhe Wang, Yu Qiao, Dahua Lin, Xiaoou Tang, and Luc Van Gool. Temporal segment networks: Towards good practices for deep action recognition. In *European conference on computer vision*, pages 20–36. Springer, 2016. 7, 8

[49] Wei-Yao Wang, Hong-Han Shuai, Kai-Shiang Chang, and Wen-Chih Peng. Shuttlenet: Position-aware fusion of rally progress and player styles for stroke forecasting in badminton. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pages 4219–4227, 2022. 3, 4

[50] Xiaolong Wang, Ross Girshick, Abhinav Gupta, and Kaiming He. Non-local neural networks. In *Proceedings of the*

IEEE conference on computer vision and pattern recognition, pages 7794–7803, 2018. 1, 7, 8

[51] Kevin Zhu, Alexander Wong, and John McPhee. Fencenet: Fine-grained footwork recognition in fencing. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 3589–3598, 2022. 3, 4