ASPERA: Exploring Multimodal Action Recognition in Football Through Video, Audio, and Commentary

Takane Kumakura, Ryohei Orihara, Yasuyuki Tahara, Akihiko Ohsuga and Yuichi Sei
The University of Electro-Communications, Graduate School of Informatics and Engineering Departments, Department of Informatics 1-5-1 Chofugaoka, Chofu, Japan
kumakura.takane@ohsuga.lab.uec.ac.jp, orihara@acm.org, {tahara, ohsuga, seiuny}@uec.ac.jp

Keywords: Action Spotting, Multimodal Learning, Transformer, Markov Chain, Soccer, Football, Live Broadcasting, Deep Learning, Machine Learning, Artificial Intelligence.

Abstract: This study proposes ASPERA (Action SPotting thrEe-modal Recognition Architecture), a multimodal football action recognition method based on the ASTRA architecture that incorporates video, audio, and commentary text information. ASPERA showed higher accuracy than models using video and audio only, excluding invisible actions in the video. This result demonstrates the advantage of this multimodal approach. Additionally, we propose three advanced models: $\text{ASPERA}_{\text{srnd}}$ incorporating surrounding commentary text within a ±20-second range, $\text{ASPERA}_{\text{cln}}$ removing irrelevant background information, and $\text{ASPERA}_{\text{MC}}$ applying a Markov head to provide prior knowledge of football action flow. $\text{ASPERA}_{\text{srnd}}$ and $\text{ASPERA}_{\text{cln}}$, which refine the text embedding, enhanced the ability to accurately identify the timing of actions. Notably, $\text{ASPERA}_{\text{MC}}$ with the Markov head demonstrated the highest accuracy for invisible actions in the football video. $\text{ASPERA}_{\text{srnd}}$ and $\text{ASPERA}_{\text{cln}}$ not only demonstrate the utility of text information in football action spotting but also highlight key factors that enhance this effect, such as incorporating surrounding commentary text and removing background information. Finally, $\text{ASPERA}_{\text{MC}}$ shows the effectiveness of combining Transformer models and Markov chains for recognizing actions in invisible scenes.

1 INTRODUCTION

Football, also known as soccer in some countries, is popular worldwide, with football clubs existing in 135 countries (FIFA, 2024). Additionally, the 2022 World Cup attracted approximately 1.63 million spectators, and according to Mordor Intelligence (Intelligence, 2024), the football market size is estimated to reach USD 741.45 million in 2024. Due to football's immense popularity, research in sports analytics focusing on understanding and analyzing player movements and situations in the game has become increasingly active in recent years. Sports analytics is utilized in various applications, such as team strategy development, player performance evaluation, scouting, referee's decision, and highlight generation. For example, manually creating video summaries requires trimming and editing approximately 90 minutes of video from both halves, demanding significant time and effort. Therefore, enabling automated generation not only reduces time and effort but also allows for efficient tactical reviews and immediate video delivery. As a result, technologies for automatically recognizing player actions from broadcast videos have become a highly active research area.

In this study, we address Action Spotting(Deliège et al., 2021), a Temporal Action Detection (TAD) task that identifies the temporal occurrence of specific actions within football videos. As shown in Figure 1, the aim is to estimate the exact moments when actions such as *Goals*, *Corner Kicks*, and *YCs* occur.

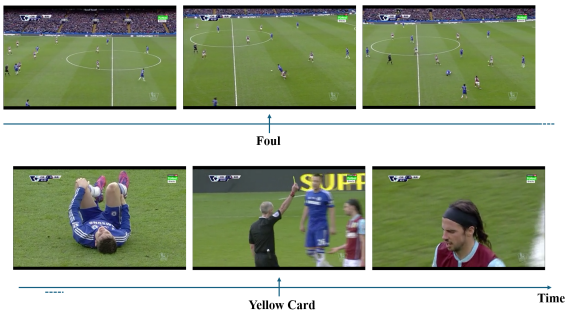

Figure 1: Example of Action Spotting.

In existing research, many methods have been proposed to tackle Action Spotting, mainly with vi-

Kumakura, T., Orihara, R., Tahara, Y., Ohsuga, A. and Sei, Y.
ASPERA: Exploring Multimodal Action Recognition in Football Through Video, Audio, and Commentary.
DOI: 10.5220/0013300700003890

In *Proceedings of the 17th International Conference on Agents and Artificial Intelligence (ICAART 2025) - Volume 2*, pages 646-657
ISBN: 978-989-758-737-5; ISSN: 2184-433X

sual features extracted from broadcast football videos. However, due to the presence of replays and the difference in camera angles, some actions remain unrecorded in the video. Actions that occur in the actual game but missing from the video are defined as invisible actions, and the challenge lies in their lower detection accuracy compared to visible ones. Therefore, recent studies aim to improve the accuracy of invisible actions by utilizing modalities such as graphs(Cartas et al., 2022) and audio (Gan et al., 2022; Shaikh et al., 2022; Vanderplaetse and Dupont, 2020; Xarles et al., 2023).

For example, Cartas et al.(Cartas et al., 2022) represented football players and referees as nodes in a graph, modeling their temporal interactions as a sequence of graphs to focus on game actions. This approach was motivated by their observation that match video captures not only the field but also spectators in the stadium, coaches on the bench, and replays. Additionally, Gan et al.(Gan et al., 2022) proposed a Transformer-based(Vaswani et al., 2017) multimodal football scene recognition method for both visual and auditory modalities. Video frames are fed into a Visual Transformer, and audio spectrograms are fed into an Audio Transformer. By performing late fusion on the estimation results from both transformers, they handle both visual and auditory modalities. Furthermore, VanderPlaetse et al.(Vanderplaetse and Dupont, 2020) set a ResNet pre-trained on ImageNet as the Visual Stream and VGGish(Hershey et al., 2017) pre-trained on AudioSet as the Audio Stream. After extracting features from each stream, they fused the two modalities using seven methods and compared the results. Similarly, Xarles et al.(Xarles et al., 2023) proposed a multimodal approach that utilizes audio and visual modalities. They extracted features from audio log-Mel spectrograms using VGGish and merged them with visual features. These combined features were then used as input to a Transformer Encoder.

However, although these methods pay attention to the excitement, atmosphere, and voices of the audience and commentators, they mainly emphasize the acoustic properties of the audio over the actual content of the commentary. The reason is that the SoccerNet-v2(Deliège et al., 2021) dataset, which is used in many existing studies on Action Spotting, has significant language variations between match videos, and some videos have no commentary at all, as shown in Table 1. Since audio features vary considerably across different languages, the model learns features based on acoustic characteristics rather than the commentary content. This relationship has been indirectly demonstrated by previous studies, where adding audio features to video-only models improved accuracy. However, these improvements can be attributed to the acoustic properties of the audio rather than the semantic content of the commentary.

Table 1: We identified the language breakdown of the SoccerNet-v2 dataset using FasterWhisper in this study.

language	English	Spanish	Russian	German	French	Norwegian Nynorsk	Italian	Turkish
train	185	153	112	73	60	7	3	2
valid	71	46	38	18	20	0	0	2
test	59	55	42	30	8	2	2	0
challenge	26	14	26	16	14	0	0	0
total	341	268	218	137	102	9	5	4

language	Korean	Polish	Latin	Welsh	Māori	Croatian	Hungarian	None
train	0	1	1	2	0	0	0	1
valid	1	0	0	1	1	0	0	2
test	0	0	0	0	0	0	0	2
challenge	0	0	0	0	0	2	2	0
total	1	1	1	3	1	2	2	5

Moreover, there are two types of actions regarding commentary in the match: (1) actions that are difficult to identify without commentary, and (2) actions that are clear from video alone; however, commentary helps improve accuracy.

For type (1), even when one *Yellow Card* is shown in the video, there are cases where *Yellow Cards* are issued to players from both teams. Additionally, although only a *Red Card* is displayed in the video, there are instances where two *Yellow Cards* are given, leading to a *Red Card (YC→RC)*. In these situations, the commentary may include expressions like "*Yellow Card* to [Player], and to [Player]", indicating that cards are given to both players or "Second *Yellow Card*, so *Red*!" indicating a *YC→RC*.

For type (2), the commentator may still exclaim "*Goal*!!" even when a *Goal* is apparent in the video. While such events can be identified from video alone, incorporating commentary information can lead to more accurate predictions.

Based on these challenges and observations of match conditions, this study proposes ASPERA (Action SPotting thrEe-modal Recognition Architecture), a multimodal football action recognition method that leverages the Transformer-based ASTRA, Action Spotting TRAnsformer for Soccer Videos(Xarles et al., 2023), model on three modalities: visual information, audio information, and textual information from the commentary. By incorporating the commentary as an additional source of information, the method aims to enhance accuracy in action spotting, with a particular emphasis on recognizing invisible actions. In this study, ASPERA is used as a baseline model for three advanced models, and these models with improvements over ASPERA are introduced as $\text{ASPERA}_{\text{srnd}}$, $\text{ASPERA}_{\text{cln}}$, and $\text{ASPERA}_{\text{MC}}$.

First, in ASPERA, text segments annotated with the start and end times of speech are generated when extracting the commentary text from audio data. However, each generated text segment may contain irrelevant information. For example, this could include cases where audience cheers are transcribed as text or where commentators' predictions are captured. Moreover, as shown in Figure 2, a single text segment strongly related to a non-occurring action—such as a *Free-kick*—can lead to misrecognition. Even if the surrounding commentary pertains to an actually occurring *Yellow Card*, the non-occurring *Free-kick* may be recognized instead of the *Yellow Card*. Therefore, the model reduces the influence of text segments at specific moments by incorporating commentary text from the surrounding 20-second time frame. This is introduced as ASPERA$_{\text{srnd}}$.


```
    {
        "start": 1931.02,
        "end": 1935.38,
        "text": " He said earlier on he didn't think Mourinho would wait
too long,"
    },
    {
        "start": 1935.38,
        "end": 1938.46,
        "text": " he's been known to react to him being early in
games."
    },
    {
        "start": 1942.06,
        "end": 1943.26,
        "text": " That's probably..."
    },
    {
        "start": 1943.26,
        "end": 1947.14,
        "text": " Well, Pellegrini will feel an unnecessary free-kick to
give away,"
    },
    {
        "start": 1947.14,
        "end": 1952.18,
        "text": " so soon after the goal, and to collect a booking for
Vincent Kompany as well."
    },
    {
        "start": 1952.18,
        "end": 1954.82,
        "text": " He needs to regain his composure this season,
Kompany."
```


Figure 2: An example of a generated text segment. Although a *Yellow Card* was issued in 1946s, there is a mention of a *Free-kick*—which was actually absent—between 1943.26s and 1947.14s.

Furthermore, the commentator's utterance at each second often contains background information unrelated to the actions during the match. Therefore, irrelevant background information, such as coaches' comments and recent match records, was excluded from the analysis. This is introduced as ASPERA$_{\text{cln}}$.

Then, in existing research (Xarles et al., 2023), when observing the recognition results of models using only visual information and models utilizing visual and audio information, flows of actions that were absent in the dataset—such as a *Direct Free-kick* leading to a *Red Card*—were recognized. To address this issue, this study aims to improve accuracy by providing prior knowledge of action flows—such as *Throw-in* often following a *Ball out of play*, *Kick-off* often following a *Goal*, and *Indirect Free-kick* often following an *Offside*—as a Markov chain. Specifically, in cases with invisible scenes, the model is expected to improve the recognition accuracy of invisible actions by providing prior knowledge on which action is likely to occur based on past or subsequent actions. This is introduced as ASPERA$_{\text{MC}}$.

2 RELATED WORKS

2.1 ASTRA

Xarles et al.(Xarles et al., 2023) proposed ASTRA, a Transformer-based model designed for Action Spotting. ASTRA achieved the third-highest score on the Challenge Set of the SoccerNet 2023 Action Spotting Challenge. ASTRA tackles the challenge of lower accuracy in invisible actions compared to visible actions by combining audio modality with video modality to improve performance on invisible actions.

For the video modality, Baidu Soccer Embeddings (Zhou et al., 2021) are fed into a Position-wise Feed-Forward Network (PFFN), which enables parallel processing at each frame. For the audio modality, audio features are extracted using VGGish (Hershey et al., 2017), pre-trained on AudioSet (Gemmeke et al., 2017), from log-Mel spectrograms. The feature-aligned video and audio embedding are then concatenated and processed through a Transformer Encoder-Decoder architecture. The resulting embedding is fed into two heads: (1)a classification head (Λ_s) for temporal position classification and (2) an uncertainty-aware displacement head (Λ_d) for prediction refinement.

2.1.1 Classification Head

The classification head processes the decoder's query outputs at 0.5-second intervals through a sequence of operations: two linear layers followed by ReLU acti-

vation, and finally a sigmoid function. This generates per-class probability scores every 0.5 seconds, indicating the likelihood of each action occurring. This 0.5-second interval is referred to as the *feature clock*.

2.1.2 Uncertainty-Aware Displacement Head

The uncertainty-aware displacement head processes the decoder queries through two linear layers with ReLU activation, followed by parallel linear layers that generate estimated means and variances for each action class. This architecture models displacement as a Gaussian distribution, capturing temporal uncertainty in the predictions.

The estimated displacements serve to refine the classification head's predictions by probabilistically adjusting the temporal locations for each *feature clock*.

2.2 Baidu Soccer Embeddings

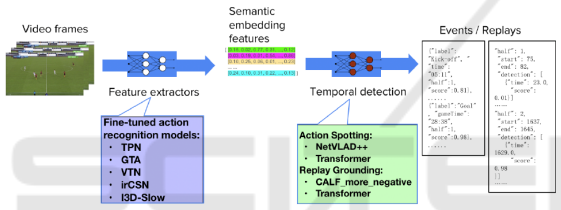

Figure 3: The architecture of the two-stage approach by Zhou et al.(Zhou et al., 2021).

Zhou et al.(Zhou et al., 2021) proposed a two-stage framework for event detection in football broadcast videos, as illustrated in Figure 3. The first stage employs multiple fine-tuned action recognition models—TPN(Yang et al., 2020), GTA(He et al., 2020), VTN(Neimark et al., 2021), irCSN(Tran et al., 2019), and I3D-Slow(Feichtenhofer et al., 2019)—to extract high-level semantic features. The second stage utilizes NetVLAD++(Giancola and Ghanem, 2021) and a Transformer as temporal detection modules for Action Spotting.

2.3 Whisper and FasterWhisper

Radford et al.(Radford et al., 2023) introduced a weakly supervised sequence-to-sequence Transformer model trained on large-scale internet audio data. The model was trained on multiple audio processing tasks, including multilingual speech recognition, speech translation, speaker identification, and voice activity detection. FasterWhisper(SYSTRAN, 2024), an optimized implementation of the Whisper architecture, enhances performance through the CTranslate2 inference engine, 8-bit quantization, and various optimization techniques. In GPU-based processing, FasterWhisper demonstrates up to 4x speedup compared to the original Whisper model, while simultaneously achieving significant reductions in both GPU and CPU memory consumption.

2.4 Markov Chain

Markov chains are used to model probabilistic processes where the states of a system change discretely. A key feature is the "Markov property", which means that the next state depends only on the current state, without influence from any previous states. This allows complex state transitions to be represented using simple transition probabilities. For example, in a football match, immediately following a *Goal*, the next action is a *Kick-off* with approximately 87% probability according to the SoccerNet dataset. This transition illustrates the Markov property, where the next action, *Kick-off* is determined solely by the current state of *Goal*, independent of past events. Thus, the flow of a football match can be represented by a simple state transition matrix, as shown in Table 2, expressing the transition probabilities between actions throughout the game.

A transition probability matrix enables the numerical representation of the probability that each action will transition to the next during a match. This approach captures the system's dynamic behavior within a Markov chain framework. In recent years, the concept of Markov chains has been incorporated into deep learning and utilized as an auxiliary method to enhance the predictive capabilities of models.

For instance, Markov chain-based methods have been proposed for continuous action recognition. Lei et al.(Lei et al., 2016) proposed a hybrid architecture that combines convolutional neural network(CNN) and Hidden Markov Model(HMM) to model the statistical dependencies between neighboring sub-actions. This method leverages CNN's high-level feature learning capabilities to extract action features and uses HMM to model the transitions between these features.

Furthermore, the concept of Markov chains has been applied in Transformer models as well. Zhang et al.(Zhang and Feng, 2023) proposed the Hidden Markov Transformer (HMT), which models translation initiation timing as a hidden Markov model in simultaneous machine translation (SiMT) tasks. By selecting the optimal start point from multiple candidate timings, they achieved high-accuracy simultaneous translation.

Table 2: A portion of the transition probability matrix between actions during a football match.

Before \ After	Penalty	Kick-off	Goal	Substitution	...	Total
Penalty	0	0	0.361	0	...	1
Kick-off	0	0.019	0.004	0.021	...	1
Goal	0	0.872	0.019	0.067	...	1
Substitution	0	0.091	0	0.136	...	1
Offside	0	0.002	0	0.058	...	1
...	...	...	...	...	...	...

3 PROPOSED METHOD

We propose ASPERA (Action SPotting thrEe-modal Recognition Architecture), a Transformer-based multimodal football action recognition method that utilizes three modalities: visual information, audio information, and textual commentary content. We obtained English commentary text for all match videos in the dataset by transcribing the broadcast audio of SoccerNet using FasterWhisper and translating non-English content using GPT-3.5 Turbo(OpenAI, 2024a). Next, embedding generated using Text Embedding Large 3(OpenAI, 2024b) for the commentary text was incorporated into the existing ASTRA model(Xarles et al., 2023) for training.

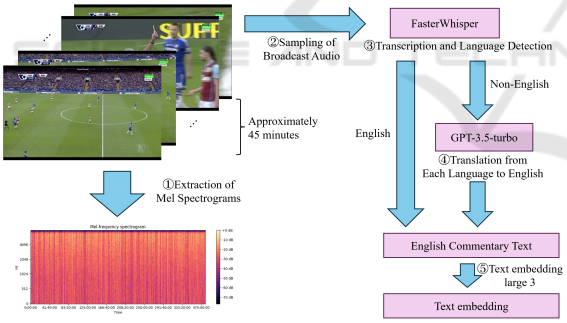

Figure 4: Data preparation flow for the audio and text modalities.

First, the data preparation flow for Mel spectrograms and text embedding is shown in Figure 4. Then, ASPERA based on ASTRA is indicated in the green dotted line portion of Figure 5.

3.1 Creation of the Audio Modality Dataset

To perform training using audio information, Mel spectrograms were created from the broadcast audio of SoccerNet-v2. Specifically, the sampling rate of the original audio files was set to 16,000 Hz, and the

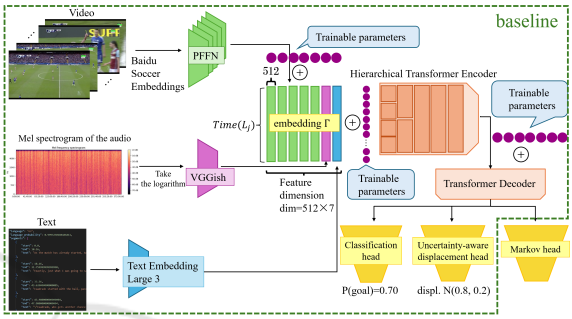

Figure 5: ASPERA adds a text modality to the ASTRA model.

audio channels were set to mono to reduce the size of the audio data and improve processing efficiency.

Next, to align the sequence length of Mel spectrograms with that of the broadcast video, the audio files were downsampled to a rate of 100. To compute Mel spectrograms, the parameters for the Short-Time Fourier Transform were set as follows: n_{fft} was set to 512, hop_{length} was determined by the following equation (1), and the number of Mel filter banks N_{mels} was set to 128. In this context, $len(y)$ denotes the sequence length of the broadcast video. The power spectrogram was then converted to decibel units to facilitate numerical processing. For segments of broadcast audio with missing or interrupted sound, zeros were added to the audio files for consistent processing.

$$hop_{length} = \left\lfloor \frac{len(y) - n_{fft}}{len(y) - 1} \right\rfloor + 1 \quad (1)$$

These procedures were applied to all match videos in the SoccerNet-v2 dataset, creating the dataset for the audio modality by extracting Mel spectrograms.

3.2 Creation of the Text Modality Dataset

The ASTRA model was extended to incorporate commentary content by converting audio commentary into text embedding. Initially, the audio data was

processed with the same steps as the audio modality dataset to reduce its size and enhance processing efficiency.

Audio transcription and language detection were performed using FasterWhisper, with particular emphasis on preserving temporal information. The transcription data was stored with timestamps marking the beginning and end of each utterance, alongside their corresponding spoken text segments. This format enables precise temporal tracking of all utterances. For non-English matches identified by FasterWhisper's language detection, translation was performed using the GPT-3.5 Turbo(OpenAI, 2024a) API. To preserve temporal alignment, translations were processed sequentially for each text segment. The translation prompt specified the football commentary context and included language-specific examples.

In this study, we extracted each transcribed text segment at one-second intervals if that second falls between the beginning and end times of the utterance, to enhance temporal accuracy. Each one-second interval is defined as a *text clock*, and the transcribed text segment at each *text clock* is referred to as *sec-text*. Then, for each *sec-text*, we obtained the text embedding using Text Embedding Large 3(OpenAI, 2024b) and considered them as *sec-text embedding*. Finally, by obtaining and concatenating a D-dimensional *sec-text embedding* over approximately 45 minutes, or around 2700 seconds, we obtained a total of approximately $2700 \times D$ text embedding for each match.

3.3 Training

ASPERA proposed in this study extends ASTRA by adding a text modality. Specifically, the main modification to the ASTRA model involves fusing text embedding with other modalities before the Transformer Encoder. The architecture is shown in Figure 5. The Baidu Soccer Embeddings were divided into five parts along the feature dimension and passed through a Position-wise Feed-Forward Network (PFFN) to obtain video embedding. The spectrograms obtained in Section 3.1 were converted into log-Mel spectrograms and passed through the VGGish model to obtain audio embedding. The approximately $2700 \times d$ text embedding obtained in Section 3.2 was then concatenated with these before the Hierarchical Transformer Encoder. This allows the model to learn dependencies between different modalities.

Additionally, trainable temporal positional embedding and feature positional embedding were added to the text embedding before concatenation. This enables the model to learn considering temporal and feature positional information.

3.4 The Refinement of ASPERA

We propose three advanced models that extend ASPERA.

First, ASPERA$_{\text{srnd}}$ uses text embedding which adds the text embedding of the commentary text from 20 seconds before and after each *text clock* to the *sec-text embedding*. Next, ASPERA$_{\text{cln}}$ excludes background information unrelated to actions, such as coaches' comments and match records, from ASPERA$_{\text{srnd}}$. Then, ASPERA$_{\text{MC}}$ introduces a Markov head to incorporate prior knowledge of action flows to ASPERA$_{\text{MC}}$.

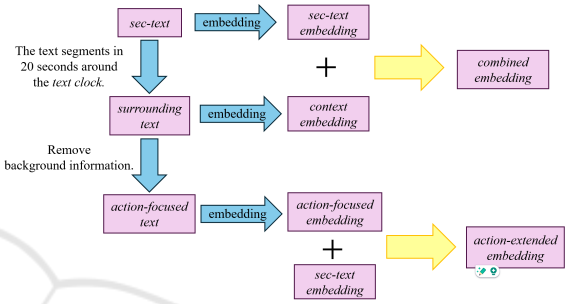

Figure 6: The text embeddings of ASPERA, ASPERA$_{\text{srnd}}$, ASPERA$_{\text{cln}}$ and ASPERA$_{\text{MC}}$.

3.4.1 ASPERA$_{\text{srnd}}$

In ASPERA$_{\text{srnd}}$, we combined the transcribed texts whose text clock within ±20 seconds fell between their start and end times, designating this as the *surrounding text*, and used Text Embedding Large 3 to create the *context embedding*. By adding this *context embedding* to the *sec-text embedding*, we obtained the *combined embedding*, as shown in Figure 6.

3.4.2 ASPERA$_{\text{cln}}$

ASPERA$_{\text{cln}}$ excluded background information unrelated to actions. The *surrounding text* from ASPERA$_{\text{srnd}}$ was provided to GPT-4o mini with the following instruction:

"The following is a football commentary. Please remove any text unrelated to the Action Spotting task, which involves identifying the occurrence of specific actions in a match, such as team records and player achievements. Condense the remaining text into a single sentence containing only the important information related to the match progress and actions".

This process allowed us to obtain the *action-focused text*. Then, the *action-focused text* was used to create an *action-focused embedding* by utilizing Text Embedding Large 3. Similar to ASPERA$_{\text{srnd}}$, by adding this *action-focused embedding* to the *sec-text*

embedding, we obtained the *action-extended embedding*, as shown in Figure 6.

3.4.3 ASPERA$_{MC}$

In ASPERA$_{MC}$, we considered the action flow by introducing a Markov head to the output of the Transformer Decoder in our proposed model shown in Figure 5. The Markov head outputs the confidence of each action at each *feature clock*, similar to the classification head. The actions are 18 types of actions including the background class representing no action. Considering only the transitions of actions that exceeded a threshold at each *feature clock*, we designed the loss function as shown in equations 2 and 3.

$$\text{lossM} = (\text{class}_{\text{before}} + \text{class}_{\text{after}}) \cdot (1 - \text{trans}_{\text{prob}})^2 \quad (2)$$

$$\text{loss} = \text{lossC} + \text{lossD} + \lambda \cdot \text{lossM} \quad (3)$$

$trans_{prob}$ represents the transition probability matrix between actions during a football match, as shown in Table 2. $class_{before}$ denotes the occurrence frequency of the action that occurred immediately before, and $class_{after}$ denotes the occurrence frequency of the current action. $(1 - trans_{prob})^2$ reduces the loss for transitions between actions with high transition probabilities and increases the loss for transitions between actions with low transition probabilities. Additionally, Multiplying with $(class_{before} + class_{after})$ increases the loss or transitions between frequently occurring actions, despite their low transition probabilities. Furthermore, the loss in ASPERA$_{MC}$ was designed by adding λ proportion of *lossM* to the loss obtained by summing *lossC* from the classification head and *lossD* from the uncertainty-aware displacement head, as proposed in ASTRA(Xarles et al., 2023).

4 EVALUATION

4.1 Dataset

This study uses SoccerNet-v2 as the dataset for action spotting in football. SoccerNet-v2 is a dataset consisting of 550 football matches held from 2014 to 2017 in the Premier League, UEFA Champions League, Ligue 1, Bundesliga, Serie A, and La Liga. It has been provided for various tasks.

For action spotting, annotations for 17 types of football actions are publicly available for 500 matches, while annotations for the remaining 50 matches are accessible only to the organizers as a challenge dataset. Out of the 500 matches, 300 are designated as training data, 100 as validation data, and 100 as test data, and in this study, the test data is evaluated according to the metrics outlined in Section 4.3.

4.2 Implementation Details

The hyperparameters of the model are determined according to the ASTRA settings. Specifically, the model implementation uses PyTorch, and the Adam optimizer is applied. The initial learning rate is set to 5×10^{-5} with an initial warm-up of three epochs, followed by cosine decay over 50 epochs. This model uses 50-second clips with an embedding dimension of $d = 512$ as input. Differences from ASTRA in this model include the number of embedding $\mathcal{E}$, the embedding dimension of the text embedding, and the positional embedding for the text embedding. For the embedding, we used a total of $|\mathcal{E}| = 7$, comprising five Baidu Soccer Embeddings for visual data, one audio embedding obtained by passing log-Mel spectrogram through VGGish for audio data, and one text embedding for textual data. The embedding dimension for the text embedding is set to $d = 512$.

4.3 Evaluation Metrics

Average-mAP was used as the evaluation metric for this method. This metric quantifies the area under the mAP curve for different tolerance values, denoted by δ. The mAP represents the mean Average Precision across all action classes. Average Precision is a summarized value of the Precision-Recall curve, where precision is plotted on the vertical axis and recall on the horizontal axis. In action spotting, the detection results must match the ground truth within a specific time range, which is why different tolerance values δ are set. SoccerNet adopts the metrics of tight Average-mAP and loose Average-mAP for Average-mAP. The metric of tight Average-mAP uses a δ range of 1 to 5 seconds, while loose Average-mAP uses a δ range of 5 to 60 seconds. This study also used tight Average-mAP and loose Average-mAP for evaluation. In addition, each action class was evaluated with tight Average-mAP and loose Average-mAP. All reported metrics represent the average values obtained by training each model five times with different random seeds.

4.4 Evaluation of the Proposed Models

Since the audio spectrograms, commentary text, and text embedding were created in this study, we trained and evaluated seven cases: (i) visual modality only, (ii) visual and audio modalities, (iii)video and commentary text modalities, (iv) ASPERA (visual, audio, and commentary text modalities), (v)ASPERA$_{srnd}$, (vi)ASPERA$_{cln}$, and (vii)ASPERA$_{MC}$. The results are shown in Table 3, 4, and 5. In the tables, tight Average-mAP is abbreviated as "tight" and loose Average-mAP as "loose".

Table 3: Average-mAP for all actions, visible actions, and invisible actions.

Model	All		visible		invisible	
	tight	loose	tight	loose	tight	loose
ASTRA(video)	66.35	77.96	71.83	82.25	36.43	52.40
ASTRA(video+audio)	66.19	77.98	71.61	82.03	37.28	52.59
video+commentary text	66.17	77.97	71.70	82.28	36.66	52.44
ASPERA	66.93	**78.39**	72.21	**82.47**	36.64	52.55
$\text{ASPERA}_{\text{srnd}}$	**67.13**	78.18	**72.50**	82.34	36.936	52.766
$\text{ASPERA}_{\text{cln}}$	67.05	78.24	72.39	82.35	37.02	52.958
$\text{ASPERA}_{\text{MC}}$	66.98	78.02	72.45	82.26	**37.42**	**53.30**

Table 4: The metric of tight Average-mAP for all actions, including both visible and invisible actions, in each football action class (*Penalty, Kick-off, Goal, Sub, Offside, SonT, SoffT, Clearance*, and *BOOP*).

	Penalty	**Kick-off**	**Goal**	**Sub**	**Offside**	**SonT**	**SoffT**	**Clearance**	**BOOP**
ASTRA(video)	**87.29**	67.28	84.01	53.67	61.09	61.11	65.63	65.45	80.47
ASTRA(video+audio)	86.93	67.28	82.26	**55.84**	62.28	60.77	66.18	66.04	80.85
ASTRA(video+text)	85.87	67.86	83.66	55.07	60.32	61.05	65.30	65.24	80.41
ASPERA	86.44	67.49	83.75	55.24	**64.25**	62.16	**66.64**	66.17	80.48
$\text{ASPERA}_{\text{srnd}}$	86.52	68.30	83.94	55.57	62.19	**62.18**	66.35	**66.58**	81.37
$\text{ASPERA}_{\text{cln}}$	87.23	68.38	84.04	55.24	62.14	61.85	66.20	66.44	81.30
$\text{ASPERA}_{\text{MC}}$	86.21	**68.63**	**84.17**	55.74	62.06	62.05	66.20	66.20	**81.44**

Table 5: The metric of tight Average-mAP for all actions, including both visible and invisible actions, in each football action class (*Throw-in, Foul, Indirect FK, Direct FK, Corner, YC, RC*, and *YC→RC*).

	Throw-in	**Foul**	**Indirect FK**	**Direct FK**	**Corner**	**YC**	**RC**	**YC → RC**
ASTRA(video)	78.26	77.02	55.53	73.78	83.58	64.59	40.41	28.75
ASTRA(video+audio)	78.77	77.34	55.45	73.43	83.96	64.27	38.77	24.75
ASTRA(video+text)	78.85	77.31	55.69	73.69	83.25	64.90	38.48	27.97
ASPERA	78.89	77.68	56.00	73.75	84.01	65.24	40.96	26.53
$\text{ASPERA}_{\text{srnd}}$	79.21	**78.23**	56.45	**74.01**	84.38	65.40	**41.17**	29.36
$\text{ASPERA}_{\text{cln}}$	79.17	78.21	**56.49**	74.00	83.99	65.11	40.57	**29.81**
$\text{ASPERA}_{\text{MC}}$	**79.49**	**78.23**	56.20	73.67	**84.56**	**65.49**	40.07	28.27

4.4.1 Evaluation on All Actions, Visible Actions, and Invisible Actions Using Tight and Loose Average-mAP

Table 3 showed that ASPERA achieved the highest accuracy in both tight Average-mAP and loose Average-mAP for all actions and visible actions compared to the original ASTRA, which use only video and video+audio modalities. The result confirms the effectiveness of multimodal learning that utilizes three modalities—video, audio, and commentary text—for action spotting.

Comparing the model utilizing only video with the model utilizing both video and commentary text, some metrics showed a decrease in accuracy. This suggests that ASPERA improved accuracy by effectively capturing cross-modal relationships. $\text{ASPERA}_{\text{srnd}}$ demonstrated high accuracy primarily in tight Average-mAP, while $\text{ASPERA}_{\text{cln}}$ showed stable and high accuracy across both tight and loose Average-mAP. This suggests that considering the *surrounding text* improves tight Average-mAP, while excluding background information enhances overall detection accuracy. $\text{ASPERA}_{\text{MC}}$ showed the highest accuracy for invisible actions, achieving improvements of 0.78 in tight Average-mAP and 0.75 in loose Average-mAP compared to ASPERA. This suggests that introducing a Markov chain as prior knowledge of football action sequences enhances the accuracy of recognizing invisible actions. This effectiveness is attributed to modeling temporal dependencies that allow for predicting the next likely action based on previous actions and addressing the lack of sufficient visual data in recognizing invisible actions.

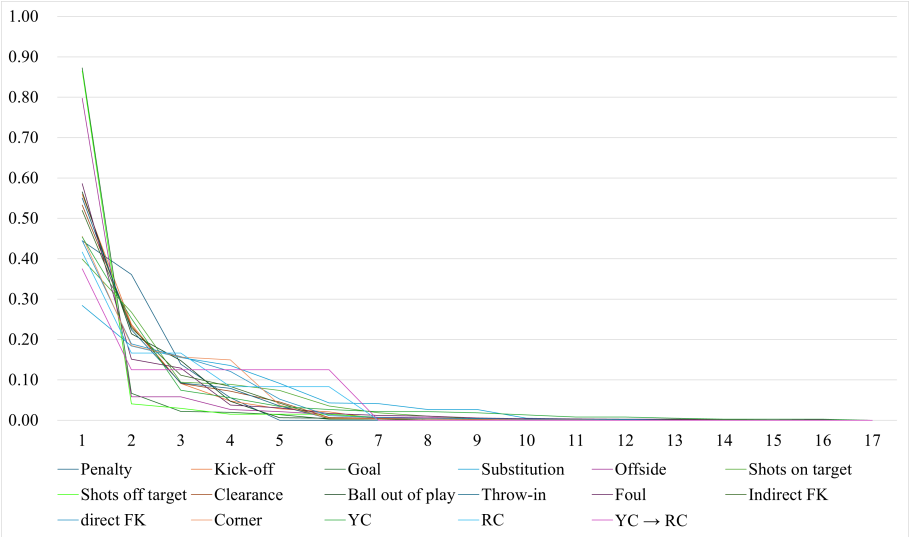

Figure 7: The transition probability matrix between each football class. The football classes are ordered along the horizontal axis based on the magnitude of their transition probabilities. The vertical axis shows the values of the transition probabilities.

4.4.2 Evaluation Using Tight Average-Map for Each Football Action Class

Tables 4 and 5 show the results of tight Average-mAP for all actions, including both visible and invisible actions, across all football action classes. *Sub, SonT, SoffT, BOOP, FK, YC*, and *RC* stand for *Substitution, Shots on target, Shots off target, Ball out of play, Free-kick, Yellow Card*, and *Red Card*, respectively. The Average-mAP for each football action in visible and invisible actions is published at http://www.ohsuga.lab.uec.ac.jp/information/average-mAP_ICAART.pdf.

ASTRA(video+text) and ASPERA. From Table 4, adding commentary text to ASTRA(video) or ASTRA(video and audio) improves accuracy for *Kick-off*, *Offside*, *SonT*, *SoffT*, *Clearance*, *Throw-in, Foul, Indirect FK, Corner, YC, RC*, and *YC→RC*. On the other hand, adding commentary text to ASTRA(video+text) or ASPERA decreases accuracy for *Penalty, Goal, Sub, BOOP, Direct FK*, and *YC→RC*. Based on the following two observations, this is believed to be due to the occurrence of the same words near different football action classes.

The first observation is the similarity of text embeddings spoken within five seconds of each football action class. The two graphs at the top of Figure 8 show examples where accuracy improved by considering commentary, for *Foul* and *Offside*. The two graphs at the bottom show examples where accuracy decreased, for *Penalty* and *YC→RC*. From these graphs, it can be inferred that the football actions where accuracy decreased by considering commentary had low similarity between text embeddings. This indicates that, for football actions with low accuracy, the commentary contained information unrelated to the game flow or insufficiently conveyed the game flow by itself.

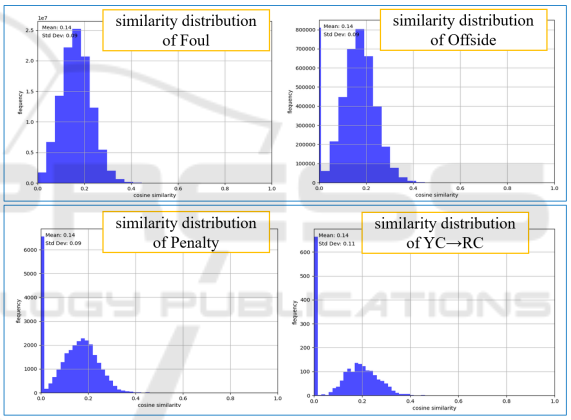

Figure 8: Similarity of text embeddings spoken within five seconds of each football action. The vertical axis represents frequency, and the horizontal axis represents similarity.

The second observation is the frequency of tokens spoken within five seconds of each football action. In Figure 10, the upper graph shows the frequency when Goal was recognized based on tight Average-mAP, the lower left graph shows the frequency when *BOOP* was recognized, and the lower right graph shows the frequency when *BOOP* was predicted but did not occur within the five-second window. Here, when *Goal* was recognized, the token "Goal" ranked first, and when *BOOP* was recognized, the token "Goal" ranked second. And when *BOOP* did not occur, the token "Goal" ranked second. This observation suggests

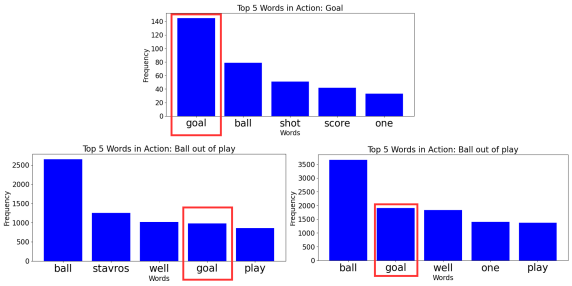

Figure 9: The frequency of each token spoken within five seconds of each football action. The vertical axis shows frequency, and the horizontal axis lists the top five tokens with the highest frequency.

the accuracy decreased because the token "Goal" frequently appeared in other classes.

ASPERA$_{srnd}$. The metric of tight Average-mAP for each action class improved in ASPERA$_{srnd}$ compared to ASPERA, except for *Offside* and *SoffT*. This is likely due to *sec-text* not fully capturing the game flow. Considering *surrounding text* provides additional game flow information, which reduces the impact of *sec-text*. Here, we examined the frequency of tokens spoken within five seconds of *Offside* and *SoffT*.

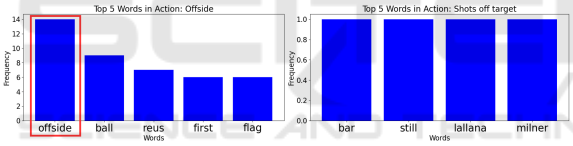

Figure 10: The frequency of each token spoken within five seconds before and after *Offside* and *Shots off target*. The vertical axis shows frequency, and the horizontal axis lists the top five tokens with the highest frequency.

It was found that, when *Offside* was correctly detected based on tight Average-mAP, the token "Offside" appeared most frequently and was not among the top five for any other football action classes. When *Offside* was not detected based on tight Average-mAP metric, the token "Offside" did not appear in the top five. Additionally, when considering commentary, the accuracy for *Offside* significantly improved. This suggests that the token "Offside" had a major impact on accuracy improvement. It is believed that considering *surrounding text* reduced the impact of this token, leading to a decrease in the accuracy of ASPERA$_{srnd}$ for *Offside*.

For *SoffT*, no token appeared repeatedly in the five-second window around its correct detection based on tight Average-mAP. Additionally, *surrounding text* for *SoffT* often described strategies and player situations. Below are examples of *surrounding text* for successful and failed detection of *SoffT* based on tight Average-mAP. The underlined part corresponds to *sec-text*. The decrease in accuracy is likely due to *surrounding text* for *SoffT* containing little relevant information.

- successful detection: That was the play of Chelsea's goal although it had a second part and let's say it is a prolongation with that recovery almost on the side of the area but the start of the Chelsea's play was like this let's see there's Casar Casar who has moved well towards Ivanovich. Ivanovic's center is not good now, it's way too far, he has to be very careful. Barley because what is that simply with that pass Gaby what has caused is that everything team would go up again this is what we are saying the Barley has its lines in a point of the field in which Chelsea with relative ease because they are very.
- failed detection: But look where Barley is installed. That is to say, obviously there will be bad stretches. He is going to suffer a lot and he is going to suffer the onslaughts of Chelsea. He will have to lower his center of gravity, right? Its center of gravity, the team. But while he can, he keeps the lines at a good height for the team. Boyd arrives. He didn't think twice. Filipe was being closed down. That is, the idea is not to lock themselves in. Obviously, if a team locks itself in, it is almost impossible to achieve something positive. You can lock it up but you have to unfold it at some point. Barley is trying, look, here it is again, don't huddle too much.

ASPERA$_{cln}$. The metric of tight Average-mAP for each football action class in ASPERA$_{cln}$ improved for *Penalty, Kick-off, Goal, Indirect FK*, and *YC→RC* compared to ASPERA$_{srnd}$. On the other hand, the accuracy decreased for *Sub, Offside, SonT, Shots off target, Clearance, BOOP, Throw-in, Foul, Direct FK, Corner, YC*, and *RC*. This is likely due to removing irrelevant information, which preserved the overall game flow while excluding specific football action details. Below are examples of *surrounding text* and *action-focused text* when *Sub* occurred. The underlined part corresponds to *sec-text*. *Sub* is mentioned in *surrounding text*, while in *action-focused text*, only the *YC* that happened just before is mentioned, and *Sub* is not referenced. Although the accuracy for many football actions decreased under tight Average-mAP, the accuracy increased under loose Average-mAP as shown in Table 3. This indicates that ASPERA$_{cln}$ is effective under loose Average-mAP.

- *surrounding text*: And a yellow card for Keitli for that tackle. About Cés. The previous action that was a clear yellow. And the referee does it very well here. He takes down his license plate number and then shows him a yellow card. Very ag-

gressive in some phases the Barley, leaving these entries a little rough. Good, William for Cuadrado. Cuadrado hasn't shined, to be honest. He has not been at a great level. It will be getting into the team's dynamics. It has been rumored this week. Who had an English teacher. Sculpture. There have been many jokes on Twitter. And he had to put a photo with his real English teacher.

- *action-focused text*: A yellow card was issued to Keitli for a tackle on Cés, while Cuadrado has not been performing at a high level.

ASPERA$_{MC}$. The metric of tight Average-mAP for each football action class in ASPERA$_{MC}$ improved for *Kick-off, Goal, Sub, SonT, BOOP, Throw-in, Foul, Corner, YC*, and *YC→RC* compared to ASPERA$_{cln}$. On the other hand, the accuracy decreased for *Penalty, Offside, Clearance, Indirect FK, Direct FK*, and *RC*. This is due to the design of the Markov loss, which emphasizes the occurrence frequency and transition probabilities of each football action, leading to improved accuracy for actions involving transitions with high occurrence frequency or extreme transition probabilities. Here, "extreme" refers to transition probabilities farther away from 0.5. Figure 7 shows the transition probability matrix between each football class. The accuracy of *Goal*, *Sub*, and *YC→RC* improved, as they have extreme transition probabilities compared to other football actions. On the other hand, *Penalty* and *Direct FK*, where accuracy decreased, had transition probabilities closer to 0.5 compared to other football actions. Additionally, *Penalty*, a less frequent action that also saw a decrease in accuracy, had many frequent transitions involving actions with fewer occurrences, which likely led to the decrease in tight Average-mAP.

5 CONCLUSION

In this study, we propose ASPERA, a multimodal football action recognition method by applying the Transformer-based architecture ASTRA to three modalities: video, audio, and commentary text. ASPERA, which was trained using these three modalities—video, audio, and commentary text—achieved improvements of 0.26 in tight Average-mAP and 0.80 in loose Average-mAP over models with video and audio modalities.

In addition to ASPERA, we developed three advanced models with enhanced text handling. ASPERA$_{srnd}$ incorporated *surrounding text*, including transcription within a ±20-second range around each *text clock* to reduce the effects of each *sec-text*. ASPERA$_{cln}$ further refined this by removing non-action-related background information, such as coaches' comments, which allowed a more targeted focus on action-relevant data. ASPERA$_{MC}$ introduced Markov head to ASPERA$_{cln}$, adding prior knowledge of football action flow via a Markov chain.

As a result, ASPERA$_{srnd}$ improved tight Average-mAP by leveraging background context around actions, and ASPERA$_{cln}$ achieved stable high accuracy across all metrics by focusing on action-related information. ASPERA$_{MC}$ showed the highest accuracy in detecting invisible actions, with increases of 0.78 in tight Average-mAP and 0.75 in loose Average-mAP, due to the predictive benefit of prior action flow patterns.

These results demonstrate how different combinations of modalities and additional information in each model affect the accuracy of action spotting. ASPERA$_{srnd}$ achieved high accuracy in tight Average-mAP, ASPERA$_{cln}$ showed stable high accuracy across metrics, and ASPERA$_{MC}$ was advantageous for invisible actions, suggesting a tailored model choice for different applications.

Future research directions include three main areas. First, optimizing the *surrounding text* feature representation may further improve action spotting, especially by pre-processing it similarly to video features before Transformer encoding. Second, combining Transformers and Markov chains holds promise for action recognition in invisible scenes, potentially extending to other sports and general activity recognition. The third point is player evaluation. Currently, research in player evaluation mainly focuses on analyzing scoring contributions by predicting player trajectories from video, as in Teranishi et al.(Teranishi et al., 2022); however, commentary texts can provide information such as which players were noticed, who scored goals, who made good plays, and who excelled in passing. Therefore, utilizing live commentary texts enables more detailed player evaluations.

ACKNOWLEDGEMENTS

This research was supported by JSPS KAKENHI Grant Numbers JP22K12157, JP23K28377, and JP24H00714. For the English translation review in this paper, we used ChatGPT o1-preview and Claude 3.5 Sonnet.

REFERENCES

Cartas, A., Ballester, C., and Haro, G. (2022). A graph-based method for soccer action spotting using unsu-

pervised player classification. In *Proceedings of the 5th International ACM Workshop on Multimedia Content Analysis in Sports*, MM '22. ACM.

Deliège, A., Cioppa, A., Giancola, S., Seikavandi, M. J., Dueholm, J. V., Nasrollahi, K., Ghanem, B., Moeslund, T. B., and Van Droogenbroeck, M. (2021). SoccerNet-v2: A dataset and benchmarks for holistic understanding of broadcast soccer videos. In *2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, pages 4503–4514.

Feichtenhofer, C., Fan, H., Malik, J., and He, K. (2019). Slowfast networks for video recognition. In *2019 IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 6201–6210.

FIFA, I. (2024). FIFA publishes Professional Football Report 2023. https://inside.fifa.com/legal/news/fifa-publishes-professional-football-report-2023. Accessed: 04/06/2024.

Gan, Y., Togo, R., Ogawa, T., and Haseyama, M. (2022). Transformer based multimodal scene recognition in soccer videos. In *2022 IEEE International Conference on Multimedia and Expo Workshops (ICMEW)*, pages 1–6.

Gemmeke, J. F., Ellis, D. P. W., Freedman, D., Jansen, A., Lawrence, W., Moore, R. C., Plakal, M., and Ritter, M. (2017). Audio Set: An ontology and human-labeled dataset for audio events. In *2017 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 776–780.

Giancola, S. and Ghanem, B. (2021). Temporally-aware feature pooling for action spotting in soccer broadcasts. In *2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, pages 4485–4494.

He, B., Yang, X., Wu, Z., Chen, H., Lim, S.-N., and Shrivastava, A. (2020). GTA: Global temporal attention for video action understanding. *arXiv preprint arXiv:2012.08510*.

Hershey, S., Chaudhuri, S., Ellis, D. P. W., Gemmeke, J. F., Jansen, A., Moore, R. C., Plakal, M., Platt, D., Saurous, R. A., Seybold, B., Slaney, M., Weiss, R. J., and Wilson, K. (2017). CNN architectures for large-scale audio classification. In *2017 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 131–135.

Intelligence, M. (2024). FOOTBALL MARKET. https://www.mordorintelligence.com/industry-reports/football-market. Accessed: 04/06/2024.

Lei, J., Li, G., Zhang, J., Guo, Q., and Tu, D. (2016). Continuous action segmentation and recognition using hybrid convolutional neural network-hidden markov model model. *IET Computer vision*, 10(6):537–544.

Neimark, D., Bar, O., Zohar, M., and Asselmann, D. (2021). Video transformer network. In *2021 IEEE/CVF International Conference on Computer Vision Workshops (ICCVW)*, pages 3156–3165.

OpenAI (2024a). GPT-4o. https://platform.openai.com/docs/models/gpt-4o.

OpenAI (2024b). text embedding large 3. https://platform.openai.com/docs/models/embeddings.

Radford, A., Kim, J. W., Xu, T., Brockman, G., McLeavey, C., and Sutskever, I. (2023). Robust speech recognition via large-scale weak supervision. In *International conference on machine learning*, pages 28492–28518. PMLR.

Shaikh, M. B., Chai, D., Islam, S. M. S., and Akhtar, N. (2022). MAiVAR: Multimodal audio-image and video action recognizer. In *2022 IEEE International Conference on Visual Communications and Image Processing (VCIP)*, pages 1–5.

SYSTRAN (2024). Faster whisper. https://github.com/SYSTRAN/faster-whisper. Accessed: 2024-06-10.

Teranishi, M., Tsutsui, K., Takeda, K., and Fujii, K. (2022). Evaluation of creating scoring opportunities for teammates in soccer via trajectory prediction. In *International Workshop on Machine Learning and Data Mining for Sports Analytics*, pages 53–73. Springer.

Tran, D., Wang, H., Feiszli, M., and Torresani, L. (2019). Video classification with channel-separated convolutional networks. In *2019 IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 5551–5560.

Vanderplaetse, B. and Dupont, S. (2020). Improved soccer action spotting using both audio and video streams. In *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, pages 3921–3931.

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, L., and Polosukhin, I. (2017). Attention is all you need. In *Proceedings of the 31st International Conference on Neural Information Processing Systems*, NIPS'17, page 6000–6010, Red Hook, NY, USA. Curran Associates Inc.

Xarles, A., Escalera, S., Moeslund, T. B., and Clapés, A. (2023). ASTRA: An action spotting transformer for soccer videos. In *Proceedings of the 6th International Workshop on Multimedia Content Analysis in Sports*, pages 93–102.

Yang, C., Xu, Y., Shi, J., Dai, B., and Zhou, B. (2020). Temporal pyramid network for action recognition. In *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 588–597.

Zhang, S. and Feng, Y. (2023). Hidden markov transformer for simultaneous machine translation. *ArXiv*, abs/2303.00257. https://api.semanticscholar.org/CorpusID:257255341.

Zhou, X., Kang, L., Cheng, Z., He, B., and Xin, J. (2021). Feature combination meets attention: Baidu soccer embeddings and transformer based temporal detection. *CoRR*, abs/2106.14447.