Detecting and Classifying Punches in Olympic Boxing Matches

Talha Kaba [*1] **Abdulkadir Parlak** [*1]

Abstract

Computer vision has become an essential tool in sports analytics, enabling performance monitoring without requiring invasive sensors. In boxing, detecting punches through video analysis can provide valuable insights for coaches, referees, and analysts. This paper presents a system for automatic punch detection in boxing matches using a single static camera. We propose a convolutional neural network (CNN) model trained on labeled boxing footage to classify frames into eight distinct punch types, including head punches, body punches, blocks, and misses with both hands. The system leverages spatial features from frames to enhance detection accuracy. Experimental evaluation shows that the proposed approach achieves an overall accuracy of 84%, with an F1 score of 91% for right-hand blocks and 86% for left-hand misses. The results demonstrate the system's potential for real-time boxing match analysis, offering automated punch detection and event labeling for enhanced sports analytics.

1. Introduction

Sports analytics has transformed how games are played, coached, and analyzed. In boxing, performance analysis requires evaluating punches based on type, target, and effectiveness. Traditional methods rely on manual scoring by judges or wearable sensors, which can be subjective or intrusive.

Advances in computer vision enable non-intrusive video-based punch detection, analyzing footage to identify and classify punches. Such systems could revolutionize boxing analytics, helping coaches assess performance, referees make better decisions, and broadcasters enhance audience engagement with real-time statistics.

Boxing is complex due to varied punch types, defensive maneuvers, and high-speed action. An effective detection system must distinguish between punches targeting the head or body, thrown with either hand, while recognizing hits, blocks, and misses. Rapid sequences of movements add to the challenge.

Key technical hurdles include varying camera angles, inconsistent lighting, and occlusions caused by boxer movement. A successful system must generalize across different boxers, weight classes, and fighting styles, emphasizing data diversity.

Our research focuses on developing an automated punch detection system using video analysis. We classify punches into predefined categories while leveraging deep learning techniques like convolutional neural networks (CNNs). Our goal is real-time punch detection with high accuracy.

This work contributes to sports analytics by offering a scalable, non-intrusive punch detection framework. It has potential applications in training, officiating, and enhancing viewer experiences, bridging the gap between sports science and artificial intelligence.

2. Related Work

Detecting and analyzing punches in boxing has been a challenging problem in sports analytics. Prior works have utilized both invasive and non-invasive methods for combat sports performance analysis. Invasive methods, such as wearable sensors, provide accurate and direct measurements of punch forces and velocities (Worsey et al., 2020; 2019). However, these systems are often limited by safety concerns and regulatory constraints (Ye et al., 2022). Non-invasive approaches using computer vision and RGB cameras have gained traction as they are safer and more versatile (Kasiri et al., 2015; 2017).

The study by Stefański et al. (2024) highlights the advantages of non-invasive methods, particularly their ability to analyze matches without interfering with the athletes. These systems use convolutional neural networks (CNNs) and other machine learning models to classify punches, leveraging features like motion dynamics and spatial information. Despite their potential, challenges such as occlusion, lighting variations, and data imbalance remain significant hurdles (Behendi et al., 2015).

[1]Department of Artificial Intelligence Engineering, University of Hacettepe, Ankara, Turkiye. Correspondence to: Talha Kaba <talhakaba@hacettepe.edu.tr>, Abdulkadir Parlak <abdulkadirparlak@hacettepe.edu.tr>.

A notable work by Kasiri et al. (2015) focused on depth imagery for punch classification, demonstrating promising results in distinguishing punch types. However, this approach is constrained by the requirement for specialized cameras, limiting its accessibility. Similarly, Wattanamongkhol et al. (2005) explored glove tracking systems, providing accurate punch detection but requiring manual intervention and high computational resources.

Recent advancements include hybrid approaches combining color extraction and background subtraction techniques (Stefański et al., 2024). These methods enhance punch detection accuracy by isolating regions of interest and eliminating irrelevant background information. While effective, these techniques can be computationally expensive and may require further optimization for real-time applications.

This study builds upon these foundational works by leveraging a single static camera setup, optimizing preprocessing pipelines, and focusing on lightweight CNN architectures to achieve high accuracy and efficiency. By addressing the limitations of previous methods, we aim to provide a scalable and robust solution for boxing punch detection.

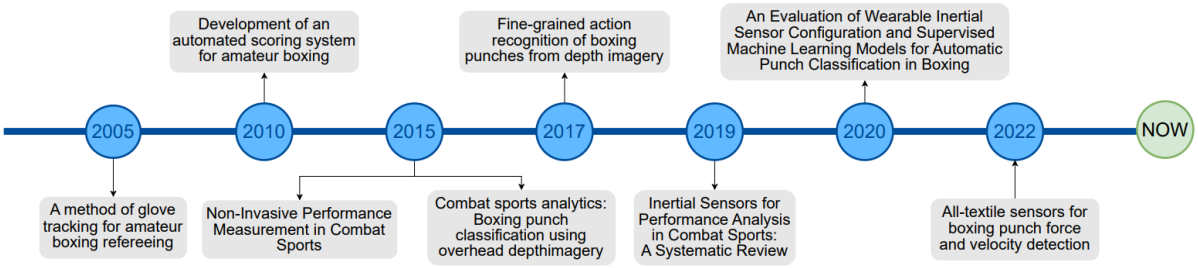

Figure 1. Timeline chart for the most relevant works to the study.

3. Methodology

The methodology for detecting punches in boxing videos consists of a structured preprocessing pipeline, including label translation, frame extraction, image cropping, and image scaling. These steps ensure that the data is properly formatted and optimized for input into the convolutional neural network (CNN) model.

3.1. Data Preprocessing

3.1.1. Label Translation

The dataset originally contained annotations in Polish, which were translated into English for clarity and compatibility with the model. A custom script, `json_label_translator.py`, automated this process by mapping each Polish label to its English equivalent. For example, "Głowa lewą ręką" was translated to "Head with left hand." This step ensured consistency in label representation throughout the dataset.

3.1.2. Frame Extraction

Frames corresponding to annotated punch events were extracted from boxing videos using the `frame_extraction.py` script. This targeted extraction reduced irrelevant data, focusing solely on frames containing critical boxing actions. This approach streamlined the dataset by isolating moments of interest.

3.1.3. Image Cropping

To maintain visual consistency and ensure the model's attention remained on relevant regions, frames were cropped into 1080x1080-pixel squares 2 centered on the bounding boxes around the boxers. This was accomplished using `crop_images.py`. Cropping standardized the input size while preserving essential spatial features related to punches.

Figure 2. Example cropped frame

3.1.4. Image Scaling

The cropped images were resized to 180x180 pixels using `scale_images.py` to fit the CNN's input size. This resizing reduced computational complexity while retaining critical visual details necessary for punch detection. The scaled images provided a balance between data efficiency and model performance. In the future, it may be considered to train the model with higher resolutions such as 360x360 or 540x540.

By following this preprocessing pipeline, the dataset was effectively prepared for training and evaluating the punch detection model, ensuring high-quality input for robust performance.

3.2. Model Training

The model used for punch detection in boxing images is a Convolutional Neural Network (CNN), designed for multi-class classification. The architecture and training process are detailed below:

3.2.1. Model Architecture

The CNN model consists of the following layers 3:

- **Conv2D Layers**: Three convolutional layers with ReLU activation. These layers use 3x3 kernels and progressively increase the number of filters (90 and 180 filters) to extract hierarchical features from the input images.
- **MaxPooling2D Layers**: After each convolutional layer, max pooling is applied with a pool size of 2x2 to reduce the spatial dimensions of the feature maps. This helps in reducing computation while retaining important features.
- **Dropout Layer**: A dropout layer with a rate of 0.5 is applied after the last convolutional layer to prevent overfitting by randomly setting half of the activations to zero during training.
- **Flatten Layer**: The feature maps are flattened into a 1D vector, which is passed into the fully connected layers.
- **Dense Layers**: The dense layers consist of 180 units with ReLU activation, followed by the final output layer with 8 units and softmax activation. The softmax activation is used to output probabilities for the eight different punch types.

3.2.2. Model Compilation

The model was compiled with the following configurations:

- **Optimizer**: Adam optimizer, chosen for its efficiency in handling large datasets.
- **Loss Function**: Sparse categorical cross-entropy, appropriate for multi-class classification tasks where labels are integer-encoded.
- **Metrics**: Accuracy is tracked during training to evaluate the model's performance.

3.2.3. Training Process

The training process is carried out on a GPU(NVIDIA Tesla P100) to accelerate the training. The dataset is split into training and validation sets, with 80% of the data used for training and the remaining 20% for validation. The model is trained for 10 epochs with a batch size of 32.

The model is trained using the `fit()` method, where the training data (`X_train`, `y_train`) is used to update the model weights, while the validation data (`X_val`, `y_val`) is used to monitor the model's performance during training. The training progress, including loss and accuracy, is recorded in the `history` object.

4. Experimental Evaluation

4.1. Dataset

The dataset, as detailed in (Stefański et al., 2024), consists of annotated boxing footage recorded in Poland. After preprocessing, it includes 180x180 images and bounding box labels for eight action classes. This balanced dataset is essential for training and evaluating the CNN model.

4.2. Model Training

A CNN with the architecture shown in Figure 3 was implemented. The model uses sparse categorical cross-entropy as the loss function, given by:

$$L = -\frac{1}{N}\sum_{i=1}^{N}\log(p_{i,y_i}), \tag{1}$$

where p_{i,y_i} represents the predicted probability for the correct class.

4.3. Results

Training achieved a final validation accuracy of 83.94% and a loss of 0.645. Table 1 summarizes the precision, recall, and F1-scores in all classes, highlighting strong performance in detecting punches and misses.

Class	Precision	Recall	F1-Score
Head with left hand	0.84	0.86	0.85
Head with right hand	0.86	0.84	0.85
Body with left hand	0.66	0.72	0.79
Body with right hand	0.71	0.64	0.68
Block with left hand	0.89	0.86	0.88
Block with right hand	0.91	0.91	0.91
Miss with left hand	0.85	0.86	0.86
Miss with right hand	0.83	0.78	0.80
Overall Accuracy		0.84	

Table 1. Classification Report

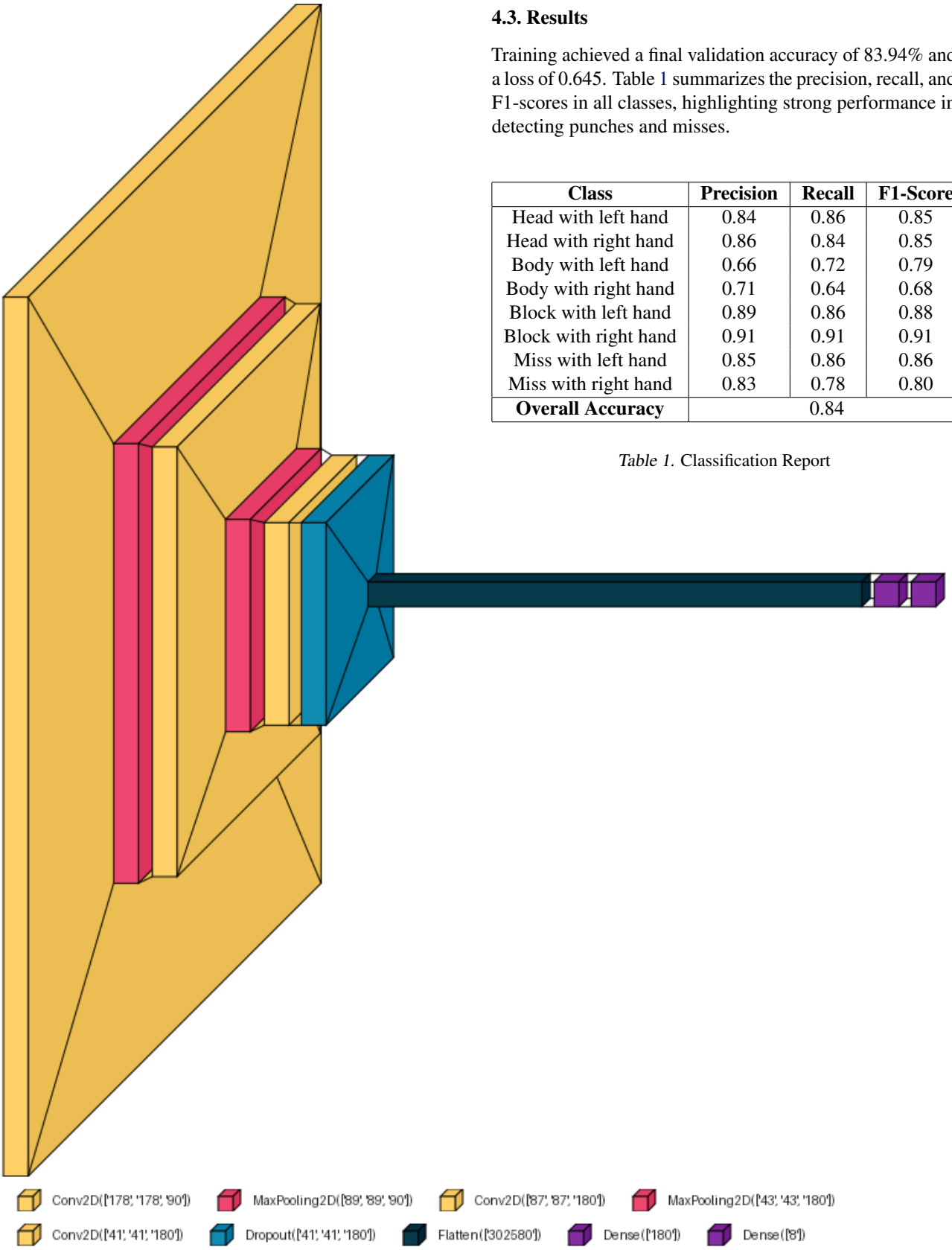

Figure 3. CNN Architecture used for punch classification.

4.4. Visual Results

Figure 4 shows the confusion matrix, providing insight into misclassification patterns.

Figure 4. Confusion Matrix

5. Conclusion

This paper presents a novel and effective approach to detect and classify punches in Olympic boxing matches using a convolutional neural network (CNN) model. Using a single static camera setup and a well-structured preprocessing pipeline, the system effectively identifies eight distinct punch types, including head & body punches, blocks, and misses. Our experimental evaluation shows that the proposed method achieves an overall accuracy of 84%. These results underscore the potential of automated punch detection systems for real-time analysis in boxing, offering significant benefits to coaches, referees, and analysts by providing detailed performance insights and enhancing the overall viewing experience.

Future work can focus on further improving model generalization across different boxers, weight classes, and styles of fight. Additionally, integrating more advanced techniques such as temporal analysis and multi-camera setups could help capture more complex interactions in boxing matches. The scalability and robustness of the proposed method make it a promising tool for further exploration in the domain of sports analytics, not only for boxing but also for other combat sports.

References

Worsey, M.T.O.; Espinosa, H.G.; Shepherd, J.B.; Thiel, D.V. An Evaluation of Wearable Inertial Sensor Configuration and Supervised Machine Learning Models for Automatic Punch Classification in Boxing. *IoT*, **2020**, *1*, 360–381.

Worsey, M.T.O.; Espinosa, H.G.; Shepherd, J.B.; Thiel, D.V. Inertial Sensors for Performance Analysis in Combat Sports: A Systematic Review. *Sports*, **2019**, *7*, 28.

Ye, X.; Shi, B.; Li, M.; Fan, Q.; Qi, X.; Liu, X.; Zhao, S.; Jiang, L.; Zhang, X.; Fu, K.; et al. All-textile sensors for boxing punch force and velocity detection. *Nano Energy*, **2022**, *97*, 107114.

Kasiri-Bidhendi, S.; Fookes, C.; Morgan, S.; Martin, D.T.; Sridharan, S. Combat sports analytics: Boxing punch classification using overhead depth imagery. In *Proceedings of the 2015 IEEE International Conference on Image Processing (ICIP)*, Quebec City, QC, Canada, 27–30 September 2015.

Kasiri-Bidhendi, S.; Fookes, C.; Sridharan, S.; Morgan, S. Fine-grained action recognition of boxing punches from depth imagery. *Comput. Vis. Image Underst.*, **2017**, *159*, 143–153.

Stefański, P.; Kozak, J.; Jach, T. Boxing Punch Detection with Single Static Camera. *Entropy*, **2024**, *26*, 617.

Behendi, S.K.; Morgan, S.; Fookes, C.B. Non-Invasive Performance Measurement in Combat Sports. In *Proceedings of the 10th International Symposium on Computer Science in Sports (ISCSS)*, Springer International Publishing: Cham, Switzerland, 2015.

Wattanamongkhol, N.; Kumhom, P.; Chamnongthai, K. A method of glove tracking for amateur boxing refereeing. In *Proceedings of the IEEE International Symposium on Communications and Information Technology, ISCIT 2005*, Beijing, China, 12–14 October 2005.