電機資訊學院 2026 作 BRAIN PLUS HAND 實作專題競賽

Automatic Player Occurrence Identification in Sports Videos Using Audio-Based Recognition and Visual Face Matching

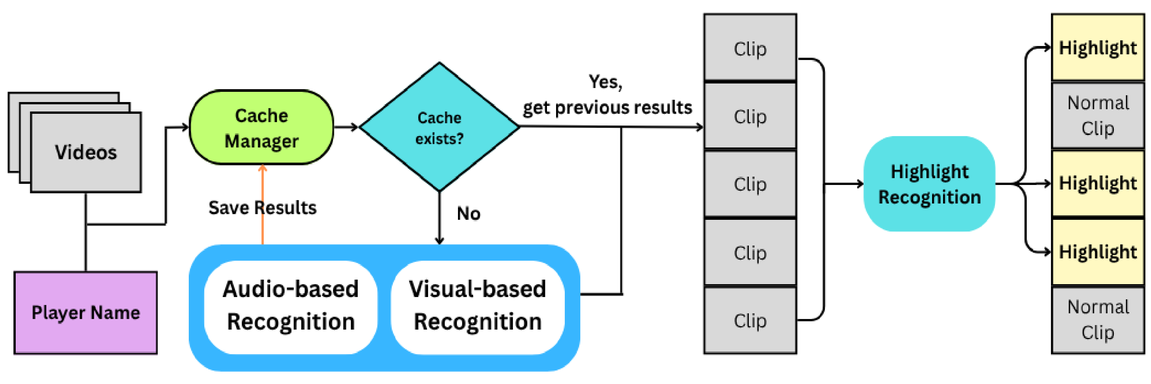
Author: 張立誠 (110006161, 電資班國際學程) 隊伍編號EECS10

INTRODUCTION

In sports broadcasting, locating all appearances of a specific player within a game remains largely a manual, time-consuming process. Broadcasters such as ELTA TV must scrub through hours of footage to find relevant segments, identify when a player was mentioned or appeared on screen, and assemble highlight reels. This manual workflow is slow, error-prone, and difficult to scale—particularly across large multi-game datasets. The goal is therefore to automatically detect and timestamp each player's appearances in sports videos using reliable multi-modal cues.

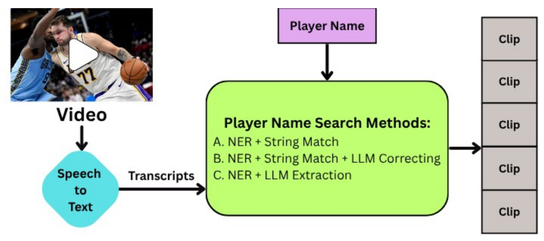
METHODOLOGY AND IMPLEMENTATION

System Overview:



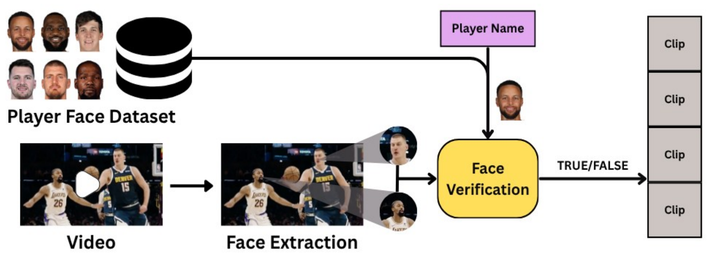
The system takes several videos and a target player name, checks a cache for existing results, and returns them if available. Otherwise, it runs the missing modules—audio-based recognition and/or visual face matching—to locate all player appearances. These candidate clips are then passed to a highlight classifier, which identifies which segments qualify as highlights.

• Audio-based Pipeline:



The video is transcribed into timestamped text, which—together with the target player name—is processed by one of three matching methods (A, B, or C). The selected method returns all audio hits, producing the final set of player-related clip segments.

• Visual-based Pipeline:



The video is processed by a face-extraction model, and each detected face is compared against the player's face dataset. Verified matches are treated as hits, and the system outputs all clip intervals where the player's face appears.

RESULTS

Table 1 Performance of Methods A–C on *Hawks vs Pacers* (highlight video).

Method	Hits	Acc ↑	FP ↓	Precision ↑
A	82/93	90.6%	0	**100%**
B	93/93	**100%**	18	83%
C	77/93	86.2%	0	**100%**

Table 1. result of audio : Three methods (A–C) were tested on NBA highlights and full-game datasets. In highlights, Method B hit 100% accuracy but had 18 false positives; Method A had 90.6% accuracy, 0 false positives, and 1.76s runtime. In full games, Method C achieved balanced results, lying in between Method A and B. LLM used: gemma3-4b

Table 2 Verification accuracy with thresholding (overall / profile / frontal).

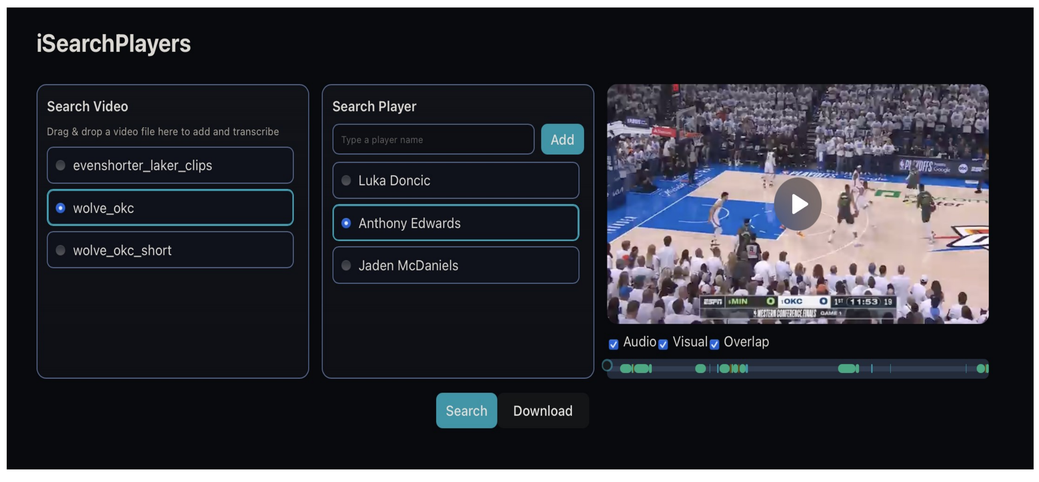
Model	Total	Profile	Frontal
ArcFace	**0.35**	0.19	**0.66**
VGG-Face	0.31	0.16	0.60
Dlib	0.27	0.12	0.56
Facenet	0.12	0.03	0.31
SFace	0.12	0.05	0.26
Facenet512	0.10	0.04	0.23
GhostFaceNet	0.10	0.02	0.26
DeepFace	0.02	0.00	0.05
DeepID	0.00	0.00	0.00
OpenFace	0.00	0.00	0.00

Table 2 & 3. result of visual : ArcFace outperformed other models (0.35 total, 0.66 frontal accuracy at threshold 0.68), and further improved to 0.53 total / 0.81 frontal without threshold, confirming its robustness for player matching.

Table 3. Verification accuracy without thresholding (nearest-neighbor).

Model	Total	Profile	Frontal
ArcFace	**0.53**	**0.38**	**0.81**
Facenet512	0.49	0.34	0.79
VGG-Face	0.45	0.32	0.69
Dlib	0.42	0.24	0.77
SFace	0.41	0.33	0.56
Facenet	0.27	0.14	0.52
GhostFaceNet	0.19	0.08	0.40
OpenFace	0.06	0.04	0.11
DeepFace	0.10	0.08	0.15
DeepID	0.06	0.09	0.00

USER INTERFACE



The interface uses a Django back end and a React.js front end. Users can drag and drop videos, select or add players, and run a search to retrieve all detected audio and visual occurrences. A download option generates a ZIP file containing all matched hits.

CONCLUSION

This work developed a player detection system integrating audio- and visual-based recognition with highlight detection. The system showed reliable performance on test videos, demonstrating the potential of multimodal fusion in sports analytics. However, transcription accuracy and face verification can be further improved, and the highlight detection model still lacks sufficient training data. Future work will focus on expanding the dataset and refining the algorithms for better robustness and accuracy.