

AI-based clipping of booking events in soccer

Robin Rognerud



Thesis submitted for the degree of
Master in Applied Computer and Information Technology (ACIT)
30 credits

Department of Computer Science
Faculty of Technology, Art and Design

OSLO METROPOLITAN UNIVERSITY

Spring 2023

AI-based clipping of booking events in soccer

Robin Rognerud

© 2023 Robin Rognerud

AI-based clipping of booking events in soccer

<http://www.oslomet.no/>

Printed: Oslo Metropolitan University

Abstract

Manual clipping is currently the gold standard for extracting highlight clips from soccer games. However, it is a costly, tedious, and time-consuming task that is impractical and unfeasible for, at least, lower-league games with limited resources. Today, manual clipping is either used to trim away undesired video frames in a custom manner per video (high cost), or by employing a preset time interval leading to non-custom static clips (low quality). To address this issue, this thesis aims to automate the generation of highlight clips for booking events, in a custom and dynamic manner. In our pipeline, we will implement logo detection, scene boundary detection, and multimedia processing. We will also do a statistical analysis of current highlight clips, and perform a subjective evaluation. Full games are used as input, where detection modules will locate possible timestamps to produce an intriguing highlight clip. Through experimentation and results from state-of-the-art research, we will use neural network architectures and different datasets to suggest two models that can automatically detect appropriate timestamps for extracting booking events. These models are evaluated both qualitatively and quantitatively, demonstrating high accuracy in detecting logo and scene transitions and generating viewer-friendly highlight clips. When looking at state-of-the-art research and the results in the thesis, the conclusion is that automating the soccer video clipping process has significant potential.

Acknowledgments

I would like to thank my supervisors Pål Halvorsen, Cise Midoglu, and Steven Hicks for all the help, patience, and the good ideas throughout the semester. A special thanks goes to my family who have been very supportive, and to my friends for spreading fun and laughter.

Contents

Abstract	i
Acknowledgments	iii
1 Introduction	1
1.1 Motivation	1
1.2 Problem Statement	3
1.3 Scope	3
1.4 Research Methods	4
1.5 Ethical Considerations	5
1.6 Main Contributions	6
1.7 Thesis Outline	6
2 Background and Related Work	9
2.1 Definitions	9
2.2 Machine Learning	10
2.2.1 Supervised, Unsupervised and Reinforcement Learning	10
2.2.2 Classification	11
2.2.3 Datasets	11
2.2.4 Overfitting	11
2.2.5 Convolution	11
2.2.6 Neural Network	12
2.2.7 Convolutional Neural Network	13
2.2.8 Weight Initialization	14
2.2.9 Binary cross-entropy	14
2.2.10 Transfer learning	14
2.2.11 Spatial and temporal features	15
2.3 Related Works	15
2.3.1 Event Detection and Annotation	15
2.3.2 Event Clipping and Highlight Generation	16
2.4 Chapter Summary	16
3 Methodology and Implementation	17
3.1 Overview	17
3.1.1 Booking Event Statistics	18
3.1.2 State-of-the-art Clipping Procedure	19
3.2 Conceptualization	20
3.3 Datasets	22
3.3.1 Overview	22

3.3.2	Logo detection dataset	22
3.3.3	Image properties	23
3.4	Implementation	23
3.5	Logo Detection	24
3.5.1	Results from Valand and Kadragic	24
3.5.2	CNN-Model	25
3.5.3	Training and parameters	26
3.5.4	Evaluation	27
3.6	Scene boundary detection	27
3.6.1	TransNetV2	27
3.6.2	Pre-trained TransNetV2	27
3.6.3	Architecture	27
3.6.4	Training and Evaluation (Technical details)	28
3.6.5	Shortcomings and misclassifications	29
3.6.6	Results from Valand and Kadragovic	29
3.7	Pipeline	30
3.7.1	Ruleset	30
3.8	Subjective Evaluation	31
3.9	Chapter Summary	32
4	Experiments and Results	35
4.1	Statistical analysis of booking events	36
4.1.1	Clipping analysis	36
4.1.2	Video-technical analysis	36
4.2	Logo Detection	37
4.2.1	Input values and logo detection	37
4.2.2	Logo transition detection	39
4.2.3	Limitations	39
4.2.4	Summary	40
4.3	Scene Boundary Detection	40
4.4	Subjective Evaluation	42
4.5	Chapter Summary	44
5	Discussion	47
5.1	Addressing the Research Questions	47
5.1.1	Target 1	47
5.1.2	Target 2	48
5.1.3	Target 3	49
5.2	Limitations	49
5.2.1	Booking Events	49
5.2.2	Dataset	50
5.2.3	Logo detection	50
5.2.4	SBD	51
5.2.5	Approach	51
5.3	Open Questions and Future Work	51
5.4	Contributions	52
5.5	Chapter Summary	53
6	Conclusions	55

List of Figures

2.1	Graphic of overfitting where the functions is the model fitting to the training set. Source: [34].	12
2.2	A 2D convolution operation. Source: [48].	12
2.3	Max pooling with 2X2 filter and a stride of 2. Source: [3].	14
3.1	The regular order of details in a booking event.	19
3.2	The conceptualization of the average [5, 11] highlight clip produced for booking events.	21
3.3	The animation appearing after bookings, which we aim to recognize when detecting booking events.	21
3.4	A shortened version of the conceptualization if a clip is shortened due to circumstances in the game.	22
3.5	A small collection of frames during the logo transition.	23
3.6	Simple convolutional neural network (CNN) model.	26
3.7	The architecture of the Stacked Dilated Deep CNN [37].	28
3.8	The evaluation on a SoccerNet validation dataset. Source: [46].	29
3.9	The final evaluation of pre-trained vs. SoccerNet-trained TransNetV2. Source: [46].	29
3.10	A flowchart of the pipeline.	30
3.11	The questions in the post-questionnaire.	33
3.12	An example of the question in the central part of the survey.	34
4.1	The start of the logo transitions, recognized by the cap in upper left corner.	38
4.2	A frame wrongly predicted as logo transition.	39
4.3	Two examples of how a scene change happens, but TransNetV2 does not detect it. The network detect scene change between a and b, and c and d.	41
4.4	4 frames showing a wrongly detected scene change. Scene change detected between subfigure b and c.	41
4.5	Answers from the pre-questionnaire in the subjective evaluation, the questions are in the caption for each subfigure.	42
4.6	Average general experience from each video.	43
4.7	Results from the subjective evaluation where the importance of each detail in a clip is rated 1 to 5, where 5 is the most important.	44
4.8	Results from subjective evaluation about the duration of clips.	44

List of Tables

3.1	The distribution of frames in each set.	23
3.2	Execution time measured on DGX2 server 3.3.1. From [46].	25
3.3	Descriptions of each video in the survey, also shows how they are randomized in the survey	32
4.1	Statistics of the details in the analyzed clips from the current clipping in Eliteserien and Allsvenskan.	36
4.2	The number of different shots (scene changes) in a highlight clip from the current clipped highlights.	37
4.3	The frequency of logo transitions in each clip.	37
4.4	A confusion matrix of the predicted frames in the test set from the CNN with $108 \times 192 \times 3$ as input.	37
4.5	A confusion matrix of the predicted frames in the test set from the CNN with $72 \times 72 \times 3$ as input.	38
4.6	Precision, recall, and F1 score for the different input types in the CNN for logo detection.	38
4.7	Precision, recall, and F1 score for the neural network, TransNetV2, utilized on Eliteserien dataset.	40
4.8	Which clip of static and AI-based were decided as the best by the majority in the comparison	43

Chapter 1

Introduction

1.1 Motivation

The development in technology has, in recent years, made information and data globally accessible. The accessibility makes nonlinear streaming services such as Netflix and HBO increasingly popular. In addition, YouTube, TikTok, and other social media/content-sharing platforms are popular, with Youtube reporting over 2 billion monthly users [55]. Smaller devices and cheaper cellular data allow viewers to watch on-demand content on the go.

Sports are significant in people's everyday life as recreational hobbies and entertainment. During 2017 the watch time of sports highlights grew by more than 80%, and searches for 'how to' sports doubled simultaneously on YouTube. Google also stated that 80% of sports events viewers used additional devices like smartphones or computers to acquire different stats, data, and live scores to further engage with the event [1, 19]. Discussing and watching content on social media like Twitter and Facebook are also popular when viewing sports events. Producing captivating near real-time clips and highlights is necessary to elicit users' interest across different platforms. These clips should be short and capture mainly the dramatics of a highlight but long enough to give a stand-alone understanding of it.

Soccer is considered the world's biggest sport, with many active players and viewers of soccer events. According to the German Bundesliga, the sport has approximately 250 million active players across 200 nations [4]. Furthermore, the estimation is that 3.5 billion fans enjoy the sport, later confirmed when 3.572 billion people tuned in to follow the 2018 FIFA World Cup Final. Moreover, over half of the global adult population viewed a single event, giving a substantial target group. Therefore, it is as crucial as ever to produce content for the population increasingly hungrier on information and data. The average person's attention span has sunk drastically due to modern technology, where a ton of information and entertainment is only a couple of touches away. The span has shrunk by 4 seconds from the years 2000 to 2013 [8]. Therefore, content-sharing applications like TikTok and Instagram have become some of the most preferred social media today, as they aim to continuously deliver short, attention-grabbing clips and content to keep users at their app for a prolonged time. As a result, the aim must be to make the clips as short as possible but contain attention-grabbing elements.

Today, manually annotating standardized clipping the different highlights in Eliteserien and Allsvenskan from a "studio" is the protocol. After an event occurs,

people note it as a highlight and a static time decides the start and end points of a clip. This manual approach requires human resources and could be more practical. Also, the pre-determined time interval does not consider the actions of an event. Therefore, automating the clipping of specific highlights is of high interest, as it could save workforce and the cost of producing highlight clips.

Automatic clipping does not only save cost, but it can also further engage users with new and fast information. For example, when a highlight is detected, the automated clipping model can make a new clip within seconds and post it on a platform. Fast and accurate representation of highlights has a wide range of target groups in soccer fans. People following a game from home, fans attending the game live, and people on the go all have different reasons to be interested in a highlight replay. The biggest challenge, however, is the diversity of soccer fans and understanding what they prefer in a highlight. In bookings, most people like seeing a few seconds before the event to view what led to it. Hence, utilizing modern technology to present near real-time clips is valuable for user engagement.

Swift production of highlights is also motivated by the fact that betting on bookings makes up a considerable portion of the betting in soccer. The European Gaming and Betting Association (EGBA) accounted for 33% of the revenue in online gambling across Europe and had a gross gaming revenue (GGR) of €11,6 billion in 2021. In the GGR, sports betting, mainly pre-match betting, represented 46% of the revenue [12]. With this much money at stake, people would want to see what caused the event and the event itself. Well-produced and quick release of the clips could further engage debate and user engagement, with many soccer fans emotionally and economically invested.

Following the work of Valand and Kadragic, we want to produce clipping of highlights containing all the essential details automatically [45–47]. There is already research on event detection itself, but many neglect producing clips, especially for bookings. There are also related research on clipping, but mostly it focuses on goal-scoring events. Thus, the novelty of booking events is to find prominent starting and ending points, as well as compressing clips containing non-essential details.

Where much of the related works focused on goal-scoring events, the novelty of this thesis is to employ the same idea on yellow and red cards. Even though the manual annotating is similar in goals and bookings, there are other actions and details in the event we can and need to exploit. For example, we can apply parts of their intelligent system like logo detection and scene boundary detection as it appears similarly in bookings and goals. However, new problems and opportunities arise as the booking events are diverse. Some problems are prolonged sequences prior to the event, lasting advantages given by a referee, different types and numbers of replays, and a quick restart of the game. While goal-scoring events have a more average way of transpiring, bookings are far more dissimilar as an event. It can prevent replays from being displayed; the camera may not be on the referee handing the card, or the event occurred too far ago to make it a short clip. Utilizing work from Valand and Kadragic on clipping goal events [45, 47], and implementing booking-specific statistics and details should make automatic clipping of highlights feasible.

1.2 Problem Statement

The procedure of manually or statically clipping highlights is the main problem we wish to resolve. Forzasys want to quickly deliver highlights of booking events to the audience after the event transpires. However, manually clipping is tedious and time-consuming, and the editors have other simultaneous tasks like annotating. Therefore, the clips are often determined and edited by a preset time interval.

Observing the game, manually noting, and creating the clip could be more effective and cheap. As mentioned earlier, the clips are often of poor quality as the pre-determined time intervals do not fit soccer's diversity as a game. This affects Forzasys' strength in delivering interesting highlight events and blocks resources useful in other areas.

The thesis aims to implement an intelligent system automating the production of a clip. We receive an annotated event where we can extract a specific period in video format. We seek to use the intelligent system to make engaging and intriguing clips for the target group. Later, we can compare the clips produced by the system against the manual ones.

To build the system, we utilize parts of the solution from Valand and Kadragic [46, 47], which again relies on state-of-the-art technology. Their attempt to automate highlight clipping dealt with goal-scoring events, including shot classifications, neural networks, and clipping as machine learning tools. However, as their solution only involved goal-scoring highlights, we must examine which tools are relevant to this thesis. To accommodate the novelty of bookings, we must implement new ideas as a whole new set of details happens in a booking event.

Therefore, the thesis's overarching purpose is to attempt to automate the production of highlight clips in booking events to aid today's manual operation. To reach this goal, we divided the purpose into three research targets for more straightforward structuring. The three research targets are defined as follows:

- **Target 1** Analyze current highlight clips and identify relevant data specific to bookings.
- **Target 2** Design and implement a pipeline detecting details of a booking event from the annotated game.
- **Target 3** Complete an evaluation of the new pipeline through automatically clipped highlights versus the current state-of-the-art in commercially deployed clipping protocols highlights.

1.3 Scope

The scope of this thesis is booking events in soccer, constituted by yellow and red cards. Creating a solution consisting of a machine learning system to make clips is the goal, and the system will harness some state-of-the-art tools already developed [37, 46, 47]. Our approach employs a convolutional neural network to recognize patterns like logos and scene boundaries, which are the central components of the solution. Thus, the main scope is collecting events, detecting details, and composing clips.

As most research in the field focuses on goal events, it is necessary to understand what clips of booking events consist of and how they are structured.

Ergo, we have to carry out an analysis of the current highlight clips. As we propose a pipeline that produces compelling clips, it is subjective to the audience if the clips are exciting. Therefore we must also perform a user survey to measure how well the audience like the clips. Thus, the main scope of this thesis relates to the research targets, i.e., collecting data and statistics on booking clips, detecting details in the clip, and developing a pipeline that composes clips.

Outside the scope is event detection of entire games, as existing research uses models to locate events. We assume this is done manually due to the 100% requirement of accuracy and that we will receive annotated events or videos. We will receive annotated events or videos and then use models to recognize the characteristics of booking events. Audio-related work, which could help detect details in events, is outside of the scope due to the length of the thesis.

We will use the and clips from Eliteserien 2022 [11], Allsvenskan 2022 [5], and complete games from Eliteserien 2020 handed by forzasys.com. As the input to the pipeline we will use in testing are the full games, we do not consider adding functionality for other datasets like SoccerNet. Due to the limited time in a short thesis, we must limit the datasets to streamline the development.

1.4 Research Methods

Computer science is a vast field that encompasses a wide range of research methods. According to the report approved and released by the ACM Education Board, we use three paradigms to provide context when approaching the discipline of computing [10]. The paradigms are:

Theory The first paradigm is theory and originates from engineering and mathematics. Mathematicians and computer scientists mainly practice this paradigm through four steps. (1) characterize the object of study, (2) hypothesize possible relationships among them, (3) determine whether the relationships are true, and (4) interpret the results. The steps are iterated if errors or inconsistencies are apparent.

Abstraction This is the paradigm where the scientific method of investigating a phenomenon occurs. This one also has four steps to iterate through if the experiment results deviate from the model's predictions. (1) form a hypothesis, (2) construct a model and make a prediction, (3) design an experiment and collect data, (4) analyze results.

Design The last paradigm is design, which is utilized in engineering and development. The steps engineers wish to iterate through here are (1) state requirements, (2) state specifications, (3) design and implement the system, and (4) test the system.

The research we will conduct fits under the paradigm of design. The motivation of the thesis is to find possible improvements and to plan and develop a pipeline to fulfill the improvement. To find the upgrades, we must state the requirements and specifications needed. And we must design, implement and test the system to show the experimental results. The requirement is to improve the standardized highlights.

To do that, we specify the use of logo detection and scene boundary detection as tools, and develop them with different evaluation metrics as testing.

We also nudge both abstraction and theory paradigms. As we use related works and other statistics and competence to consider which tools and techniques we wish to implement. This is related to the steps of finding relationships in characterized objects in the theory paradigm. The abstraction is also somewhat relevant as we use experimenting in the development of a system, and look at evaluations to produce better predictions.

Qualitative and quantitative research are two distinct methods of conducting research. Qualitative research aims to gather an in-depth understanding of the area in which the research is conducted. This is often done through interviews, focus groups, and observations, focusing on people and behavior [18]. The nature of this research is exploratory, and it wishes to learn about undocumented theories and untested hypotheses. On the other hand, quantitative is more numerical and statistical in its approach. It employs experiments, questionnaires, and statistical analysis to draw conclusions and learn relationships between data. Both methods have strengths and flaws, so it is elemental to know which method fits the research.

The research in this thesis is quantitative in nature, as it involves experimenting and a survey. The survey results in a statistical analysis which helps us gain insights into the pipeline we implement. The pipeline is done as experimental prototyping, which also hints toward quantitative research.

Thus, the research done in this thesis is quantitative and fits under the design theory. Experimental prototyping requires planning, experimenting, and testing, which are known as design theory characteristics.

1.5 Ethical Considerations

As artificial intelligence (AI) becomes more ubiquitous, it is urgent to consider the ethical implications of its development and deployment. Ethical considerations in AI include machine ethics, transparency, accountability, and privacy. In addition, it is important to ensure that AI is developed and deployed in ways consistent with our values and respect human dignity, autonomy, and well-being. As such, ethical considerations in AI have become a critical area of research and debate, with stakeholders of the subject works to ensure AI's responsible use.

Privacy is an essential ethical consideration in computer science, and the branch of AI is no different. Zhang et al. [58] state that privacy is one of the most important ethical principles, specifically the ability to control personal information. With the evolving nature of AI, this is a topic that should be scrutinized. As this thesis will utilize videos of soccer players, it is paramount to gather and maintain the consent of the involved individuals. Furthermore, it is necessary to ensure that users are aware of the potential implications of sharing their data and that they have the right to control how their data is used.

Machine learning also touches on other ethical considerations like machine ethics. These considerations are on the morality of intelligent systems and machines and how to keep human rights in robots, for example. With the growing consciousness of machines and the increasingly rapid development of AI, it is vital for ethical thinking to follow the speed of development and not be neglected [58].

Fairness is another example that touches on the public's trust in the fairness of decisions made by a system. Often, the general public is unknown of how these AI systems work and are therefore not comfortable with personal data being used. Transparency is, in this case, important [58].

Fairness and machine ethics are related to deep learning and the thesis theme. However, these last ethical considerations are not directly related to the concepts like privacy and should only be kept in mind when developing. Nevertheless, it is crucial to approach the development of AI in videos and pictures with a deep awareness of the ethical considerations involved and to work to mitigate any potential negative consequences.

1.6 Main Contributions

In this section, we will discuss the thesis's main contributions. Based on the research targets expressed in section 1.2, we can explore innovative aspects and original thoughts we can contribute. For example, we can contribute to analyzing booking events in soccer and automating highlight generation. Thus, with good results, this research can be implemented as a tool for TV productions.

As research on booking events is fresh in this field, we can contribute to understanding what defines a booking event. For studies in event detection, clipping, and summarization, goal-scoring events are dominant in the literature [23, 47]. There are studies also detecting cards, however they do not emphasize what defines a booking event. [29]. Therefore it is interesting to contribute by exploring how similar tools work in booking events.

A statistical analysis of the booking event is another main contribution of the thesis. It is established in the regular soccer audience that a booking is when a player receives a yellow or red card. However, most people do not pay close attention to what constitutes a whole booking event in the production of the videos. Therefore we must analyze which details appear when a foul occurs, and a card is handed out. The statistical analysis will aid the research field in understanding what defines the event.

Further, Gautam et al. attempted summarization of highlights for whole games [16]. In contrast to this thesis, it was based on game audio, annotations, and commentary. Meaning they did not define booking events. Also, they did not focus on how the clipping. This will be one of the main contributions in the thesis, to create compelling and intriguing clips.

To conclude, we will contribute to the research field by emphasizing how booking events are defined, its limitation and challenges, and by implementing a pipeline that clips intriguing clips. There are similar studies in this field. However, they deal with goal-scoring events or event detection for game summarization. The pipeline can be found in <https://github.com/simula/forzify>.

1.7 Thesis Outline

The remainder of this thesis is structured into 5 chapters:

- **Chapter 2 - Background and Related Work**

This chapter is an introduction and an overview of the key concepts and techniques we shall employ in the thesis. From the broad subject of machine learning to the narrow mathematics of deep learning models like CNN. We are also summarizing some related works and the history of how these concepts have developed. This chapter introduces the bedrock of the subject, which is necessary to understand before further studying the thesis.

- **Chapter 3 - Methodology and Implementation**

The methodology chapter will delve into how we conduct our research, initially describing the datasets, properties of the videos, and our approach to conceptualizing booking events. Further, we will discuss the proposed pipeline and how we will construct it succeeded by explaining how we will evaluate it. Finally, the evaluation will consider various parameters and variables presented in this chapter.

- **Chapter 4 - Experiments and Results**

This chapter presents our results gained from experimenting and survey. First, we obtain results from the statistical analysis of the current highlight clips. Then, we see results from experimenting with Convolutional Neural Network (CNN) in logo detection and TransNetV2 in scene boundary detection (SBD). Finally, we present the answers and insights gained from the survey.

- **Chapter 5 - Discussion**

This chapter presents a discussion of our findings and relates them to our research questions, as well as discuss some of the limitations of the work. The limitations and challenges will be ordered as in previous chapters, starting with current highlight clips, followed by logo detection and SBD, and ending with the survey. Furthermore, we will discuss possibilities for future work.

- **Chapter 6 - Conclusion**

This chapter summarizes and concludes the thesis. Where we start with the motivation for the research, followed by how we conducted the experiments and the results of them. Finally we conclude discussions.

Chapter 2

Background and Related Work

In the previous chapter, we identified real problems in extracting soccer highlights, and referred to solutions utilizing machine learning and modern technology. We must delve into the technicalities to comprehend our approach to the issues. Firstly, we need to understand a few basic ideas machine learning systems intend to resolve. Then, we go technically deeper to see how machine learning systems operate to figure out these ideas. This funnel-like approach provides the necessary overview for the remaining chapters.

Reviewing state-of-the-art research in the field is also necessary to acquire relevant information. We must understand outcomes in event detection, video and image processing, and other relevant ideas to utilize the technology optimally.

2.1 Definitions

In this thesis, the goal is to detect the booking events in a soccer match. However, there are similar challenges tied to booking events as goal events. We also have to account for additional challenges. Most soccer events generally have various durations related to the game's context, making it hard to define an event. Also, when comparing booking and goal highlights in Forzasys' gallery, the booking events were far more diverse. Not only in duration but also in details of the event.

We define the booking event as showing the yellow or red card to the player. Given a specified core of the event, it corresponds with annotations from highlight clips, and it is easier for the system to recognize the pattern.

Terminology used throughout this thesis, with some being gathered from the thesis of Husa [20]:

- **Soccer:** Also called association football¹, played in accordance with a codified set of rules known as the Laws Of The Game (LOTG)² by the International Football Association Board.
- **Broadcast:** A transmit of television.
- **Streaming:** Different than broadcasting, Over-the-top (OTT) delivery of content.
- **Soccer game/match:** Soccer game is any game including unofficial.

¹https://en.wikipedia.org/wiki/Names_for_association_football

²<https://www.fifplay.com/downloads/documents/laws-of-the-game-2021-2022.pdf>

- **Football club:** 'An organization of players, managers, owners or members associated with a specific football team.' [15]
- **Event:** (in the context of multimedia content, not to be confused with the entire soccer game) Also called "highlight". According to the Cambridge Dictionary an event is defined as: 'anything that happens, especially something important or unusual' [13]. In this thesis it will be focused on something important or unusual in a soccer match. For this occasion an event is distinct and the list below includes types of events that are relevant in a soccer match.
 - Goal: When the ball passes the goal line.
 - Card: When the referee hands a yellow or red card to a player.
 - Substitution: When one player is substituted off and another player goes on the pitch.
- **Booking:** When referee hands yellow or red card to a player. See event, card
- **Highlight clip:** Video clip displaying a particular event from a soccer game.
- **Tagging/annotation:** Setting timestamps on events in soccer match, adding metadata, etc..
- **Detail:** A specific part of a clip like logo transition, foul or a scene of a close-up.
- **Shot:** A sequence of frames captured by a single camera.
- **Shot/scene boundary:** A transition between two successive shots.
- **Close-up:** An image or shot that shows subject up close and detailed. Typically making the subject cover most of the frame.

2.2 Machine Learning

2.2.1 Supervised, Unsupervised and Reinforcement Learning

The three branches of Machine Learning (ML) are supervised learning, unsupervised learning, and reinforcement learning. Supervised learning uses labeled datasets, which means the model knows the correct answer and can train to be fitted accordingly. For example, a labeled dataset with images of animals would tell which images were cats or dogs. When trained, the model can compare input images of animals with the labeled ones and try to predict correctly. Supervised learning is, therefore, suitable for classification problems with a known truth or answers, such as logo detection. Unsupervised learning involves unlabeled datasets with no specified truth to train the model. Thus, the model actively tries to find patterns and structures that could prove helpful. Reinforcement learning is employing agents following specified rules to maximize learning. Algorithms and programs are typical agents attempting to find the best path to accomplish goals [39].

2.2.2 Classification

As mentioned, supervised learning is eminent in solving classification problems. In our case, we can send images labeled either “game” or “logo” when training for logo detection. When classifying into two groups of truth, it is called binary classification, which provides four possible outcomes: true positive, false positive, true negative, and false negative.

2.2.3 Datasets

Datasets are crucial for algorithms and systems to function in AI. The algorithms build on the datasets as input and base the learning on it. The complexity of machine learning algorithms requires massive datasets to learn correctly. Hence, many pictures and videos are necessary for the systems to recognize patterns mathematics and humans cannot.

This supervised learning is made possible by labeling the images. Labeling adds context information so that machine-learning models can learn from it. Often, it is humans who label unlabeled data to give correct information. For example, labeling can range from defining if there is an animal in the photo, or defining the pixels containing the animal. Using the corresponding labels to our images, we often split the datasets into three subsets for the machine learning model to employ.

The three fractions are training, testing, and validation. The training set is the greatest as it is the set machine learning models use when identifying patterns and fitting weights. The validation set helps control the algorithm in addition to finding possible improvements. This set does not update the model or any weights; we can consider it a current measurement. However, we use the validation set to tune the model and understand how well it generalizes, which helps against overfitting. This is why we also need a standalone test set to measure how well the model performs when an unknown set of images goes through the algorithm. It is an evaluation to check the model’s generalization and should be implemented in the end phase [51].

2.2.4 Overfitting

The mentioned overfitting is a notable problem occurring in neural networks. It happens when a model performs well on the training set but poorly on testing sets or unknown data. Hence, the model generalizes badly and overfits the training set’s specifics, as seen in Figure 2.1. Obviously, the goal of the model is to perform well on the unlabeled data rather than the known data it trains on.

2.2.5 Convolution

Convolution is a mathematical operation on which the Convolutional Neural Network (CNN) builds. It expresses how much overlap one function, e.g., f , has when it shifts over another, e.g., g . The definition of it is:

$$h(t) = \int_{-\infty}^{\infty} f(\tau)g(t - \tau)d\tau \quad (2.1)$$

Wang and Raj state that convolution is an element-wise product followed by a sum. Figure 2.2 shows when a 3×3 matrix (on the left) convolutes with a kernel (the

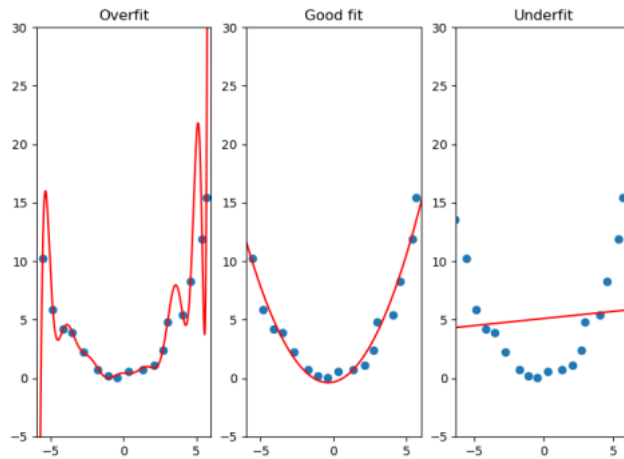


Figure 2.1: Graphic of overfitting where the functions is the model fitting to the training set. Source: [34].

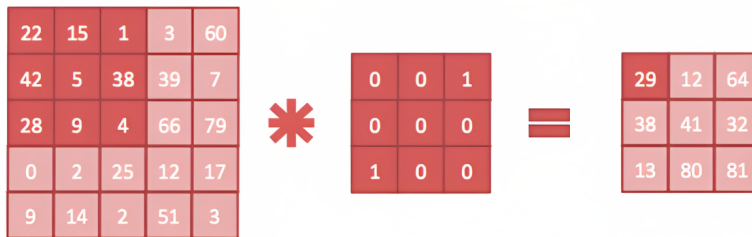


Figure 2.2: A 2D convolution operation. Source: [48].

middle matrix), it results in the top left rubric on the matrix to the right. Next, slide the 3×3 matrix one column to the right and repeat the convolutional process. The process is done when all the 3×3 matrices are convoluted [48].

Filter size and stride are necessary values we need to understand in this convolution operation. Here, the filter size corresponds to the 3×3 matrix of information from the leftmost matrix, which we can adjust to different sizes. Stride is how many columns we slide over before applying the operation. Suppose the stride is 1; then the filter size slides just one column. However, if it is two, it skips one column and slides two times before the operation [46].

The reason for applying convolution is to detect different aspects of an image. For example, utilizing different kernels obtains edge detection, blur, and sharpening, among other elements. Random kernels can also be applied to achieve intriguing transformations [48].

2.2.6 Neural Network

Neural Networks are a branch of machine learning and are vital for deep learning algorithms. As the name suggests, the brain is the inspiration, with all its neurons signaling each other. A neural network incorporates several layers of nodes, consisting of an input layer, one or several hidden layers, and an output layer. The layers can have a different amount of nodes and are, therefore, independent of each other [2, 50].

The neural networks utilize training data to improve accuracy and learn accordingly. As a result, the trained algorithms become robust classification and image recognition tools, which is helpful for the thesis.

The nodes are interconnected and can send signals based on activation functions assigned to the network. With allocated weights and variables, the activation function fire a signal to the next layer of the network if the variable exceeds a set threshold. Hence, it resembles the neurons in the brain firing signals toward output and is depicted as a feedforward network. Sigmoid and ReLU are two prevalent activation functions [50].

Also, having a loss function is necessary for measuring a prediction model. It can show how close the trained model is to being optimal. A function approximation represents how accurately it maps the input to the correct output. Given the function, we can adjust weights to fine-tune the network and eliminate loss [2, 46].

2.2.7 Convolutional Neural Network

Convolutional Neural Networks (CNN) is a type of deep learning algorithm combining the convolutional operation with the structure of a neural network. It has become increasingly popular in recent years and performs specifically well on classification and images. CNN is constructed through the layers of a neural network, with a convolutional layer, pooling layer, and fully connected layer as the main building blocks.

The convolutional layer works as expressed in the previous subsection and helps the neural network extract features. It is possible to use several convolutional layers paired with pooling layers in a neural network, as it can find progressively more detailed patterns in an image. For example, the first convolutional layer can detect edges and lines, while the next can detect figures such as squares. And lastly, it can locate patterns such as animals or buildings [3, 9].

Next, the pooling layer helps with downsampling by the feature map. This layer does the same sliding filter operation as the convolutional layer, only without weights. In other words, the kernel works as a reduction operator, as seen in Figure 2.3. The goal of pooling is to reduce complexity for future layers, which resembles lowering the resolution of an image. 2×2 is the most frequent size for pooling in combination with a stride of 2. This procedure does not keep track of spatial information but works well to control if the information is apparent [3, 9]. The two typical pooling techniques are max pooling and average pooling, where the pooling procedure chooses the max and the average of the region, respectively.

It is possible to repeat several steps of convolution and pooling, but we need a fully connected layer before the end output. The fully connected layer must have each node connected with all nodes in both the previous and the next layer. The obvious drawback of this layer is the computational cost of intertwining all these nodes. Training can therefore be time- and resource-consuming. As a result, researchers introduced dropout to eliminate the number of connections, easing the computation and learning [3].

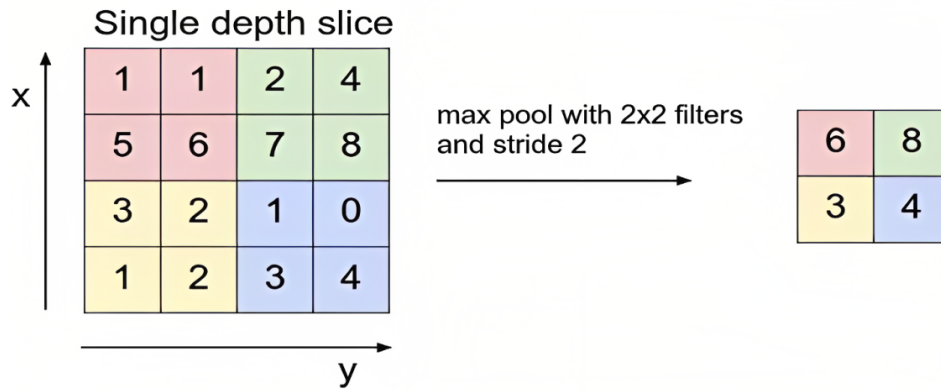


Figure 2.3: Max pooling with 2X2 filter and a stride of 2. Source: [3].

2.2.8 Weight Initialization

Weight initialization in neural networks is setting the values of the weights in a neural network before training. The weights are typically initialized to small random values, and the initialization process is an essential factor in the success of a neural network. Therefore, proper initialization techniques enable the neural network to learn effectively.

Relevant to the thesis is techniques where weights are initially selected from a random distribution but later can be modified to minimize loss and improve the model. Glorot and Bengio presented a popular technique that bases initialization on the number of input nodes and layers in a neural network. This technique grew popular and inspired several researchers to experiment with parameter norms and rectifier non-linearities building on Glorot and Bengio's work [27].

2.2.9 Binary cross-entropy

As the model in the artifact will contain binary classification, we will utilize binary cross-entropy. This loss function is advantageous for the binary classification problem as it represents the loss exponentially. Which therefore makes the function sensitive to outliers. This is binary cross-entropy:

$$H_p(q) = -\frac{1}{N} \sum_{i=1}^N y_i * \log(p(y_i)) + (1 - y_i) * \log(1 - p(y_i)) \quad (2.2)$$

Quickly explained, it computes the probability of predicting the correct label ($p(y)$) and adds the logarithm of the probability to the loss. In addition, it adds the other predicted label output to $\log(1 - p(y))$. Since the result of log between 0 and 1 is negative, we use a negative \log to receive positive results. As a consequence of using log, we now receive positive values and exponentially rising losses for wrong predictions [46].

2.2.10 Transfer learning

Unfortunately, collecting satisfactory training data is time-consuming, tedious, and often unrealistic. (source). As a result, initializing weights and training models could

result in incomplete or faulty networks. A possible solution is transfer learning, where a network can utilize already recognized patterns and initialize a model with known weights. When humans learn something new, we use earlier experiences to assist in the process. Therefore, machine learning should also aim to realize the transfer of knowledge from model to model [59].

2.2.11 Spatial and temporal features

When using machine learning to classify images, detecting temporal features is more complicated than spatial features. Different models are incredibly accurate in classifying images. However, detecting events in a video over several frames is challenging and demands more complex models. Among other use cases is event detection in soccer, which is highly relevant to the thesis. Employing models to find bookings in a game automatically is a version of temporal feature extraction, which must use state-of-the-art technology to succeed.

2.3 Related Works

2.3.1 Event Detection and Annotation

Event detection is where intelligent systems can figure out specific occurrences in a game, like goals, shots, and fouls. Advanced image and video processing are required for the system to find patterns corresponding to event heuristics. Although finding these heuristics is the first issue of event detection, we need to define what constitutes an event. For bookings, some recognizable patterns are the cards when shown to a player, animation in the video, and close-ups of players.

Some other well-known patterns used in event detection are logo detection and scene-boundary detection. As replays and logos often occur when interesting events happen, systems detecting these events are reaching good results.

ML research has recently succeeded in solving and automating numerous video-related problems. As a result, it can show viewers desired events from video segments in an efficient manner. In this section, we present chosen work and papers in the fields of event detection and event clipping [45, 47].

Event detection, also called *action detection* or *action spotting*, has lately been popular and received much attention. Several versions of two-stream CNNs have been variously employed in the problem [22, 36], and further developed to incorporate 3D convolution and pooling [6, 14]. Wang et al. [35, 49] proposed temporal segment networks (TSN), and C3D [41] explored 3D convolution learning spatio-temporal features. Tran et al. [43] later utilized (2+1)D convolutions to enable the network to learn spatial and temporal characteristics independently. Additionally, optional methods focus on identifying specific temporal instances within the timeline [17, 21, 25, 26, 28, 29, 33, 53]. Although many of these works propose interesting approaches and promising results, these technologies are not currently ready for implementation in practical, real-life scenarios. The explanation is that most of the presented models are significantly expensive and need to be more accurate. The results are used officially (e.g., live sports broadcasts) in live-action and event annotation, such as sports and betting. Hence, the results must reach 100% accuracy as it will affect persons and critical systems. Therefore, no false alarms or overlooked

events can happen, and resources must still be placed in manual annotating. In this thesis, the aim is to automate the clipping after the annotation is done, such that we can limit the resources to a minimum and still reach great accuracy through manual annotation.

2.3.2 Event Clipping and Highlight Generation

In *event clipping*, the existing work is limited. A shot detection algorithm was shown by Koumaras et al. [24], further tailored by Zawbaa et al. [57], implementing the possibility to detect transitions over several frames. Zawbaa et al. [56] further made classifications for soccer video scenes: long, medium, close-up, and audience/out of field. Other papers have performed good results in scene classification [30, 54, 56]. Replay detection is also performed by Ren et al. [32], amongst other things labeled as play, focus, and breaks. support vector machine (SVM) were later presented as effective in detecting replays, but not with artificial neural network (ANN) [56, 57]. Valand et al. [45, 47] also researched replay detection, while also looking at scene boundary detection. Audio is also proven to perform in clipping. Raventos et al. [31] used audio to find highlights, and Tjondronegoro et al. [40] did summarization by detecting when a whistle was blown. Lastly, we have seen research on spatio-temporal features [7, 36, 42].

The potential of AI-based clipping and video manipulation is certainly apparent. Especially, temporal information is interesting in producing highlight clips. However, there are still limited results in event clipping, and specially for booking events. The aim should be to deliver clips rapidly, given the need for real-time highlights in several productions.

2.4 Chapter Summary

Most importantly, we have looked at the background of the machine learning tools and techniques we utilize in this thesis. By understanding the basics of convolution and neural networks, we can understand why they are effective and popular in this research field. We have also explained some related works, such that we can place our research in a greater context. Especially the rapid growth and results of predictions and machine learning in event detection are exciting for soccer. However, as most research specialize in goal-scoring events or full games, it is interesting to delve into booking events. Now, we will look at how the building blocks of the thesis.

Chapter 3

Methodology and Implementation

Aiding the attempts to automate the standardized clipping done by Forzasys today, we wish to construct a pipeline consisting of various algorithms and models. Furthermore, we can supplement and build the pipeline by automatically detect shot types of soccer bookings in a video by researching state-of-the-art technology regarding detection techniques. In addition, we need to examine the factors supporting the development and rate of success, like the datasets, how they film and produce the clips, and the order of constructing the pipeline.

When we look at the current clips of bookings in the Eliteserien and Allsvenskan, we notice details and animations the highlighted clips wish to employ. Assuming they follow a protocol when annotating and editing the clips, we can conceptualize this protocol to extract meaningful information to find relevant technology. Moreover, we can also use the conceptualization and recognition of patterns to produce compelling clips. Exploring different clips of the same event and performing subjective surveys can propel interest in the highlight clips.

This chapter consists of the proposed pipeline to solve the detection and clipping of bookings. On that basis, we will first look at statistics and datasets. Then, partition the pipeline to comprehend the building blocks. For example, we use logo detection to recognize replays and scene boundary detection to divide the clips and identify zoomed frames. The priority is to extract highlight clips of the events with as high accuracy as possible and create compelling clips for the subjective supporter, stakeholders, and other interested bodies. Finally, we should compare the solution to the current procedure in Eliteserien and Allsvenskan through a subjective evaluation and acknowledge feedback.

3.1 Overview

When building a pipeline, it is vital to envision what constructs a highlight clip and see which details, Forzasys, for example, operates with in a game. Football is an incredibly dynamic game, making it hard to follow a convention at all times. Bookings in a game reflect this, as various aspects of a booking event influence the cameras, scenes, and animations.

Thus, we will use the current highlight pages from Allsvenskan and Eliteserien [5, 11] to analyze clips. We will review these clips in a few details, which seems to be the standard protocol of clipping on a regular booking event. This statistical approach

helps answer research target 1, enabling us to make meaningful comparisons and identify patterns in the current protocol.

3.1.1 Booking Event Statistics

Several statistics are interesting for this thesis, guiding us in understanding how the highlight clips are technically produced. We list the repeating data we explore and find when observing the highlight clips:

1. Number of shots: the total number of shots in a clip can give us a good indication of the complexity of the average shot, as well as the similarity between different clips. If the number of shots is similar across several clips, we realize they follow a strict protocol of clipping and production.
2. Shot length: We can identify different patterns in the clip if we measure the duration of the diverse shots. This statistic is also interesting when determining the protocol. Logo transition frequency: Analyzing how often a logo transition occurs is suitable for understanding when a clip is started and ended. It also emphasizes how many replays they usually operate with.
3. Logo transition frequency: Analyzing how often a logo transition occurs is suitable for understanding when a clip is started and ended. It also emphasizes how many replays they usually operate with.
4. Limited clips: One of the most important statistics and data we can gather from the current clips is how many limited clips there are. E.g., knowing how many clips are incomplete due to circumstances in the diversity of the game. It is also worth noticing the reasons for the limitation.

We have done a deeper statistical analysis on some of these points, to learn deeper how the highlight is usually produced. The results are shown in section 4.1, but main points for understanding the clips is presented below.

Elitserien and Allsvenskan are alike in the clipping and how they produce the highlight clips of the booking events. However, they differ significantly in the average duration of the clips. Where Elitserien follows a strict time of around 25-26 seconds and has an average of 25.8 seconds, Allsvenskan is more dynamic. Most clips range between 20 to 35 seconds, and some clips are as long as 80 seconds. The average is 36.2 seconds, more than 10 seconds longer than Elitserien. It shows how the different challenges of clipping a booking event also require a dynamic approach.

Table 4.2 shows most clips range from 4 to 6 shots, with the average being 4.96 for the Elitserien clips and 5.29 for the Allsvenskan clips. Usually, the clip starts from open play, then changes scene to a couple of close-ups when the foul has happened. Further, a logo transition goes into a replay which counts as one or two extra shots given the clip, before it finally ends with a logo transition out of the replay. This order of actions can be seen in figure 3.1 gathered from an Elitserien game between Brann and Tromsø. The logo transition backs this statistic happening close to always between shots 4 and 6. Also worth noticing is that all clips and events with none of the mentioned limitations have either one or two logo transitions. Therefore, they are consistently implemented in production if there is time.

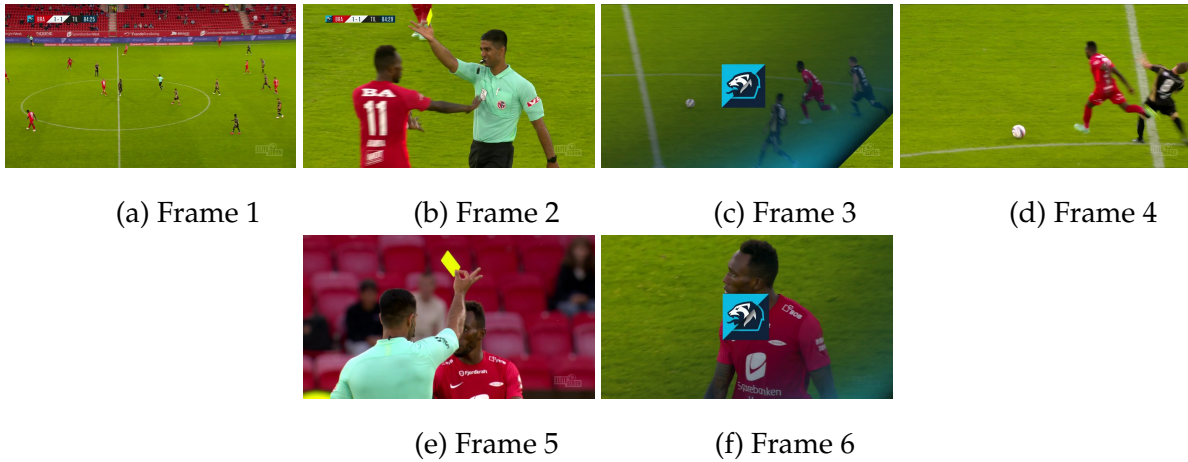


Figure 3.1: The regular order of details in a booking event.

The average frequency of logo transitions are 1.29, meaning they are very often apparent in the highlight clips. We can see how each of 0, 1, and 2 logo transitions are distributed in table 4.3. There are a few interesting takes from this table. Firstly, the majority is 2 logo transitions, followed by 1. Which means it is what the producers target to do. This is backed by looking further into the occurrences of 0 logo transitions. Almost all of them are limitations and challenges discussed in section 5.2. Thus, the aim is to exploit the logo transitions when attempting to produce compelling clips.

Besides these points, we wish to understand what constitutes the booking event itself. For example, if the camera zooms, or how they picture the foul. This will help us conceptualize the protocol of the current highlight clips.

3.1.2 State-of-the-art Clipping Procedure

When analyzing the current clips, we look at the last 50 clips from the 2022 season in both Allsvenskan and Eliteserien [5, 11]. In each clip we review, we search for notable events in presenting, like zooming and logo transitions. When noticing a regular notable event in some clips, we decide to implement it into the provisional protocol and count the occurrences in the clips. This way, we can determine which details the producers consider necessary and exciting for the public.

When reviewing the 100 clips, it became apparent both leagues have almost identical protocols, even though their designs and camera positions are diverse. With many occurrences each, these are the details we consider to be in the protocol:

Foul This is counted if we see the build-up to the foul and/or only the foul itself. We only count it as an occurrence if we see it in live action, even if there are some less ordinary fouls like mouthing and shirt celebrations. The foul in replays and in slow-motion is not counted.

Zoom This is counted if there is a scene change after the foul to a close-up shot of one or more active parts in the foul leading to the booking. It could be the player committing the foul, the player being fouled, or the referee observing the situation. It has to be zoomed in drastically from the standard camera to count.

Logo transition (Replay) This is the logo transition, mainly transpiring before replays and slow-motion shots. It is counted if the clip has one clear and complete logo transition. In other words, the clip can end before the replay is over as long as the first logo transition has transpired and is completed.

Referee showing card (Zoomed) This is counted when the referee shows the card, but only when the camera is zoomed in and focusing on this detail. It could be in the same shot as the zoom detail; one does not exclude the other.

Animation This detail is shown in Figure 3.2 and is the designed animation showing which player that was booked. It usually appears right below the scoreline and is counted if it is visible.

The most central detail is the close-up when the foul is committed, with a rate of 98% and 94% in Eliteserien and Allsvenskan, respectively. Logo transitions into replays and seeing the foul happen in real-time are details occurring in around 80% of the clips. The other two details animation and the referee showing the card reach lower rates of 40-60%.

These rates suggest the motion of the game and capturing player reactions are of higher significance than the more informative details like animation and the referee showing the card. However, it can also be justified by the limitations of booking events. There can be many reasons why the cameras and production of the game cannot capture all these details within a short enough span for a highlight clip. For example, a delayed yellow card means we are unable to capture both the card's handout and the foul happening in real time. Additionally, upon further investigation of the full games in Eliteserien, we see that the animation occurs more often than the highlight clips show us.

3.2 Conceptualization

Based on the statistics in the previous section, we can conceptualize it as a protocol the producers follow when a booking event happens. The current highlight clips are merely standard time intervals of the game where the event occurs. In addition, the clips often contain several or all of the details pointed out in the previous section. As the nature of booking events can vary, we need to present two different conceptualizations. The first one is the structure of the details when a standard event transpires, while the other one contains fewer details because of various situations in a game, shortening the event. These details and situations are discussed in section 5.2.

As figure 3.2 shows, the foul must happen before the camera and producers act. Rapidly after the foul, the cameras often zoom in on either of the players involved in the event, with the referee occasionally being amidst the situation. Then, if the break is prolonged, they do a logo transition into and out of a replay of the foul where the cameras zoom in on it. After a short stretch, when they are sure the booking has transpired, they do animation below the scoreline, like in Figure 3.3. This is a summary of the standard procedure when a booking transpires. However, there were some uncommon cases where only one or two of the five details in a standard clip happened due to a shortened event.

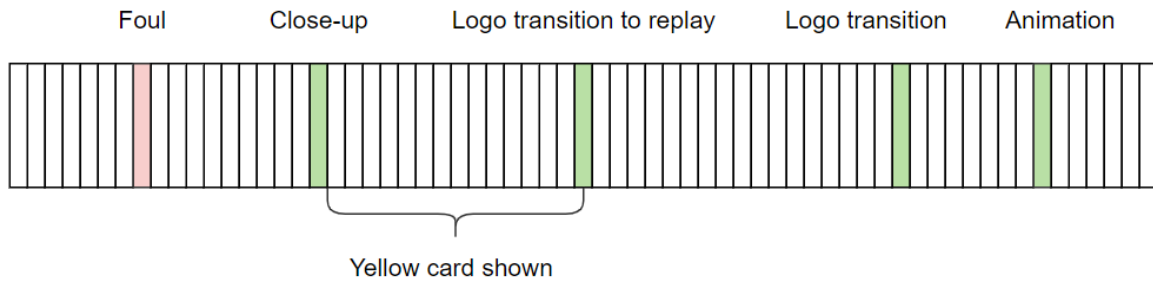


Figure 3.2: The conceptualization of the average [5, 11] highlight clip produced for booking events.



(a) Animation of yellow card in Eliteserien. From: [11].



(b) Animation of yellow card in Allsvenskan. From: [5].

Figure 3.3: The animation appearing after bookings, which we aim to recognize when detecting booking events.

The mentioned shortened event happened in 11 out of the 100 events we analyzed as challenges in the game transpired. Some of the distinct variables make it difficult to follow this protocol at all times. For example, stopping a counter-attack and time-wasting bookings often results in a quick game resume, which rejects replays for in-game play. Another problem is when an event transpires out of the focused area. For example, if there is an altercation between managers or players not participating in the present play. Sometimes camera operators cannot switch focus swiftly enough, making the event harder for viewers to follow. The producers also struggled when a player celebrated a goal by pulling the shirt off, resulting in a yellow card. The goal replay would be prioritized instead of seeing the referee handing the card or zooming in on the player pulling the shirt off. Thus, we have attempted a second part of the conceptualization, often occurring after a described issue, for example.

In Figure 3.4, the shortened version often only shows a close-up after the foul. Then, for example, abruptly returning to the live game after a quick restart. However, the producers usually notice the booking and do animation, as shown in Figure 3.3.

Using these conceptualizations, we can decide which techniques to gather and deconstruct the valuable videos and data from them. For example, according to the high replay rate and logo transitions into them, we should use logo recognition to extract the timestamps of replays. Furthermore, as the statistics show, zooming often happens, thus giving a higher rate of scene boundaries. Thus, we should attempt scene boundary detection as well. Finally, we should also be able to detect animations, as these statistics show almost after each booking event.

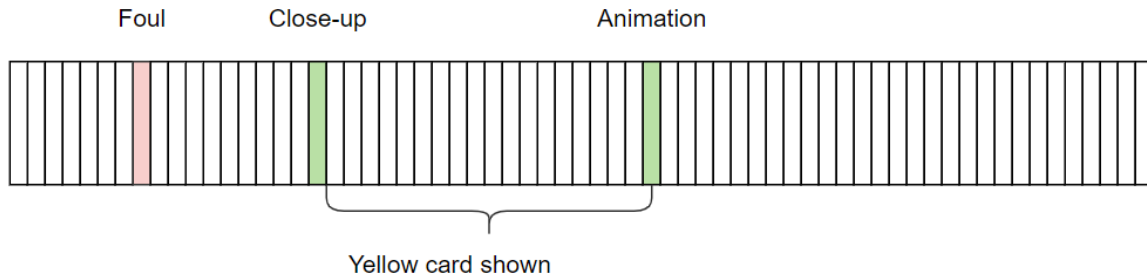


Figure 3.4: A shortened version of the conceptualization if a clip is shortened due to circumstances in the game.

3.3 Datasets

Proper datasets are essential when training and evaluating machine learning models. This is because they form the foundations upon which the model learns to recognize patterns and make predictions. The data quality is vital for hitting good accuracies and producing a reliable model. To satisfy Forzasys' needs, we must utilize data from Eliteserien and Allsvenskan.

3.3.1 Overview

In this thesis, we are accessing highlight clips from Eliteserien (2022) and Allsvenskan (2022), as well as full games from Eliteserien (2020) and goal-scoring events from Allsvenskan (2022). The current highlight clips from Eliteserien (2022) and Allsvenskan (2022) are the standardized clipped events on the corresponding web pages [5, 11] with length averaging 25 seconds, with a few being longer and shorter. These clips are only used for reviewing and analysis. The full games from Eliteserien (2020) and the Allsvenskan goal-scoring events are utilized more deeply. The Allsvenskan goal-scoring event is averaging 30 seconds in length, while the full game is usually between 120-130 minutes. We must manipulate the videos to make them valuable for technologies and models. Firstly, we divided these .mp4-files into frames. Then, due to a lack of automatic partitioning, we had to manually split the images containing logos and the pictures of the rest of the game. Finally, we divided it into three sets; training, validation, and testing.

3.3.2 Logo detection dataset

The logo detection dataset is made using clips from Eliteserien booking events and Allsvenskan clips of goal events. The Eliteserien booking events are gathered from the full games in Eliteserien. They are easy to extract, given a text-file managing annotations of the events following the video file of the game. We employ the open-cv library in python extracting frames from the videos to gather images for the datasets. Figure 3.5 shows some of the frames containing logo transitions. As we see, the frames include gradual transitions based on the Eliteserien logo. These frames begin with the diagonal blue line from the top left to the bottom right, while the logo appears in the middle. Then, the diagonal transition is gone toward the end, and the logo gradually moves to the right and out of the frame.



Figure 3.5: A small collection of frames during the logo transition.

3.3.3 Image properties

The frames from Eliteserien clips have a width and height of 1280×720 , while the Allsvenskan has a width and height of 960×540 pixels. We should later resize these values to input sizes reaching good results in similar studies [46, 47]. We will test the input sizes of $72 \times 72 \times 3$ and $108 \times 192 \times 3$ for the neural network, and explore which input values perform best on the test set.

And after manually having to divide the frames into a training set, validation set, and test set, we have gathered 8110 game images and 894 game images. This gives us a total ratio of 9.07. However, as the raw clip we use utilize in the pipeline is 60 seconds, and we expect two logo transitions in each clip with an average stretch of 25 frames, we end up with a ratio of 29.2. All info available in table 3.1. We planned to keep the test set and validation set equal in size. However, to properly test the network, we bumped the number up in the test set as we wished to see which frames the network struggled to detect.

Dataset	Game frames (G)	Logo frames (L)	Ratio (G/L)
Train	4693	693	6.77
Validation	533	85	6.27
Test	2884	116	24
Expected Input	1460	50	29.2

Table 3.1: The distribution of frames in each set.

3.4 Implementation

The pipeline is developed using Python version 3.10, TensorFlow version 2.10.0, Keras version 2.10.0, and NumPy version 1.24.2. The specs of the computer which implemented the pipeline have Intel(R) Core(TM) i5-10210U CPU @ 1.60GHz - 2.11 GHz processor and 8GB RAM.

Some tasks in the pipeline are dependent on a few installed libraries. We will list them here:

OpenCV OpenCV (Open Source Computer Vision) is a popular Python library for computer vision and image processing applications. It provides a wide range of functions for image and video processing. In this thesis, it helps us split video files into frames, for example.

FFmpeg FFMpeg is a library used for video and audio processing. It provides various functionalities such as video and audio conversions, format conversions, filtering, and streaming. FFMpeg can be used with other video processing libraries like OpenCV and MoviePy.

matplotlib Matplotlib is a Python library used for data visualization and plotting. It provides many functions for creating charts and plots, such as line plots, bar plots, and histograms. We will utilize it to present some data, like confusion matrixes.

NumPy NumPy is a Python library used for numerical computing and data analysis. It provides powerful array computing functionality for manipulating large arrays and matrices. It also provides various mathematical functions such as linear algebra, Fourier transform, and statistics.

MoviePy MoviePy is a Python library used for video editing and manipulation, built on top of other libraries such as FFMpeg and NumPy. With MoviePy, we can cut, concatenate, and add video audio. Altering video speed and size is also doable. In our pipeline, it will be used to retrieve shorter videos of booking events from the full games explained in section 3.3

3.5 Logo Detection

An essential goal of the pipeline is to recognize logo transitions into replays in the match, considering the statistics in section 3.1. With Kadragovic and Valand reaching great results with their logo detection component on parallel datasets, we have gathered pieces of their model [46, 47]. Their overall approach is to use a CNN to recognize the logo in a frame and then see if there is a stretch of consecutive detections to eliminate false positives. Eliteserien and Allsvenskan allocate 24 and 27 frames for each logo transition. Hence, surpassing a threshold of continuing detected frames makes wrongly recognized pieces almost nonexistent. We will also apply a slideshow to visualize the stretch of logo frames in the clip.

3.5.1 Results from Valand and Kadragic

Valand and Kadragovic [45–47] tested a few different implementations of CNN to see which version performed the best. They measured CNNs efficiency with precision, recall, and F1-score as metrics. They also estimated the computational cost because some implementations were slower and more expensive due to different amounts of convolutional layers, for example. The implementations they tested were simple-CNN, Residual Network (ResNet), VGG-inspired, and Support Vector Machine (SVM).

This thesis focuses on finding a cheap version of CNN that still delivers good results in measuring metrics. If we are looking at pure computational cost, the simple CNN is vastly superior, according to figure 3.2. Performing at around 340.000 fps with a DGX2 server, this version is much better than the second best, which is VGG-inspired at 80.000-90.000 fps.

Looking at the performances through the measuring metrics, VGG-inspired CNN with 54×96 pixels in grayscale performs best on the Eliteserien dataset. It

Model	Input	FPS
ResNet	$144 \times 256 \times 3$	1,798
ResNet	$108 \times 192 \times 3$	3,117
ResNet	$54 \times 96 \times 3$	12,295
Simple CNN	$144 \times 256 \times 3$	340,936
Simple CNN	$108 \times 192 \times 3$	341,897
Simple CNN	$72 \times 72 \times 3$	339,099
VGG inspired	$144 \times 256 \times 3$	94,169
VGG inspired	$108 \times 192 \times 3$	65,281
VGG inspired	$54 \times 96 \times 3$	86,594
VGG inspired	$144 \times 256 \times 1$	96,428
VGG inspired	$72 \times 72 \times 1$	85,722
SVM (Simple CNN)	$108 \times 192 \times 3$	179
SVM (Simple CNN)	$27 \times 48 \times 3$	14,023
SVM (Simple CNN)	$72 \times 72 \times 3$	1,103
SVM (Simple CNN)	$72 \times 72 \times 1$	883
SVM (VGG16)	$108 \times 192 \times 3$	442
SVM (VGG16)	$72 \times 72 \times 3$	3,172

Table 3.2: Execution time measured on DGX2 server 3.3.1. From [46].

reaches 100% precision and recall. However, taking a closer look shows us the other versions also shows promising results, with almost all other being at 97% to 99% in precision and recall. These excellent results show us practically all of their tested implementations can be fitted as a part of the logo detection module.

The discussion again states that all models performed well given the simple dataset. And, as our dataset is even more straightforward with only the logo of Eliteserien and Allvenskan apparent, we could choose any of the models. Therefore, we aim to implement the computationally cheapest model, which is the most accessible and the least time-consuming. The goal is to implement the simple CNN with input types of $72 \times 72 \times 3$ and $108 \times 192 \times 3$ as they reach good results.

3.5.2 CNN-Model

The pipeline bases its CNN model for logo detection on Valand and Kadragovic’s development [46]. Their results show promising results on several CNN implementations, as mentioned in the previous subsection. However, there is an essential change in the dataset. For bookings, only the league’s logo shows in the logo transition. Thus, it should be easier for the models to recognize them when the many different teams’ logos are not apparent. Therefore, we are exploring simple CNN, given the leading computational cost. And we are predominantly using $108 \times 192 \times 3$ as the input modes.

The implementation of the CNN model is presented in Figure 3.6, where we start with frames in the mentioned input size. We then apply two convolutional layers with 32 filters of 3×3 ., both preceding max-pooling of 2 strides of 2×2 . The 2D convolutional layers use ReLU as its activation function to further lower the model’s computational value. Then, we flatten the model and advance the data to the fully connected layer, which utilizes ReLU. This fully connected layer

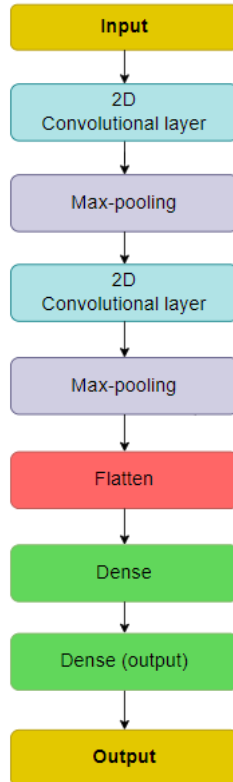


Figure 3.6: Simple CNN model.

consists of 128 input neurons. Lastly, we use a new fully connected layer with just one neuron, which operates the sigmoid function to convert the values between 0 and 1. Ultimately, we have a neural network receiving images as input which can binarily predict them. The total trainable parameters included with $108 \times 192 \times 3$ as the input size with RGB and the other filters and neurons are 4,720,801.

3.5.3 Training and parameters

When training the simple CNN, we use the Adam optimizer, which uses the Glorot uniform initialization to initialize the weights. We are also using binary cross-entropy as the loss function for the model. We are thoroughly explaining these techniques in section 2.2.

We train the simple CNN using card events from Eliteserien full games (2020) and goal events from Allsvenskan (2022). These datasets are explained in section 3.3. We save the trained model, which should be able to predict if a frame is either part of a logo transition or a frame of the game.

We are using the default learning rate of 0.001 following the use of Adam optimizer. The batch size is 32. We use Keras as the high-level API for building the network, and we set the max epochs to 20. Due to a relatively small dataset, we decide not to define steps per epoch.

3.5.4 Evaluation

Evaluating the network for logo detection, we will apply it to predict a series of clips on booking events from full games. We will revise the annotations and find the statistics on true positives, false positives, true negatives, and false negatives. The results will show how well the network detects logos in booking events.

3.6 Scene boundary detection

SBD is used to detect scene changes, and we can utilize it to decide where to clip in our pipeline. Kadragovic and Valand [46] researched the TransNetV2 framework presented by Soucek and Lokoc [37], which performs scene change detections reaching good results on varied datasets and videos. We aim to use Valand and Kadragovics results when implementing SBD in our pipeline.

3.6.1 TransNetV2

TransNetV2 [37] is a state-of-the-art scene boundary detection framework taking advantage of deep learning. The framework detects scene changes using CNNs and recurrent neural networks (RNNs) to analyze video frames and track objects over time. Implementing the framework on various levels is possible based on wishes and goals. For example, it comes with pre-trained weights and out-of-the-box possibilities based on datasets from ClipShots and TECVid IACC.3, and it is also possible to replicate the framework and train on other datasets. Valand and Kadragovic tested both the TransNetV2 model on their dataset with training and evaluation and the out-of-the-box version to see how the versions performed.

3.6.2 Pre-trained TransNetV2

The goal of the pre-trained version is to be easy to use and promotes efficient handling of large datasets. The model trains and evaluates with different big datasets. Approximately 15% of the videos come from ClipShots, while the rest is videos fabricated with synthetic transitions from the TECVid IACC.3 dataset [37]. With these datasets, the model trains on several transitions; hard cuts, gradually dissolved, and gradual transitions. It is easy to implement, as we only need to clone the TransNetV2 into our project and install it as a python package. Then, it is ready to use by calling the prediction function and adding the desired video.

3.6.3 Architecture

The model builds on a Stacked Dilated Deep Convolutional Neural Network (SDDCNN) with several DDCNN layers and spatial average pooling in order. In this case, it is 6 DDCNN layers, each consisting of four $1 \times 3 \times 3$ spatial convolutions succeeding four $3 \times 1 \times 1$ temporal convolutions with 1, 2, 4, and 8 as dilutions. Looking at the figure, we see three blocks have two cells of DDCNN and an average spatial pooling of $1 \times 2 \times 2$, which halves the spatial dimension [37].

After the DDCNN layers, it implements RGB histogram and learns similarities between frames. The histogram decides a similarity score for each frame based

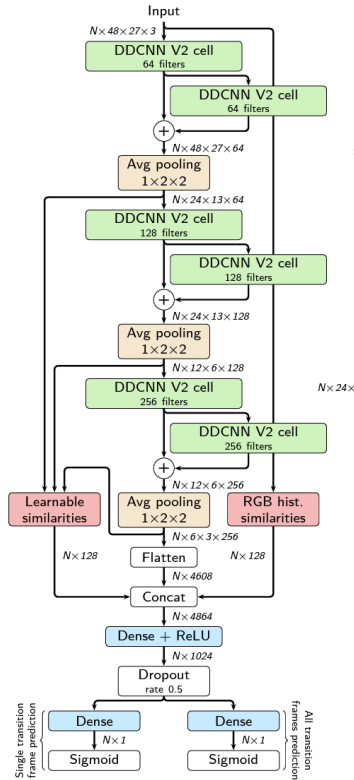


Figure 3.7: The architecture of the Stacked Dilated Deep CNN [37].

on the values of red, green, and blue colors in the pixels of the image. There are concatenated pooling layers responsible for the similarity score, which then is passed through a dense layer and further has its cosine similarity computed. Finally, it outputs a similarity score for the frame compared to the neighboring frames.

Lastly, the model has multiple classification heads, which are two prediction heads. The first head predicts only one frame in a transition, while the other head is trained to predict all the frames in the transition.

3.6.4 Training and Evaluation (Technical details)

The two classification heads is trained using standard cross-entropy loss averaged on batch. The single-frame and all-frame heads are weight by a factor of 5 and 0.1, respectively. L2-regularization of 0.0001 is also added to the loss. SGD optimizes the loss function with momentum set to 0.9 and a fixed learning rate of 0.1. They have trained the convolutional neural network for 50 epochs with 750 batches sized 16 [37].

Also worth noticing is the data augmentation Soucek and Lokoc did. They use both color and filtering possibilities from Tensorflow and the python image library PIL, as well as flipping images both left to right and top to bottom with different probabilities. However, they augment each frame in a shot in the same manner to prevent fake scene boundaries in an input sequence [37].

The training of the network took over 17 hours using a single Tesla V100 16GB GPU, and is therefore a considerable reason for utilizing the pre-trained version of it. Acquiring annotated data and utilizing the trainable version would therefore be extremely time-consuming, in regard to the span of the short thesis [37].

metric	Trained on SoccerNet		Pre-trained	
	$\delta = 4$	$\delta = 24$	$\delta = 4$	$\delta = 24$
Precision	93.80%	96.46%	97.85%	99.05%
Recall	84.09%	86.47%	97.85%	99.05%
F1 Score	88.68%	91.19%	89.33%	90.43%

Figure 3.8: The evaluation on a SoccerNet validation dataset. Source: [46].

weights	metric	All	Abrupt	Grad	Logo	Abr.&Grad.
SoccerNet	Prec.	95.66%	98.37%	97.66%	76.80%	98.22%
Pre-trained		95.96%	98.73%	98.69%	77.26%	98.72%
SoccerNet	Rec.	80.35%	96.63%	87.51%	33.24%	94.56%
Pre-trained		80.67%	95.58%	89.19%	36.06%	94.13%
SoccerNet	F1-score.	87.34%	97.49%	92.31%	46.40%	96.36%
Pre-trained		87.65%	97.13%	93.70%	49.17%	96.37%

Figure 3.9: The final evaluation of pre-trained vs. SoccerNet-trained TransNetV2. Source: [46].

3.6.5 Shortcomings and misclassifications

As scene boundary detection is difficult, we should expect to hit less than 100% in the evaluation metrics. Most of the misclassifications are due to either poor annotations in the dataset or difficulties in detecting smooth transitions. [46, 47] In general, smooth transitions may not surpass the threshold to predict a scene change. Thus, we should investigate which shortcomings we prefer for our pipeline.

3.6.6 Results from Valand and Kadragovic

Valand and Kadragovic tested the pre-trained TransNetV2 and used the model to train on their datasets of Eliteserien and SoccerNet clips. [46, 47] The evaluation is based on abrupt and gradual transitions.

When training, they use a dataset of clips from SoccerNet, specifically the 16/17 season of the Premier League. They run the training with 50 epochs and use a tolerance between 4 and 24, meaning a prediction must be within 2 and 12 frames. These are the configurations that gave them good outcomes, as the score in figure 3.5 depicts.

The results are overall positive for the pre-trained version of TransNetV2, as Figure 3.8 illustrates. When using the SoccerNet dataset for testing, the pre-trained version scores better in all classes except for abrupt transitions. This transition has a better recall score in the trained version. But as the thesis correctly concludes, precision is the most crucial metric when the recall is good. Good precision yields fewer false positives, which are more severe than true negatives because false positives can fool the system with non-existent scene changes. As a consequence of the pre-trained version scoring better precision on a dataset containing soccer clips, this is the version we intend to prioritize in the thesis.

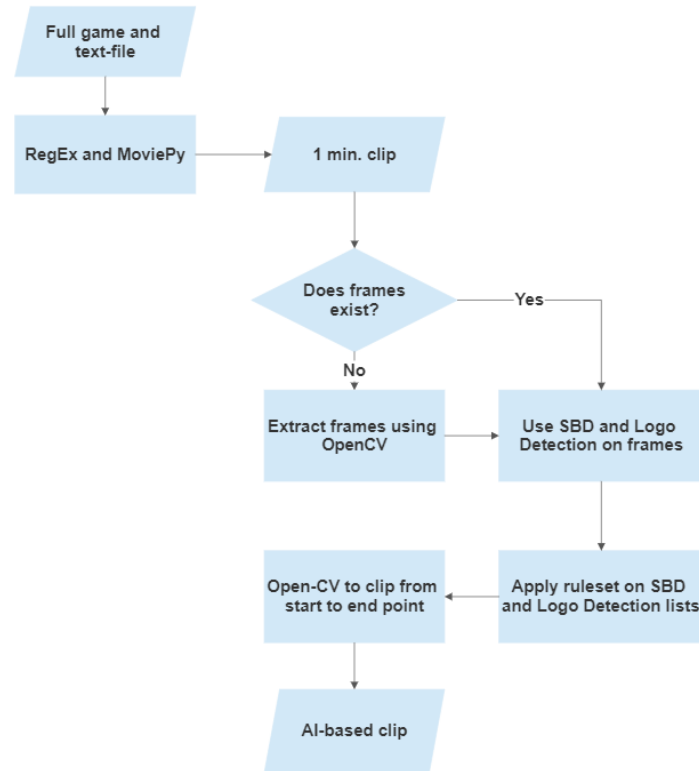


Figure 3.10: A flowchart of the pipeline.

3.7 Pipeline

So far, we have accounted for each tool we have implemented in the pipeline. Therefore, we must discuss using the tools and networks to produce compelling videos. First, we explain the flow of the pipeline and in which order the tools are implemented. Then we will present the ruleset we use to decide the highlight clips' start and end points.

The chart in figure 3.10 summarizes how the data flow through the pipeline, and we use a full game to create compelling clips of booking events. First, we use regular expressions to extract the timestamps of the bookings from the text file and use MoviePy to cut a clip of 1 minute with 30 seconds on each side of the annotated event. Then, we use Open-CV to retrieve each frame from the video; 25 fps gives 1500 frames per video. When the images are processed, we can apply the network to list when a logo transition ends or when a scene change occurs. Lastly, we use the ruleset below to decide the best starting point, and add the frames between starting point and end point together into a clip.

3.7.1 Ruleset

The ruleset we apply to the final part of the pipeline helps us clip starting and endpoints of the clips. We base the rules on the logo transitioning and clipping on a scene change at a time when action occurs. This, is also backed by the statistical analysis discussed in sections 3.1 and 4.1. Thus, we have different rules with either one, two, or no logo transitions apparent.

We decide on a starting point by looking at the statistics, where most action happens. The static clipping uses 10 seconds before the annotated booking. Most action occurs in this part, so we implement the clipping rule on the first scene change if it is within 12 seconds prior to the annotation. This attempts to catch the audience's attention while finding a good starting point, not in the middle of a shot.

The number of logo transitions influences the ending point. If two logo transitions exist in a clip, we use the last logo transition as the ending point as it ends the replay and often goes back to the open play. This means the action and most exciting details have already been shown. If only one logo transitions exist, we count two scene changes after the logo transition. The analysis statistics show that one or two shots after the logo transition is a replay, which is a wanted detail to include. If there is no logo transition apparent after a booking event, the clip is usually limited, as discussed in section 5.2. Then we follow the static and clip 15 seconds after the annotation.

This renders us with a fallback option of 12 seconds prior, and 15 seconds after the annotated clip. It resembles static clipping, which can catch most of the action occurring in the limited clips.

3.8 Subjective Evaluation

To get quantifiable data on how well the pipeline performs compared with the current highlight clipping, we run a survey comparing AI-based highlight clips versus statically clipped highlight clips. We will describe which data we will use, followed by how the survey is set up. We will also explain the pre- and post-questionnaire and what insights we wish to receive from the survey.

The clips to the survey are a way to test the pipeline, which is why we use the same input data. It is full game videos averaging ca. 2 hours, and its metadata is from a text file. The data is described in section 3.3. The pipeline uses this input to produce a highlight clip. We use the same annotation as found in the metadata text file to create a statically clipped clip 10 seconds prior to the annotated card event and 15 seconds after. Thus, we have two videos of the same event clipped differently.

To randomize the booking events, we chose three random full games handed from forzasys.com, which were not used for training and testing the networks in the pipeline. Furthermore, from these three complete games, we randomly picked ten events. Thus, we extracted AI-based and statically clipped highlights of the booking events in table 3.3. Each game is recognized by a stamp, where the first four numbers says which game it is and the last four says which second of the clip it is. In this case, the game 1271 is Bodø-Glimt-Stabæk, 1773 is Rosenborg-Stabæk, and 1824 is Brann-Tromsø.

We published the survey in Google Forms, and we anonymized the respondents. The main part of the survey had ten comparisons of each video pair, as seen in figure 3.12. Then, the user was asked to rate their general experience of each clip before deciding which one they preferred. The clips were randomized like in table 3.3, so the respondents did not know which clips were clipped in whichever way. Additionally, we had pre- and post-questionnaires. Before the central part, we asked about gender, age, and whether the respondents were familiar with soccer and video editing. The possible answers are in figure 4.5. As seen in figure 3.11, the post-

Video	Description	S	AI
1271-s5235	Yellow card, foul, close-up of player and replay	A	B
1773-s3188	Yellow card, tackle, close-up of referee and replay	B	A
1773-s5682	Yellow card, close-ups but only see foul in replay	B	A
1773-s5893	Yellow card, clip has all details	A	B
1773-s6145	Yellow card, cant see the foul and short clip	B	A
1773-s6569	Yellow card, misses close-ups	B	A
1824-s2328	Yellow card, missing foul	B	A
1824-s5906	Yellow for time-wasting, few details which is interesting	A	B
1824-s6076	Yellow card, misses foul and is limited	A	B
1824-s6449	Yellow card for mouthing, interesting with few details	B	A

Table 3.3: Descriptions of each video in the survey, also shows how they are randomized in the survey

questionnaires are related to the booking events. With the videos fresh in memory, they can answer what they feel contributes best to an exciting clip of the details from section 3.1 and how long they think a clip should be. We have also made optional comment rubrics in the survey if the respondents wish to clarify or add answers. This helps to gain deeper insights into the answers given.

Other options for the survey were also discussed. It could have been advantageous to have five comparisons preceded by the same ratings only for single videos. It could have given more honest reviews of the general experience, as the thought of comparing clips would not color the answer. However, we did not see this as an issue, and the comparisons are the main idea of the survey. We wanted to understand how the pipeline performed compared to todays clipping procedure.

3.9 Chapter Summary

To implement the pipeline for clipping highlights, we have accounted for all the tools and how they are implemented. The statistical analysis was key to finding the critical details to catch in a booking event. Replays, fouls in real-time, and close-ups were the most prevalent. Further, we discussed the datasets, the two neural networks we implement and then summarized them by presenting the order of the pipeline.

Both TransNetV2 and the simple CNN implemented leave us with reasonable computational costs and perform well enough to give compelling highlight clips. The following chapter will present how intriguing they are and additional results.

How important are the features below for a highlight clip?
Please rate from 1 (not important) to 5 (very important)


	1	2	3	4	5
Close-ups (zoom)	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>
Foul in real time	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>
Card handed by referee	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>
Replay	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>
Booked player information	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>

Do you find the benchmark duration of around 25 seconds OK?

- The clips should be shorter
- The duration is suitable
- The clips should be longer

Figure 3.11: The questions in the post-questionnaire.

Clip 1B



Clip 1B: How was your overall experience with this clip? *

Poor 1 2 3 4 5 Excellent

1 2 3 4 5

Which clip do you prefer? *

Clip 1A


Clip 1B

Comments (optional)

Svaret ditt _____

(a) 1

Clip 1B



Clip 1B: How was your overall experience with this clip? *

Poor 1 2 3 4 5 Excellent

1 2 3 4 5

Which clip do you prefer? *

Clip 1A

Clip 1B

Comments (optional)

Svaret ditt _____

(b) 2

Figure 3.12: An example of the question in the central part of the survey.

Chapter 4

Experiments and Results

This chapter will present the pipeline's findings and statistics implemented in the thesis. We will provide a detailed and thorough investigation of the data and results we have gathered from our experiments. Furthermore, the results from this chapter will lay the foundation for answering how, if feasible, to replace the statically standardized highlight clipping in soccer videos.

In chapter 3, we described the approach to improving the clipping for highlights in Eliteserien and Allsvenskan. To summarize, we started learning how they clip and produce the highlights today and condense this knowledge to a conceptualization. Then, we began implementing a pipeline to detect details like logo detection and scene changes to be able to edit the clips. Thus, this is also the order we seek to employ in this chapter.

Firstly, we will present the statistics and data from manually inspecting the current highlight clips accessed on the highlight pages of both Eliteserien and Allsvenskan [5, 11]. Then, we will attempt to interpret and understand the data to analyze and identify trends, patterns, and possible relationships. As section 1.6 describes, the statistical analysis is a novelty, which means it is vital to appropriately understand how the event transpires.

Then, we wish to show the accuracy of the CNN model in logo detection and the TransNetV2 implementation we did in scene boundary detection. Thus, we will use a confusion matrix to determine how well the CNN performed and see the scene boundary detection's precision, recall, and F1 score. The precision, recall, and F1 score is our predominant evaluation metrics for these key concepts, as they are valuable metrics for understanding and weighting the statistical results.

Finally, we will present the statistics from the survey comparing the newly clipped highlights to the statically clipped ones. The metrics from this survey are explained in section 4.4.

To sum up, the statistics from studying the current highlight clips are valuable and can be used to improve the models we implement and gain insights valuable for Forzasys and further development of clipping. It is also important to interpret the results from each neural network for possible improvements. The results from the final subjective evaluation will help us answer how the pipeline in its entirety is performing, thus it will aid in answering if the research targets in 1.2.

4.1 Statistical analysis of booking events

4.1.1 Clipping analysis

This section will present the statistics of the current highlight clips from Eliteserien and Allsvenskan [5, 11]. We have manually analyzed and reviewed the clips regarding several details discussed in section 3.1. When investigating the clips, we checked if the details were apparent, and also examined technical aspects like number of shots and length per shot. The technical aspects we looked at are described in section 3.1.1

These statistics are an integral part of the results, and we aim to provide comprehensive insight into what a highlight of a booking event consists of.

The table 4.1 present the details most often occurring in a highlight clip of a booking. It shows how often the details occur in the 50 highlight videos analyzed in each league. Further presentation is in section 3.1

Detail	Eliteserien	Allsvenskan
Foul	80%	88%
Logo	72%	80%
Zoom	98%	94%
Card	50%	46%
Animation	48%	64%

Table 4.1: Statistics of the details in the analyzed clips from the current clipping in Eliteserien and Allsvenskan.

4.1.2 Video-technical analysis

To gain insights into how the highlight of a booking event is produced, it is interesting to review the technical aspects such as number and length of the shots in a clip. Therefore, the analysis was done with the same clips as in the previous subsection regarding technical aspects, such as measuring the number of shots and the average duration of the clips. We explained these technical details are in section 3.1.1

In table 4.2, we see how many clips in each league consisted of the number of shots. We manually counted the shots in each clip, and ended up with a majority of clips in the range between 4 and 6 shots. In figure 3.1 we showed the pattern we recognized in this range of shots. Further motivation of how we utilize this information is found in section 3.1

We also counted the frequency of logo transitions, which can be seen in table 4.3. We notice the most occurrences is clearly 2, followed by 1, and then some occurrence of 0. The one occurrence of 3 logo transitions can be overlooked, as it happened so far ahead of the foul it would not be captured by the pipeline.

To conclude, these statistics underpin the dynamics of soccer and the challenges arising when clipping the highlights. As opposed to highlights for goals where the games are paused for a while regardless of the event, this is why the approach for bookings is exciting and necessary to research.

Number of shots	Eliteserien	Allsvenskan
1	0	0
2	2	1
3	5	4
4	12	6
5	12	19
6	13	13
7	2	4
8	3	1
9	1	0
10	0	2

Table 4.2: The number of different shots (scene changes) in a highlight clip from the current clipped highlights.

Number of logo transition	Eliteserien	Allsvenskan
0	7	10
1	21	11
2	22	28
3	0	1

Table 4.3: The frequency of logo transitions in each clip.

4.2 Logo Detection

4.2.1 Input values and logo detection

To provide insightful results of the logo detections from the implemented CNN, we will use precision, recall, and F1 scores. We will also present the results in a confusion matrix, showing how many correct predictions the neural network makes. These evaluation metrics will reveal how many correct frames with logos and games the implemented model can predict.

We tested two different inputs for the neural network, which expected the best results [47]. $108 \times 192 \times 3$ and $72 \times 72 \times 3$ are the two different input types with similar-looking results. In this confusion matrix, we find the true labels on the left axis and the predicted labels on the bottom axis. If the CNN predicts correctly, the upper left and bottom right boxes should dominate.

		True labels	
		Game	Logo
Predicted labels	Game	2869	15
	Logo	15	101

Table 4.4: A confusion matrix of the predicted frames in the test set from the CNN with $108 \times 192 \times 3$ as input.

In the matrixes, both input types predict all the images labeled game reasonably precisely. However, $108 \times 192 \times 3$ is a more precise input for predicting the logo class, with 101 correct versus 88. Further, we see it wrongly detects logos more often,

		True labels	
		Game	Logo
Predicted labels	Game	2879	5
	Logo	28	88

Table 4.5: A confusion matrix of the predicted frames in the test set from the CNN with $72 \times 72 \times 3$ as input.



Figure 4.1: The start of the logo transitions, recognized by the cap in upper left corner.

but it wrongly detects fewer game frames than $72 \times 72 \times 3$.

Input type	Precision	Recall	F1 score
108 X 192 X 3	0.994	0.990	0.994
72 X 72 X 3	0.998	0.989	0.994

Table 4.6: Precision, recall, and F1 score for the different input types in the CNN for logo detection.

The evaluation metrics display promising results on the test set, consisting of two full yellow card clips with different teams. Our model can tolerate lower recall as we have a threshold of finding the continuous frames of the logo transition. Despite this, we cannot have too many incorrectly predicted frames, but a recall of 0.99 is satisfactory for detecting the transitions.

Considering the promising precision, recall, and F1-score of both the input values, we must estimate which errors the model tolerates the best. When predicting a game on a logo frame, the error can lead to miss logo transitions. But upon further investigation, it is regularly the 2-3 earliest frames in the logo transition that the model misses. These frames, in figure 4.1, only have a small percentage of the transition in the upper left corner. This is the starting point which is difficult for the model to recognize.

On the other hand, when detecting logos in a game frame, other errors are likely happening. For example, without a high threshold of continuous logo frames, we can suddenly add a whole logo transition incorrectly. It can mess up a whole highlight clip and have more severe implications than the other misclassification.

Therefore, we are more tolerant of the misclassifications done by the CNN with $108 \times 192 \times 3$ as input. Because the precision, recall, and F1 score are almost



Figure 4.2: A frame wrongly predicted as logo transition.

identical, and it detects the logos more correctly, this is the preferred input for the network.

4.2.2 Logo transition detection

Further, we can measure how many logo transitions the model finds when looking at all the videos used in our survey. With these results, we can look at how well the threshold is when finding the threshold of consequent logos to define a stretch/transition. For example, in the ten videos added to the survey, the model found all the logo transitions and sequences of logo frames. This means the threshold we set on 11 continuous frames is reasonable for finding the correct stretches. It is also sufficient to overlook misclassifications and helps the model tolerate the wrongly detected logo. However, they are more prone to errors when frames are mispredicted. Also, a higher threshold is equally prone to error as it will be challenging to define sequences even with minor mispredictions.

We have experimented with lower thresholds, such as 9 and 5. These values were tested in the same video clips used in the survey, explained in section 4.4 to check if it detected the same logo transitions. It were quite successful, but for some occurrences it misclassified two logo frames to game frames, resulting in three detected logo transitions where it only was one. Thus, it rendered the generated clips completely useless. On that basis, we use a threshold of 11.

4.2.3 Limitations

The limitation we see in the confusion matrixes 4.5 and 4.4 is where the game frames are wrongly predicted as logo frames. When examining these frames, we notice that the common factor is the majority of blue, both in the foreground and background, as exemplified by figure 4.2. A reason can be that the dataset is too shallow, and the model is unable to differentiate between similar-looking colors. It can also be addressed through different preprocessing of the images by adding more contrasts and normalizing them.

The other limitation is that the early frames of the logo transition are detected as game frames, as mentioned earlier in this section. This can be due to the simplicity of the CNN. By adding layers and neurons, the network can be more receptive to minor features, such as the little blue cap in the upper left corner at the start of the logo transition [52].

4.2.4 Summary

After looking at which errors each of the input values was prone to, the $108 \times 192 \times 3$ became the preferred input for the network. We can detect most of the logo transitions happening in the highlight clips with high precision, recall, and F1 score and the threshold of 11 when detecting sequences of continuous logo frames. On this basis, we can use the implemented neural network to successfully aid the pipeline in deciding how to clip exciting booking events.

4.3 Scene Boundary Detection

The SBD is the next part aiding the pipeline in clipping the highlight clips. TransNetV2, which we utilize, is a state-of-the-art neural network detecting scene changes in videos. By using TransNetV2, we can find different scene changes helping us find suitable locations to clip. More advanced details about the neural network can be found in section 3.6.

To measure how well the TransNetV2 performs, we can calculate how many scene changes it can detect when knowing the ground truth in some videos. We know the network cannot detect all scene changes; specifically, it struggles with transitions happening over several frames [37, 47]. Despite this, booking events primarily consist of abrupt scene changes as the camera and producers wish to capture the action. The events also have some occurrences of smooth transitions. Nevertheless, since they range over very few frames, usually 3 or 4, and are very few, we decided not to separate between abrupt and smooth transitions as scene changes. The logo transition is another smooth transition the network struggles to pick up. However, they are trivial as we have another module in our pipeline detecting it. Therefore, the neural network should perform well on our clips.

The clips we will use to measure TransNetV2s performance are the same ten clips used in the survey, which are booking events gathered from the full Eliteserien games discussed in section 3.3. Below, we can see the precision, recall, and F1 score when manually measuring how many scene changes the network detected correctly.

Logo transition	Precision	Recall	F1 score
Not counted	0.987	0.79	0.877
Counted	0.987	0.97	0.98

Table 4.7: Precision, recall, and F1 score for the neural network, TransNetV2, utilized on Eliteserien dataset.

The network reaches pleasingly recall and precision scores. And the results resemble other experiments done with TransNetV2 [37, 47]. The network performs well and can detect 73 out of 92 scene changes in these ten clips. When delving



Figure 4.3: Two examples of how a scene change happens, but TransNetV2 does not detect it. The network detect scene change between a and b, and c and d.



Figure 4.4: 4 frames showing a wrongly detected scene change. Scene change detected between subfigure b and c.

deeper into which scene changes the network miss, we notice 16 of the 19 it misses are logo transitions we are already detecting with high precision in the pipeline. This means the network realistically only misses 3 of 92 wanted scene boundary detections. Looking at these scene changes in figure 4.3, we observe equal backgrounds in the frame and many equal-colored pixels. The difference between the frames does not surpass the threshold in which the network measures the probability of a scene change, discussed by Soucek and Lokoc [37, 38].

Further, the model impresses on the false positives, where it only wrongly predicts scene change once. This is essential for the pipeline, as we wish to avoid clipping on random timestamps in a video. Which further can give a weird and incomplete experience for the soccer audience. In figure 4.4, we see a sudden change in background color. This make TransNetV2 reach the threshold of color change, and thus wrongly predicts a scene change. This is an example of how outside noise can influence the networks performance. TransNetV2 also impresses on false negatives, reaching 14907 correct predictions of no scene change. The network is trained on massive datasets and is well-adjusted to sense if the succeeding frame is within the similarity score explained in section 3.6. The false negative results is essential for the F1-scores in table 4.7

Experimenting with TransNetV2 on Eliteserien and Allsvenskan datasets has yet to be done. As a result, we could have gained poor statistics with higher amounts of false positives and true negatives. Then it could have been necessary to train the network on a dataset of only Eliteserien frames. However, this would have required a massive effort in harvesting the data, as it requires a new system or must be done manually.

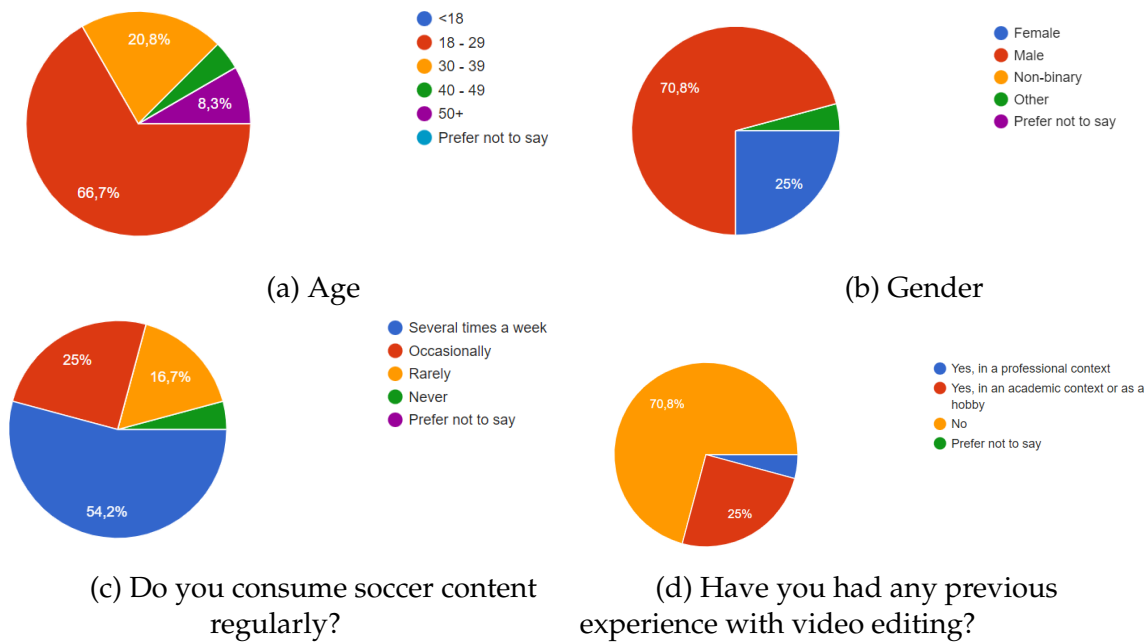


Figure 4.5: Answers from the pre-questionnaire in the subjective evaluation, the questions are in the caption for each subfigure.

4.4 Subjective Evaluation

The survey described in section 3.8 aimed to compare statically clipped soccer highlights versus AI-based clipped highlights. We had a range of questions from which we gained interesting results.

To explore the survey chronologically, we look at the results from the pre-questionnaire in figure 4.5 firstly. We asked questions about gender, age, familiarity with soccer, and familiarity with video editing. These piecharts show us a relatively young group of respondents, with 00% in the 20-29 age range. Further, the majority, 00%, is males, and 00% watch soccer often or occasionally. This means the audience understands and is used to seeing booking events. Finally, 00% of the respondents are familiar with video editing. However, we have respondents in several age ranges, both females and males, and with different experiences in both soccer and video editing. To generalize, we have a wide range of respondents, but with an emphasis on young soccer-interested males.

In the figure 4.8 we see which video of static or AI-based was decided as the best for the booking event. We see an even distribution, telling us the AI-based is competing with som clips being decided as better. However, there is still development needed to outcompete the statically clipping totally.

Further, we can gain insights by looking at the average score in general experience from how each clip is rated. Figure 4.6 displays how each clip is rated from 1 to 5 based on the general experience. The low-scoring clips often experience challenges and limitations mentioned in section 3.1. Looking at table 4.6, we know the clips E, F, G, and I are missing details, meaning the pipeline struggles to find good start and endpoints. It is also reflected in one of several comments: "Clip A onlly show the replay and it is not possible to see whcih player who gets the card. Clip B should have gave a better overview of the situation first, than the close up of the

Video	Static	AI-based
1271-s5235		X
1773-s3188	X	
1773-s5682	X	
1773-s5893	X	
1773-s6145	X	
1773-s6569		X
1824-s2328	X	
1824-s5906		X
1824-s6076		X
1824-s6449	X	

Table 4.8: Which clip of static and AI-based were decided as the best by the majority in the comparison

tackle." Thus, there should be potential for improvement here. When looking at the scores overall, we see the AI-based clips score better than static for clips A, H, I, and J. We also notice potential improvements for the other clips due to low scores related to the statically clipped videos. Interestingly, the AI-based has outscored clips that has all details, few details, and have clear limitations because of time-wasting and mouthing. This means the the AI-based are competing with, and potentially replaces the static in the future. Further discussions is in chapter 5.

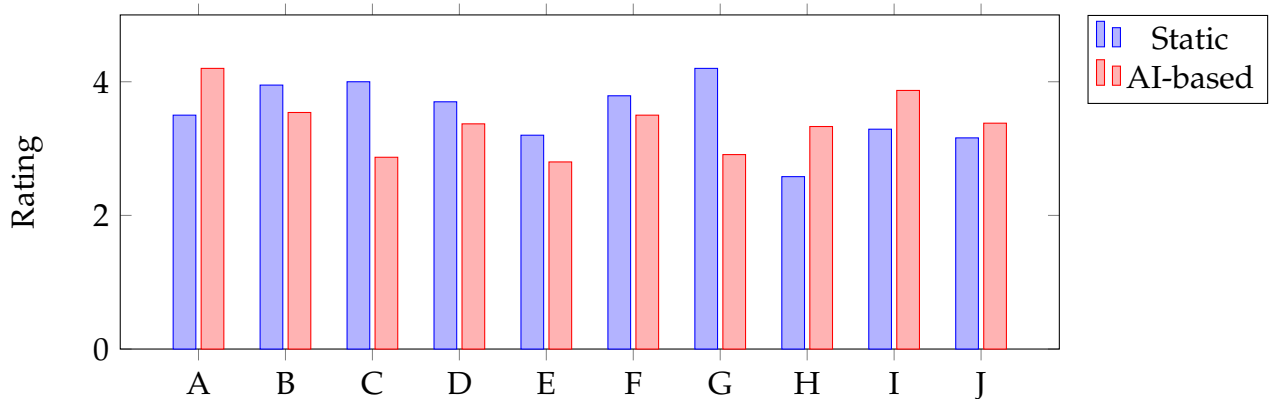


Figure 4.6: Average general experience from each video.

In the post-questionnaire, the respondents were asked about the importance of details in highlight clips and if the benchmark duration of around 25 seconds is suitable. The questions can be found in figure 3.11. The results from the details question are in figure 4.7, and we see how many have rated each detail with a value between 1 and 5, where 5 is the most important. Firstly, the replay, foul, and close-up have many high ratings of 4 and 5. This is expected if we compare it to how often the details occur in the current highlight clips, found in figure 4.1. However, the card handout and animation have many ratings of 3 and 4. Thus, we understand that the audience also finds it meaningful to include them. Even though the card handout has the most ratings of 2, it also has many ratings it 5. Looking at the rest, it is easy to conclude that the importance is very subjective to each person. There are many variances as all details have ratings from 2 to 5.

When asked about duration, most respondents, 88%, were happy with the benchmark duration of around 25 seconds, as seen in figure 4.8. However, there were some comments we should note for future work. One respondent answered: "The duration of 25 seconds is good, as long as the first 3-5 seconds catch your attention". As we saw in section 1.1, we know the attention span for humans is around 8 seconds and still decreasing. Thus, we must consider the exciting part when developing new clipping systems. Some other comments were also adamant that it is situation-based. 25 seconds is enjoyable if the relevant action is compressed into this time. However, too much unnecessary time of open play quickly makes the clip tedious. To conclude, a duration of 25 seconds is primarily suitable, but it is situation-dependent. Thus, further development of AI-based clipping for highlights could make the duration more dynamic, thus following the game's dynamism.

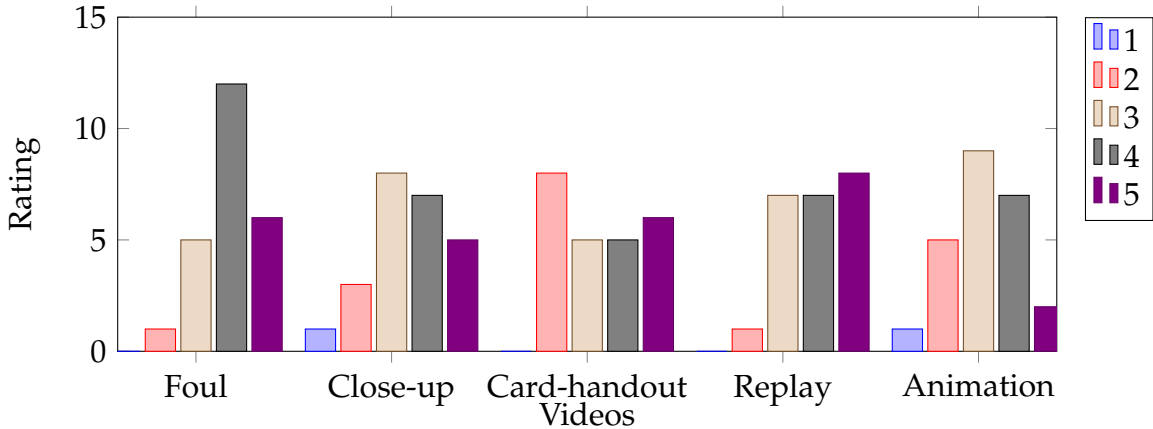


Figure 4.7: Results from the subjective evaluation where the importance of each detail in a clip is rated 1 to 5, where 5 is the most important.

4.5 Chapter Summary

In this chapter, we have experimented and seen how well the different tools in the pipeline have performed. We have analyzed how the current booking events are produced and clipped. Further, we saw how well the logo detection module performs. And we have experimented with the scene boundary detection module.

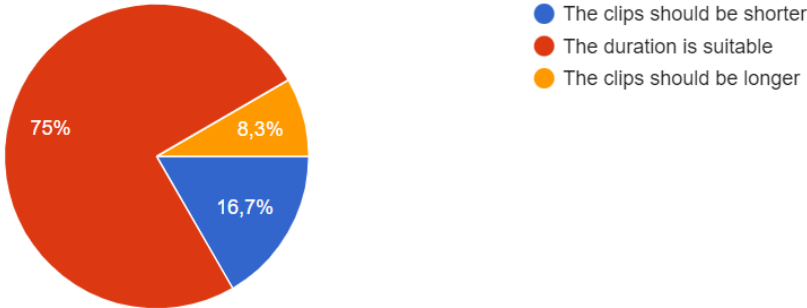


Figure 4.8: Results from subjective evaluation about the duration of clips.

These tools have shown promising results and good performances, essential for the pipeline to deliver good results and exciting highlight clips. As we see in table 4.7, realistically, we can reach precision, recall, and F1 score of 0.99, 0.97, and 0.98, respectively.

Finally, we extracted results from the user survey, which helped us gain insight into how the soccer audience experienced the clips. These particular areas of insight are great reference points for improving the pipeline. Therefore, the following section is reserved for discussing the survey answers and understanding where to improve.

Chapter 5

Discussion

5.1 Addressing the Research Questions

In the preliminary chapter, we proposed three research targets that we address based on the obtained results. To achieve these research targets' goals, we have described the backgrounds of machine learning and booking events. We then designed a methodology based on related works and the background chapter, incorporating qualitative and quantitative methods. Further, we experimented with the planned tools and machine learning techniques to observe their performance on datasets relevant to the Eliteserien and Allvenskan videos. And finally, we implemented the tools to a pipeline that automates the clipping of booking events and delivers highlight clips. We randomly chose ten booking events for the pipeline to clip and pushed a survey comparing automated clipping versus static clipping.

The results and answers from the survey are what support us in discussing the research targets. The targets were:

- **Target 1** Analyze current highlight clips and identify relevant data specific to bookings. The statistics should be intuitive and give insight into Forzasys' existing system.
- **Target 2** Design and implement a pipeline detecting and deconstructing a booking event from the annotated game. The pipeline should utilize the results of effective logo detection and scene boundary detection modules from the thesis of Valand and Kadragovic and other state-of-the-art research.
- **Target 3** Complete an evaluation of the new pipeline through automatically clipped highlights versus the current state-of-the-art in commercially deployed clipping protocols highlights. The evaluation should emphasize both how intriguing and how complete a clip is.

5.1.1 Target 1

The aim to gain insights and understanding about the current clipping system by Forzasys today was accomplished by manually inspecting 100 clips. We focused on details specified in section 3.1 when analyzing the clips. The results from this analysis are presented in section 4.1.

The approach to accomplish this target was to analyze the clips manually, as no system or program exists to assist in this analysis for Forzasys. Because reviewing the booking events is a novelty in this field of research, this quantitative study was adamant in finding what details the booking event consisted of.

Another approach could have been to request a description from Forzasys and the producers of the televised pictures of their task when a booking happened. What do they focus most on out of the details obviously apparent in such an event? Nevertheless, the main target was to get an overview of how often the details happened such that a conceptualization could be formed.

With the results from both Eliteserien and Allsvenskan being similar, the conceptualization in section 3.2 is efficiently designed. We have created two versions, one where the booking event transpires as expected. And the other one is where a limitation of the event has occurred, for example, an advantage or a booking for reckless mouthing. Thus, we have accomplished the first research target after successfully creating conceptualizations and have acquired an understanding of the booking event through a quantitative study.

5.1.2 Target 2

Our second research target was to design and implement a pipeline to detect and deconstruct details from annotated games or raw booking event clips. To accomplish this target, we have experimented with our logo and scene boundary detection modules.

According to the statistics from section 4.1, and the earlier mentioned conceptualization, it became evident that logo transitions, zoomed-in frames, and catching the foul in real time were the details emerging most often in the current highlight clips. Thus, these details are what the audience of Eliteserien, Allsvenskan, and many other leagues desire in a highlight of bookings. Therefore, we decided to implement logo transition and scene boundary detection to aid in automatically finding these points in the videos.

The two detection modules each performed satisfactorily to make them eligible for the pipeline, as seen in section 4.2 and 4.3. But are we happy with results not reaching 100%? This leads to a discussion of the use cases for the pipeline. As stated in section 1.1, all events and annotations in soccer must be noticed, as betting and statistics impact personal life and economics. However, this pipeline utilizes already annotated events or raw clips of already noticed highlights to clip them automatically. This means the eventual errors in our pipeline are not as dramatic as those named earlier. Therefore, we can tolerate some errors. Additionally, with satisfactory results in the evaluation metrics of both logo detection and SBD module, we do not experience too many errors.

With these results, we have constructed a pipeline that deconstructs a highlight into several possible timestamps for clipping. With the newly produced highlight clips receiving good scores in the evaluation, the pipeline produces intriguing and complete highlight clips. In this case, we have achieved the second research target.

5.1.3 Target 3

Our third and final target is to complete the pipeline evaluation through the newly clipped videos and compare them to static clips in a user survey. The survey should evaluate both how complete and intriguing a clip is.

The survey consists of a pre-questionnaire to gather information, the central part where they rate the videos, and a post-questionnaire to see what they find essential in the highlight clips.

The central part, where we ask them to rate their experience and to compare the statically clipped versus the automatically clipped highlight, helps us understand which clip is the most complete and intriguing. But, does the survey only provide insights for comparing the two clips? In hindsight, we could have added one or two extra questions about what they found either problematic or fitting to each clip. This could have helped us be more profound in defining what is necessary to include in a highlight of the booking event.

Furthermore, another approach could have been to run five comparisons of static versus automatic. And then done five of only the automatically clipped, with a different set of questions asking about more than the overall experience. Questions about the entertainment value, which segment in the video is most exciting, and how much of the play before the foul should be shown would help gain valuable insights.

However, a compact and effective survey keeps the respondents more perceptive when answering the questions. A too-long questionnaire is tedious and makes each answer less deliberate. Thus, it should have been contemplated further if the survey added supplemental questions or parts.

To conclude, we gain valuable insights into how complete and intriguing a clip is. Even though we could have added questions to gain further insights, we can claim the overarching research target was achieved.

5.2 Limitations

When proposing a pipeline leveraging ML, there are challenges and limitations arising. Both in the tools individually and in the harmony between them.

5.2.1 Booking Events

Soccer is a highly dynamic game and can be influenced by various factors. Booking events have the same freedom and constraints as the rest of the game. When analyzing these situations and clips, we must consider potential limitations that can impact the flow of the booking situation. For example, yellow cards for reckless mouthing, time-wasting, and the referee giving advantages are some booking events yielding very diverse durations of a highlight clip. Considering both the real-time foul and the logo transition into a replay, they can be over a minute between them after an advantage given by the referee. This example renders the proposed ruleset in the pipeline unable to produce a clip containing both details. However, the ruleset must be strict in order to provide exciting and intriguing clips for the regular booking events. This is a complex challenge to which further research may find solutions. Possible approaches to this problem are suggested in section 5.3.

The booking events are also prone to errors from the production and filming crew not catching the action on the pitch. In these cases, it is difficult to identify the key moments of a booking. Our pipeline relies on the presence of the details described in section 3.1 to provide compelling clips. For example, if the pipeline only detects a couple of scene changes and no logo transition, it can quickly drop to the fallback in the ruleset. This means it becomes a statically clipped highlight. Nevertheless, as the input is pre-determined in our case, there are no other approaches to ensure the proper production of the games.

Due to the dynamism of soccer, it is arguably too challenging to provide automated solutions for clipping the highlights in bookings. As a result, we may spend more time and cost attempting to automate the clipping than we use manually or statically clipping the highlights. On the other hand, employing tools that leverage dynamism could also be the future. Utilizing machine learning to deconstruct booking events can possibly tame the dynamism we see in the game. However, the proposed pipeline and ruleset for clipping need refining to tolerate all the different limitations and challenges that can emerge. Therefore, we can argue that the time and cost are burdens worth taking in the long run.

5.2.2 Dataset

Another limitation were the datasets discussed in section 3.3. When dealing with Eliteserien and Allsvenskan, we dont have too much data to pick from. Nevertheless, the biggest limitation were the need to manually curate a dataset of frames for the simple CNN in logo detection. It was time-consuming to move frames to the different sets. However, we reached good performance as seen in section 4.2. Another approach could have been to start with SoccerNet or other datasets, which has easier tools to retrieve data. But, as Forzadays works with Eliteserien and Allsvenskan, it was the predominant choice.

5.2.3 Logo detection

In the logo detection module, we have implemented a simple CNN. The computational cost is the main advantage of lowering the complexity of a neural network. Creating and posting the highlights as rapidly as possible should be the goal, as the interest and demand of seeing an event are highest instantly after the situation has transpired. Thus, a low computational cost can cut the time comparing this network to a more complex one. However, this must not happen at the expense of performance. The golden mean, in this case, is to reach low computational cost, while maintaining the accuracy in predictions [44]. In similar experiments, pleasant performance was reached with simple CNNs, SVMs, and VGG-inspired CNNs [47]. Although these experiments had similar datasets, our datasets are simpler, as described in section 3.3. Therefore, we were confident that the simple CNN could be sufficient for detecting the logos.

The results in section 4.2 suggest a fruitful performance of the neural network. Nevertheless, the trend of missing the two-three earliest frames in the logo transition shows potential for improving the network. The first approach would be to add some complexity to the network. As discussed in the previous paragraph, we can not lower the computational cost too much. However, a bit more complexity is fair to the

network. This could also be positive for scalability if it expands to additional leagues in the future.

Another approach could be to experiment with another dataset containing a more extensive set of frames. Since we have to harvest the data manually from clips, it is time-consuming. For future work, however, it would be advisable to spend more time providing an extensive enough dataset for the CNN to learn.

5.2.4 SBD

Using a state-of-the-art network that is extensively trained makes it too tough to propose a competing network in a short thesis. We have to rely on the performance demonstrated in resembling experiments [47], and by Soucek and Lokoc themselves [37, 38].

5.2.5 Approach

After reviewing the subjective evaluations, we can say the approach have been at least somewhat effective. We have been able to create compelling clips competing with the statically clipped clips. However, we see clear limitations and possible improvements in the ruleset, and how the details were treated. We spent a big portion of the time on finding the most important details in booking events, and as reviewing booking events rather than goal-scoring events it was crucial. However, we could have spent more time analyzing the details, and adjusted the ruleset better to the statistical analysis. With a more dynamic ruleset and knowledge of the details, we could have retrieved more of the interesting shots in the clips.

Else, following related works was smart in choosing the tools we worked with [37, 38, 46, 47]. We managed to reach good performance on a small dataset, which is due to the well-described results in earlier research. With more time we could have explored with the other details, like animation as well.

5.3 Open Questions and Future Work

This thesis has shown the potential for a machine learning approach to replace manual labor in clipping highlights of booking events. According to the survey, we saw a pipeline producing compelling clips competing with clips currently produced by Forzasys. The pipeline can perform even better and perhaps exceed the current version clipping with some adjustments.

In the pipeline, we added logo detection and scene boundary detection, which accounts for the details of logo transition, close-ups, and fouls in real time. The card hand-out and animation with information about the recipient are details that can be further refined through detection techniques. Such techniques could aid in deconstructing the clip better, enhancing the accuracy and efficiency of the process and making it more reliable. The more details of the conceptualization we can detect, the easier it will be to place them and create a compelling clip. However, adding networks for detection also adds computational cost. Therefore, a tradeoff study must be done between controlling the time it takes to produce the clip along with computational cost against accuracy and how compelling the clips are.

Another idea for future work is to add audio detection to the pipeline. State-of-the-art research sees results in measuring different types of audio in the detection of soccer events. Raventos et al. [31] used whistle detection to find events of interest in a soccer game. Gautam et al. [16] used audio intensity and in-game commentary in an effort to automate soccer game summarization, which shows potential for the future of audio-related automation. Thus, future work can incorporate these techniques to gain further insights into the booking events, making clipping easier. Even though it is a research field in development, it is exciting and relevant for additional work connected to this thesis.

We could also experiment further with the ruleset in the pipeline to remove some of the errors caused by the limitations mentioned in section 5.2. Our ruleset does not consider skipping parts between the start and endpoints in a video. With further insights potentially gained from detections in previous paragraphs, we could have enough control over the details to consider skipping the play when an advantage occurs. This way, the duration is still acceptably limited while we concentrate the clip to contain the most intriguing parts. It requires experimenting and the cost of computing, but it would be an exciting development if it were to succeed.

In summary, the insights gained in booking events are very interesting for future work. And the pipeline could benefit from incorporating additional detection techniques. However, it requires scrutinizing the tradeoff between cost, time, and accuracy.

5.4 Contributions

During this thesis, we have contributed to the research with a few different answers others can use and build upon.

Our statistical analysis provided valuable insights into the relationships between different details in a booking event, allowing us to answer our research questions with high confidence. With booking events being fresh in the research field, it was necessary to get statistical backing on the essential details of the event.

We also implemented a pipeline that automated most of the data processing, video editing, and decided the timestamps, making it easier for others to replicate our work. Even though the pipeline could not outcompete the static clipping, this experimental prototyping reached promising results, demonstrating the potential for further pipeline development. This pipeline is available as open-source software in <https://github.com/simula/forzify> and can help fellow researchers replicate the work.

Another contribution is the subjective evaluation showing the results in section 4.4. It was essential to get mainly quantitative but also some qualitative data showing what details and clips the AI-based handled well and could have handled better. In hindsight, we did the subjective evaluation very late, which gave us a period of stress at the end. The questions could probably have been further discussed to gain further insights. However, the answers gave a good portion of understanding and learning, helping us show what the future works can look like.

To summarize, we have contributed with statistical analysis, an open-source pipeline available at <https://github.com/simula/forzify>, and subjective evaluation showing how the pipeline clips performed on an audience. Furthermore, we can

say that we have contributed to the field by answering the research targets set in section 1.2.

5.5 Chapter Summary

In this chapter, we summarized our results and compared them to the research targets from section 1.2. We concluded the targets we set were generally achieved, even though we saw clear potential improvements. We also discussed the limitations and challenges of booking events, and how they impacted our results. These discussions led us to explain how the challenges can be explored in future work. We also added some interesting thoughts like adding audio, and additional details like animation.

Chapter 6

Conclusions

Today's task of manually clipping the highlights of booking events is tedious and time-consuming. As the company responsible for the clips wishes to post them rapidly after the event, the clips are often clipped by a preset time interval. This could result in poorly timed clipping, which makes it dull and deficient. Therefore, this thesis focused on exploring the automation of highlight generation through a machine-learning approach. We employed logo detection, scene boundary detection, and visual media manipulation to assemble a pipeline.

Through state-of-the-art detection models, we were able to reach promising results. In the logo detection module, we implemented a simple CNN and tested it with inputs $72 \times 72 \times 3$ and $108 \times 192 \times 3$ on an Eliteserien dataset. The results were very similar in reaching F1-scores of 0.99 for both inputs. However, looking at the results, it became clear that $108 \times 192 \times 3$ prevailed. For the Scene Boundary Detection (SBD) module, we used the pre-trained version of TransNetV2 [37]. We knew the model struggled with detecting transitions over several frames, so we had to do logo detection separately. If we remove these logo transitions from the results, TransNetV2 scored 0.98 in F1-score. Thus, both detection modules performed well and could be implemented into a pipeline.

The pipeline is constructed to take full Eliteserien-games and their metadata as input. Then, through video and image manipulation, the pipeline can employ the detection modules to give a list of frame numbers. The pipeline then uses a ruleset to decide the clip's start and endpoints given these lists, which it then produces a clip of. The code can be found here: <https://github.com/simula/forzify>

Furthermore, through a user survey, we gained insights into how a soccer audience experienced the pipeline-generated clips. These clips were compared to statically generated clips, and the results were divided. Anyway, we learned how to refine the ruleset to create more compelling clips, and we also gained insights into what is essential for an audience to see in a booking event.

The divided results from the comparison in the survey also show that it depends on how the booking event transpires in the game. There are many limitations and challenges, such as reckless mouthing, shirt celebrations, and waste of time which gives a variety of differences in the booking event. The dynamics of soccer make the clipping of a booking event difficult and exciting.

Thus, by answering all the research target from section 1.2. We have fulfilled the research target, and made it possible for future development to build on this thesis. To further conclude, automating the clipping of highlights in booking events is hard

as a range of events can occur. However, this thesis shows some promising results and has suggestions for improving the proposed pipeline.

Bibliography

- [1] *3 ways online video is changing what it means to be a sports fan*. 2018. URL: <https://www.thinkwithgoogle.com/intl/en-gee/marketing-strategies/video/sports-fans-video-insights/> (visited on 04/01/2023).
- [2] Oludare I. Abiodun, Aman Jantan, Abiodun Esther Omolara, Kemi Victoria Dada, Nachaat AbdElatif Mohamed and Humaira Arshad. 'State-of-the-art in artificial neural network applications: A survey'. In: *Heliyon* 4 (2018).
- [3] Saad Albawi, Tareq Abed Mohammed and Saad Al-Zawi. *Understanding of a Convolutional Neural Network*. Tech. rep. Department of Computer Engineering, Istanbul Kemerburgaz University, 2017.
- [4] *All you need to know about soccer*. n.d. URL: <https://www.bundesliga.com/en/faq/all-you-need-to-know-about-soccer> (visited on 06/01/2023).
- [5] *Allsvenskan | Highlights*. n.d. URL: <https://highlights.allsvenskan.se/> (visited on 02/2023).
- [6] 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 (CVPR)*. 2017, pp. 4724–4733.
- [7] Joao Carreira and Andrew Zisserman. *Quo Vadis, Action Recognition? A New Model and the Kinetics Dataset*. 2018. arXiv: 1705.07750 [cs.CV].
- [8] Microsoft Canada Consumer Insights. 'Attention spans'. In: ed. by Microsoft Canada. Microsoft, 2015, pp. 6–9. URL: https://www.sergiogridelli.it/wp-content/uploads/2015/07/AttentionSpans_report.pdf.
- [9] *Convolutional Neural Networks*. n.d. URL: <https://www.ibm.com/topics/convolutional-neural-networks> (visited on 24/01/2023).
- [10] Peter J. Denning, Douglas E. Comer, David Gries, Michael C. Mulder, Allen Tucker, A. Joe Turner and Paul R. Young. *COMPUTING AS A DISCIPLINE*. <https://dl.acm.org/doi/pdf/10.1145/63238.63239>. 1989.
- [11] *Eliteserien | Highlights*. n.d. URL: <https://highlights.eliteserien.no/> (visited on 02/2023).
- [12] *Europe's Gambling Revenues Stabilised Above Pre-Pandemic Levels In 2022 - New Data*. 2022. URL: <https://www.egba.eu/news-post/europes-gambling-revenues-stabilised-above-pre-pandemic-levels-in-2022-new-data/> (visited on 29/12/2022).
- [13] *Event*. <https://www.youtube.com/about/press/>. (Accessed on 30/03/2023). 2023.
- [14] C. Feichtenhofer, A. Pinz and A. Zisserman. 'Convolutional Two-Stream Network Fusion for Video Action Recognition'. In: (2016). URL: <https://ieeexplore.ieee.org/document/7780582/> (visited on 13/06/2020).

- [15] *Football club (association football)*. [https://en.wikipedia.org/wiki/Football_club_\(association_football\)](https://en.wikipedia.org/wiki/Football_club_(association_football)). (Accessed on 30/03/2023). 2023.
- [16] Sushant Gautam, Cise Midoglu, Saeed Shafiee Sabet, Dinesh Baniya Kshatri and Pål Halvorsen. 'Soccer Game Summarization Using Audio Commentary, Metadata, and Captions'. In: *NarSUM '22*. New York, NY, USA: Association for Computing Machinery, 2022, pp. 13–22. ISBN: 9781450394932. DOI: 10.1145/3552463.3557019. URL: <https://doi.org/10.1145/3552463.3557019>.
- [17] Silvio Giancola, Mohieddine Amine, Tarek Dghaily and Bernard Ghanem. 'SoccerNet: A Scalable Dataset for Action Spotting in Soccer Videos'. In: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR) Workshops*. June 2018, pp. 1711–1721. DOI: 10.1109/cvprw.2018.00223.
- [18] Lissie Hoover. *What Is Qualitative vs. Quantitative Study?* <https://www.gcu.edu/blog/doctoral-journey/what-qualitative-vs-quantitative-study>. June 2021. (Visited on 01/06/2021).
- [19] *How online video has transformed the world of sports*. 2019. URL: <https://www.thinkwithgoogle.com/consumer-insights/consumer-trends/sports-fan-video-consumption/> (visited on 04/01/2023).
- [20] Andreas Husa. 'Automated Thumbnail Selection for Soccer Videos using Machine Learning'. MA thesis. UiO, 2022.
- [21] H. Idrees, A. R. Zamir, Y. Jiang, A. Gorban, I. Laptev, R. Sukthankar and M. Shah. 'The THUMOS challenge on action recognition for videos "in the wild"'. In: *Computer Vision and Image Understanding* 155 (2017), pp. 1–23.
- [22] 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 (CVPR)*. 2014, pp. 1725–1732. DOI: 10.1109/CVPR.2014.223.
- [23] Muhammad Zeeshan Khan, Summra Saleem, Muhammad A. Hassan and Muhammad Usman Ghanni Khan. 'Learning Deep C3D Features For Soccer Video Event Detection'. In: *2018 14th International Conference on Emerging Technologies (ICET)*. 2018, pp. 1–6. DOI: 10.1109/ICET.2018.8603644.
- [24] Harilaos Koumaras, Georgios Gardikis, George Xilouris, Evangelos Pallis and Anastasios Kourtis. 'Shot boundary detection without threshold parameters'. In: *J. Electronic Imaging* 15 (Apr. 2006), p. 020503. DOI: 10.1117/1.2199878.
- [25] Tianwei Lin, Xiao Liu, Xin Li, Errui Ding and Shilei Wen. 'BMN: Boundary-Matching Network for Temporal Action Proposal Generation'. In: *Proceedings of IEEE International Conference on Computer Vision (ICCV)*. Oct. 2019.
- [26] Tianwei Lin, Xu Zhao, Haisheng Su, Chongjing Wang and Ming Yang. 'BSN: Boundary Sensitive Network for Temporal Action Proposal Generation'. In: *Proceedings of the European Conference Computer Vision (ECCV)*. 2018.
- [27] Meenal V. Narkhede, Prashant P. Bartakke and Mukul S. Sutaone. 'A review on weight initialization strategies for neural networks'. In: *Artificial Intelligence Review* 55 (2022).

- [28] Olav Andre Nergård Rongved, Markus Stige, Steven Alexander Hicks, Vajira Lasantha Thambawita, Cise Midoglu, Evi Zouganeli, Dag Johansen, Michael Alexander Riegler and Pål Halvorsen. 'Automated Event Detection and Classification in Soccer: The Potential of Using Multiple Modalities'. In: *Machine Learning and Knowledge Extraction* 3.4 (2021), pp. 1030–1054. ISSN: 2504-4990. DOI: 10.3390/make3040051. URL: <https://www.mdpi.com/2504-4990/3/4/51>.
- [29] Olav A. Norgård Rongved, Steven A. Hicks, Vajira Thambawita, Håkon K. Stensland, Evi Zouganeli, Dag Johansen, Michael A. Riegler and Pål Halvorsen. 'Real-Time Detection of Events in Soccer Videos using 3D Convolutional Neural Networks'. In: *2020 IEEE International Symposium on Multimedia (ISM)*. 2020, pp. 135–144. DOI: 10.1109/ISM.2020.00030.
- [30] Muhammad Rafiq, Ghazala Rafiq, Rockson Agyeman, Seong-Il Jin and Gyu Sang Choi. 'Scene Classification for Sports Video Summarization Using Transfer Learning'. In: *Sensors* 20 (Mar. 2020), p. 1702. DOI: 10.3390/s20061702.
- [31] Arnau Raventos, Raul Quijada, Luis Torres and Francesc Tarres. *Automatic Summarization of Soccer Highlights Using Audio-visual Descriptors*. 2014. arXiv: 1411.6496 [cs.LG].
- [32] Reede Ren and Joemon M. Jose. 'Football Video Segmentation Based on Video Production Strategy'. In: *Proceedings of ECIR - Advances in Information Retrieval*. 2005, pp. 433–446.
- [33] Shaoqing Ren, Kaiming He, Ross Girshick and Jian Sun. 'Faster R-CNN: Towards Real-Time Object Detection with Region Proposal Networks'. In: *Advances in Neural Information Processing Systems 28*. Ed. by C. Cortes, N. D. Lawrence, D. D. Lee, M. Sugiyama and R. Garnett. Curran Associates, Inc., 2015, pp. 91–99.
- [34] Olav Rongved. 'Automatic event detection in soccer videos'. MA thesis. UiO, 2020.
- [35] Zheng Shou, Dongang Wang and Shih-Fu Chang. 'Temporal Action Localization in Untrimmed Videos via Multi-stage CNNs'. In: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. 2016, pp. 1049–1058. ISBN: 978-1-4673-8851-1. DOI: 10.1109/CVPR.2016.119.
- [36] K. Simonyan and A. Zisserman. 'Two-Stream Convolutional Networks for Action Recognition in Videos'. In: (Nov. 2014). URL: <https://arxiv.org/abs/1406.2199> (visited on 13/06/2020).
- [37] Tomáš Souček and Jakub Lokoč. 'TransNet V2: An effective deep network architecture for fast shot transition detection'. In: *arXiv preprint arXiv:2008.04838* (2020).
- [38] Tomáš Souček, Jaroslav Moravec and Jakub Lokoč. 'TransNet: A deep network for fast detection of common shot transitions'. In: *CoRR* abs/1906.03363 (2019). arXiv: 1906.03363. URL: <http://arxiv.org/abs/1906.03363>.
- [39] *SuperVize Me: What's the Difference Between Supervised, Unsupervised, Semi-Supervised and Reinforcement Learning?* 2018. URL: <https://blogs.nvidia.com/blog/2018/08/02/supervised-unsupervised-learning/> (visited on 17/01/2023).

- [40] Dian Tjondronegoro, Yi-Ping Phoebe Chen and Binh Pham. ‘Sports video summarization using highlights and play-breaks’. In: *Proceedings of ACM SIGMM International Workshop on Multimedia Information Retrieval (MIR)*. 2003, pp. 201–208. DOI: 10.1145/973264.973296.
- [41] Du Tran, Lubomir Bourdev, Rob Fergus, Lorenzo Torresani and Manohar Paluri. ‘Learning Spatiotemporal Features with 3D Convolutional Networks’. In: *Proceedings of IEEE International Conference on Computer Vision (ICCV)*. 2015, pp. 4489–4497. DOI: 10.1109/ICCV.2015.510.
- [42] Du Tran, Lubomir Bourdev, Rob Fergus, Lorenzo Torresani and Manohar Paluri. *Learning Spatiotemporal Features with 3D Convolutional Networks*. 2015. arXiv: 1412.0767 [cs.CV]. URL: <https://arxiv.org/abs/1412.0767>.
- [43] 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 (CVPR)*. 2018, pp. 6450–6459. DOI: 10.1109/CVPR.2018.00675.
- [44] Amin Ullah, Khan Muhammad, Weiping Ding, Vasile Palade, Ijaz Ul Haq and Sung Wook Baik. ‘Efficient activity recognition using lightweight CNN and DS-GRU network for surveillance applications’. In: *Applied Soft Computing* 103 (2021), p. 107102. ISSN: 1568-4946. DOI: <https://doi.org/10.1016/j.asoc.2021.107102>. URL: <https://www.sciencedirect.com/science/article/pii/S1568494621000259>.
- [45] Joakim O. Valand, Kadragic Haris, Steven A. Hicks, Vajira Thambawita, Cise Midoglu, Tomas Kupka, Dag Johansen, Michael A. Riegler and Pål Halvorsen. ‘Automated Clipping of Soccer Events using Machine Learning’. In: *Proceedings of the IEEE International Symposium on Multimedia (ISM)*. 2021. DOI: 10.1109/ISM52913.2021.00042.
- [46] Joakim Olav Valand and Haris Kadragic. ‘Machine learning-based approach for automated clipping of soccer events’. MA thesis. UiO, 2021.
- [47] Joakim Olav Valand, Haris Kadragic, Steven Alexander Hicks, Vajira Lasantha Thambawita, Cise Midoglu, Tomas Kupka, Dag Johansen, Michael Alexander Riegler and Pål Halvorsen. ‘AI-Based Video Clipping of Soccer Events’. In: *Machine Learning and Knowledge Extraction* 3.4 (2021), pp. 990–1008. ISSN: 2504-4990. DOI: 10.3390/make3040049. URL: <https://www.mdpi.com/2504-4990/3/4/49>.
- [48] Haohan Wang and Bhiksha Raj. *On the Origin of Deep Learning*. Tech. rep. Language Technologies Institute, School of Computer Science Carnegie Mellon University, 2017.
- [49] 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: *Proceedings of the European Conference Computer Vision (ECCV)*. 2016, pp. 20–36. ISBN: 978-3-319-46484-8.
- [50] *What are neural networks?* n.d. URL: <https://www.ibm.com/topics/neural-networks> (visited on 24/01/2023).
- [51] *What is data labeling for machine learning?* n.d. URL: <https://aws.amazon.com/sagemaker/data-labeling/what-is-data-labeling/> (visited on 21/01/2023).

- [52] *What is machine learning*. n.d. URL: <https://www.ibm.com/topics/machine-learning> (visited on 17/01/2023).
- [53] Huijuan Xu, Abir Das and Kate Saenko. 'R-C3D: Region Convolutional 3D Network for Temporal Activity Detection'. In: *Proceedings of IEEE International Conference on Computer Vision (ICCV)*. 2017.
- [54] Peng Xu, Lexing Xie, Shih-Fu Chang, A. Divakaran, A. Vetro and Huifang Sun. 'Algorithms and system for segmentation and structure analysis in soccer video'. In: *Proceedings of IEEE International Conference on Multimedia and Expo (ICME)*. 2001, pp. 721–724. DOI: 10.1109/ICME.2001.1237822.
- [55] *YouTube for Press*. n.d. URL: <https://blog.youtube/press/> (visited on 05/01/2023).
- [56] Hossam Zawbaa, Nashwa El-Bendary, Aboul Ella Hassanien and Tai-Hoon Kim. 'Event Detection Based Approach for Soccer Video Summarization Using Machine learning'. In: *International Journal of Multimedia and Ubiquitous Engineering (IJMUE)* 7 (Jan. 2012).
- [57] Hossam M. Zawbaa, Nashwa El-Bendary, Aboul Ella Hassanien and Ajith Abraham. 'SVM-based soccer video summarization system'. In: *Proceedings of the World Congress on Nature and Biologically Inspired Computing*. 2011, pp. 7–11. DOI: 10.1109/NaBIC.2011.6089409.
- [58] Yi Zhang, Mengjia Wu, George Yijun Tian, Guangquan Zhang and Jie Lu. 'Ethics and privacy of artificial intelligence: Understandings from bibliometrics'. In: *Knowledge-based Systems* 222 (2021).
- [59] Fuzhen Zhuang, Zhiyuan Qi, Keyu Duan, Dongbo Xi, Yongchun Zhu, Hengshu Zhu, Hui Xiong and Qing He. *A Comprehensive Survey on Transfer Learning*. Tech. rep. Institute of Computing Technology, Chinese Academy of Sciences (CAS), 2020.