SmartCrop: AI-Based Cropping of Soccer Videos

Sayed Mohammad Majidi Dorcheh[‡§], Mehdi Houshmand Sarkhoosh[‡§], Cise Midoglu[*§], Saeed Shafiee Sabet[§], Tomas Kupka[§], Michael A. Riegler[*], Dag Johansen[¶], and Pål Halvorsen[*‡§]

[*]SimulaMet, Norway [‡]Oslo Metropolitan University, Norway [§]Forzasys, Norway [¶]UIT The Arctic University of Norway

***Abstract*—In the rapidly evolving landscape of digital platforms, the need for optimizing media representations to cater to various aspect ratios is palpable. In this paper, we pioneer an approach that utilizes object detection, scene detection, outlier detection, and interpolation for smart cropping. Using soccer as a case study, our primary goal is to capture the frame salience using object (player and ball) detection and tracking using AI models. To improve the object detection and tracking, we rely on scene understanding and explore various outlier detection and interpolation techniques. Our pipeline, called SmartCrop, is efficient, and supports various configurations for object tracking, interpolation, and outlier detection to find the best point-of-interest to be used as the cropping center of the video frame. An objective evaluation of the performance of individual pipeline components has validated our proposed architecture and the need for object, scene, outlier detection, and interpolation. Moreover, a crowdsourced subjective user study, assessing the alternative approaches for cropping from 16:9 to 1:1 and 9:16 aspect ratios, confirms that our proposed approach increases the end-user Quality of Experience (QoE).**

***Index Terms*—AI, video, cropping, aspect ratio, social media, association football/soccer**

I. Introduction

As the digital era progresses, the consumption of media has expanded beyond traditional platforms, with social media and various handheld devices emerging as primary outlets. Soccer, revered as the world's most popular sport, serves as a quintessential example of content that is consumed voraciously across various platforms, from large television screens to compact smartphones. Each platform and view mode might present a different aspect ratio, necessitating an adaptive presentation that remains consistent in its delivery of content to audiences, irrespective of the viewing settings [1].

Challenges emerge when considering dynamic content such as soccer games, where the ball, as the most central element, must remain consistently visible and undistorted. Traditionally, video cropping was performed manually using software tools like Adobe Premiere Pro, Final Cut Pro, or Avid Media Composer. These editors would select and maintain the area of interest in a frame-by-frame manner, making it a tedious and time-consuming operation. With the sheer volume of content and the demand for real-time broadcasting, this manual approach is unsustainable. Hence, the domain has witnessed a shift towards automated solutions. Deep learning models promise efficient video retargeting, yet the complexity of dynamic sports broadcasts poses unique challenges.

In general, we are aiming for automated production of events and highlight summaries, including event detection [2], automated clipping [3], thumbnail selection [4], text summarization [5] and now direct publication in various media. The goal of this work is to harness the capabilities of AI in an automatic cropping pipeline tailored for soccer highlights to be published on social media. This endeavor aims not just at maintaining the ball's visibility, but also at enhancing the overall viewing experience for users, ensuring that soccer fans worldwide receive the best possible visual narrative, regardless of their viewing platform of choice. We present SmartCrop, which works as an end-to-end pipeline for delivering differentiated media representations. It is relying on a fine tuned version of YOLOv8 for object detection, and tracking the ball through an extended logic including outlier detection and interpolation, for calculating an appropriate cropping-center for video frames, as shown in Figure 1. In short, results from both objective and subjective experiments show that SmartCrop increases the end-users' Quality of Experience (QoE).

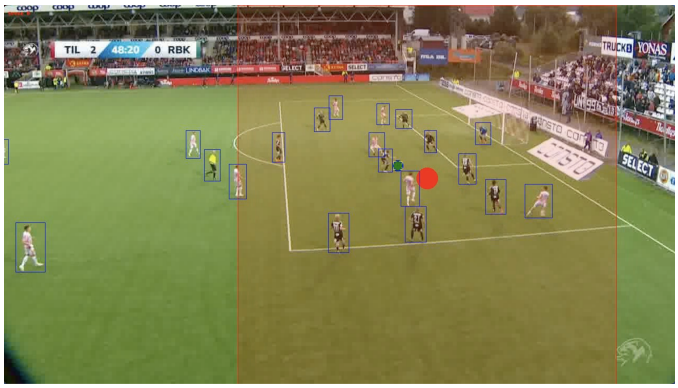


Fig. 1: Smart cropping using object detection and tracking. The red dot marks the calculated cropping-center point in the frame, and the red square marks the cropping area.

II. Background and Related Work

A. Video Aspect Ratio Adjustment

Algorithms tailored for the adjustment of the video aspect ratio can be broadly delineated into several categories. **Content-adaptive reshaping (warping)** lays emphasis on selectively modifying disparate areas of an image. By adopting a grid mechanism, pivotal regions of the image are retained, while the less significant zones undergo alterations based on deduced scaling parameters [6], [7]. **Segment-based exclusion (cropping)** isolates a specific portion of an image or frame, disregarding any visual constituents beyond the designated demarcation [8]–[11]. Moreover, **seam extraction** concentrates

on discerning and excising seams comprising non-essential pixels. Such interconnected lines of pixels within an image are judiciously pruned [12], [13].

Furthermore, hybrid techniques exist that amalgamate features from two or more of the foundational strategies, such as the integration of cropping with content-adaptive reshaping [14], or the fusion of seam extraction with segment-based exclusion [15]. A significant contribution has been made by Apostolidis and Mezaris [16], who proposed a method that leverages cropping to retarget videos to different aspect ratios, emphasizing its aptness when the priority is to minimize semantic distortions. Notably, they have presented the first-of-its-kind publicly available benchmark dataset for video retargeting, annotated by multiple human subjects, which serves as a pivotal resource for performance evaluations in this domain.

B. Object Detection

As media content grows in diversity and volume, the need for efficient, accurate, and real-time object detection models becomes paramount. **Single Shot Multibox Detector (SSD)** model [17] detects objects in media files through a single deep neural network. By discretizing bounding boxes into varying scales and aspect ratios, and eliminating redundant steps, SSD offers efficiency and speed to support real-time media applications, with benchmarks such as a 72.1% mAP on VOC2007. **Focal Loss for Dense Object Detection** [18] addresses the challenges of dense object detection in cluttered media scenes (crowded events, bustling city footage). This approach enhances accuracy. When combined with RetinaNet detector, it offers a balance between speed and precision. Moreover, the **Real-Time Object Detection with Region Proposal Networks (Faster R-CNN)** model [19] integrates a Region Proposal Network for enhanced efficiency. This integration significantly reduces detection time, proving invaluable for media applications that require immediate object localization, such as live sports broadcasts or instant content tagging. **YOLO** [20] marked a paradigm shift in how object detection was approached by treating it as a regression problem. Bounding boxes and class probabilities are predicted from images in one pass, positioning it as a preferred choice for live media broadcasts and real-time content analysis. By the model, media applications can be enabled to process up to 45 frames per second. In the context of this paper, **YOLOv8** has a specialized ball detector tailored to address the intricate demands of ball detection in dynamic environments. The unique challenges inherent in soccer broadcasts, such as swift player movements, sudden changes in camera perspectives, and frequent occlusions, necessitated adjustments to YOLOv8's neural network layers. The primary goal was to ensure the precise tracking of soccer actions throughout a game. Moreover, the model benefited from advanced data augmentation techniques that mimic various challenging in-game situations, aiming to bolster its generalization ability. Such refinements guarantee that, regardless of a scene's intricacies, the prominence of ball activities remains uninterrupted, enhancing the viewer's experience.

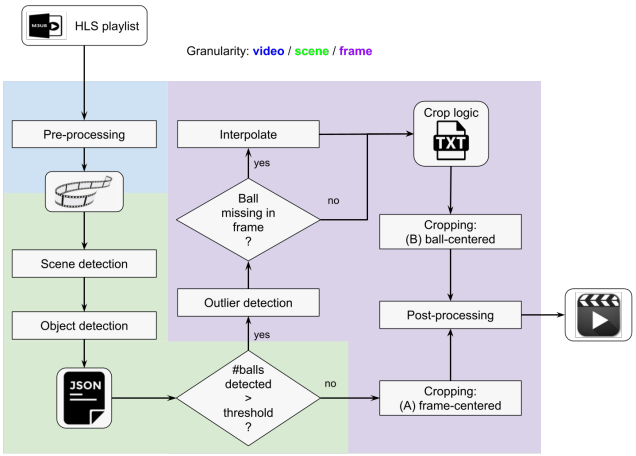


Fig. 2: SmartCrop pipeline overview.

C. Maintaining Points of Interest

While object detection models such as YOLOv8 offer promise in dynamic scenarios such as soccer broadcasts, ensuring the consistent visibility of central elements, like the ball, is non-trivial. Key challenges include **inconsistent detection**, where factors like rapid lighting changes, player occlusions, and camera angles can hinder the ball's accurate detection, even in advanced models like YOLOv8. Furthermore, **adapting to different aspect ratios** requires ensuring that the ball remains undistorted and centrally positioned. Techniques such as content-adaptive reshaping or seam extraction, if not well-integrated, can compromise visibility. Employing outlier detection and interpolation techniques to **rectify detection errors** introduces another layer of complexity. These, while meant to enhance accuracy, can occasionally introduce inaccuracies themselves due to the game's dynamic nature.

III. SmartCrop

The fundamental principle driving the SmartCrop pipeline is for the Point of Interest (POI) to be used as the center point of the cropping area (Figure 1). In our soccer scenario, the ball serves as the POI. When the ball is visible within the frame, it is the primary focal point for the cropping, when it is absent, we refine the state-of-the-art object detection and tracking techniques we employ with outlier detection and interpolation, to select an appropriate alternative focal point. As depicted in Figure 2, the SmartCrop pipeline consists of 7 modules with intermediate logic in between. These are detailed below.

A. Pre-Processing Module

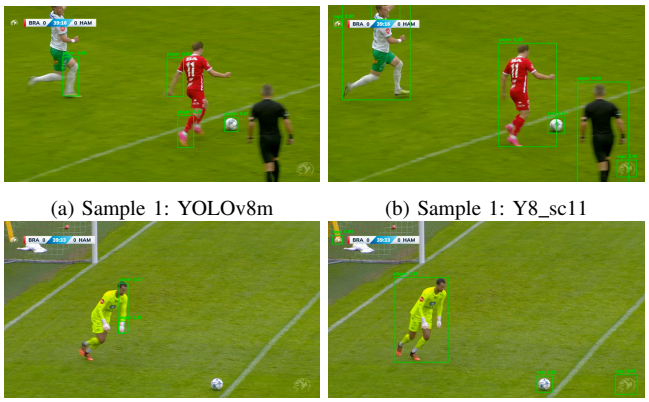
In general, all video can be processed by our pipeline, and here we have used HTTP Live Streaming (HLS) as input. Furthermore, to reduce the time spent on object detection, our pipeline searches inside the HLS manifest and selects the lowest quality stream available, which has the smallest resolution. This stream is then converted to H.264 encoded videos, encapsulated within an MP4 container. Then, the pre-processing model loads the requested (object and scene detection) models.

B. Scene Detection Module

SmartCrop processes the entire video and then narrows its focus to operate on individual scenes. The initial step following pre-processing is scene detection, facilitated by the TransNetV2 model [21]. This model segments the video into distinct scenes, allowing each to serve as a separate unit for further analysis such as interpolation and outlier detection. This targeted approach enables SmartCrop to adapt to the unique attributes of each scene, enhancing the precision of video cropping.

C. Object Detection Module

The object detection module of the smart cropping pipeline is facilitated through the integration of a YOLOv8 medium architecture model. We opted for the medium architecture after considering several trade-offs. Larger architectures offer slightly better performance but at the cost of significantly increased computational time, making them less ideal for real-time applications. On the other hand, smaller models like YOLOv8 nano provide quicker processing times but suffer from reduced accuracy, which is not suitable for the precision required in tracking dynamic elements like a soccer ball. The YOLOv8 model has been trained on a specialized soccer dataset, to detect key elements such as the ball and the players in soccer broadcasts, and is capable of identifying both the ball and the players, making it instrumental for our smart cropping logic.



(a) Sample 1: YOLOv8m (b) Sample 1: Y8_sc11

(c) Sample 2: YOLOv8m (d) Sample 2: Y8_sc11

Fig. 3: Visual comparison of the object detection performance of YOLOv8m and our Y8_sc11 model.

Due to the low object detection accuracy of the YOLOv8 (`Y8-M`) model, we curated a dataset of 1,500 images for retraining, which includes images from the Norwegian Eliteserien, and Swedish Allsvenskan and Superettan leagues, as well as 250 images from a publicly available soccer dataset [22]. As detailed in Table I and visually demonstrated in Figure 3, this approach led to significant performance improvements, notably with the (`Y8_sc11`) configuration.

D. Outlier Detection Module

This module uses various techniques to identify and remove anomalous data points from the processed information.

Scn.	Model	Ver.	Ep.	NMS	ImgSz	BatSz	Pat	Opt	LR0	LRf	Drp
1	Y8_sc01	N.	100	F	640	16	50	adamW	0.01	0.01	0.4
2	Y8_sc02	S.	100	F	640	16	50	adamW	0.01	0.01	0.4
3	Y8_sc03	M.	100	F	640	16	50	adamW	0.01	0.01	0.4
4	Y8_sc04	L.	100	F	640	16	50	adamW	0.01	0.01	0.4
5	Y8_sc05	XL	100	F	640	16	50	adamW	0.01	0.01	0.4
6	Y8_sc06	M	100	F	640	16	50	adamW	0.01	0.01	0.4
7	Y8_sc07	M	250	F	640	16	50	adamW	0.01	0.01	0.4
8	Y8_sc08	M	250	F	640	16	100	adamW	0.001	0.01	0.4
9	Y8_sc09	M	300	F	640	16	100	auto	0.001	0.01	0.5
10	Y8_sc10	M	200	T	1280	96	100	adamW	0.001	0.01	0.5
11	Y8_sc11	M	200	T	1280	96	100	adamW	0.001	0.2	0.5

TABLE I: Training scenarios for YOLO configurations.

Recognizing that the ball's position data may contain outliers, we incorporate three primary outlier detection methods to enhance the robustness of our system.

Average Method: An outlier is identified based on the average and standard deviation using:

$$\text{Outlier if } |x - \mu| > \alpha \times \sigma \tag{1}$$

Where x is the position of an individual ball, μ is the mean of all the balls, σ is the standard deviation, and α is a threshold, threshold sets the limit for outliers, often 2 or 3. In this method, x is an outlier if its scaled distance from μ exceeds α.

Z-Score Method: Outliers are identified by calculating the z-score using:

$$z = \frac{x - \mu}{\sigma} \tag{2}$$

Where z is the Z-score, x is the position of an individual ball, μ is the mean, and σ is the standard deviation. x is considered an outlier if $|z| >$ threshold, commonly set to 2 or 3.

Interquartile Range (IQR) Method: Outliers are identified using quartiles. The equation is:

$$\text{IQR} = Q3 - Q1 \tag{3}$$

Where $Q1$ is the first quartile and $Q3$ is the third quartile.

Outliers are defined as values outside the range:

$$[Q1 - k \times \text{IQR}, Q3 + k \times \text{IQR}] \tag{4}$$

Where k is a scaling factor, often 1.5 for mild outliers or 3 for extreme outliers.

E. Interpolation Module

After outlier detection has been performed, interpolation methods are employed to estimate the position of the ball in instances where it is not detected (i.e., to estimate missing or new data points in the dataset).

Linear Interpolation: To interpolate missing ball positions, between known ball positions (x_1, y_1) and (x_2, y_2), use:

$$x = x_1 + \frac{x_2 - x_1}{y_2 - y_1}(y - y_1) \tag{5}$$

Here, (x_1, y_1) and (x_2, y_2) are known ball positions. This formula estimates the X-coordinate (x) based on a known Y-coordinate (y) within this range.

Polynomial Interpolation: To interpolate missing ball positions, the polynomial functions $P(x)$ and $Q(y)$ are used:

$$P(x) = a_0 + a_1 x + a_2 x^2 \tag{6}$$

$$Q(y) = b_0 + b_1 y + b_2 y^2 \tag{7}$$

The coefficients a_0, a_1, a_2 and b_0, b_1, b_2 are calculated from known positions before and after missing frames. $P(x)$ and $Q(y)$ are then used to estimate the missing x and y positions.

Ease-in-out Interpolation using Sigmoid: To interpolate missing ball positions, the following equations are applied:

$$\text{eased\_t} = \frac{1}{1 + e^{-12 \times (t - 0.5)}} \tag{8}$$

$$x(t) = x_1 + \text{eased\_t} \times (x_2 - x_1) \tag{9}$$

$$y(t) = y_1 + \text{eased\_t} \times (y_2 - y_1) \tag{10}$$

Here, t ranges from 0 to 1, and x_1, x_2, y_1, y_2 are known positions before and after missing frames. The term eased_t provides a smooth transition between known points, influenced by the sigmoid function.

Heuristic-based Smoothed POI Interpolation: This method aims to address the limitations of simple interpolation techniques by introducing a dynamic approach speed β, which adapts based on the angle θ between the vectors of consecutive ball positions. The idea is to start slowly when approaching a new point of interest (POI) and then gradually speed up, especially if the POI is moving in a consistent direction. This makes the transition smoother and more realistic. The algorithm for calculating β and the new POI coordinates (x, y) is as follows:

$$\beta \begin{cases} \min\left(\frac{c_{\text{last}} + c}{120}, 0.20\right) & \text{if } \theta < 30^\circ \\ \min\left(\frac{c}{120}, 0.20\right) & \text{otherwise} \end{cases} \tag{11}$$

$$\Delta x = \min\left((x_{\text{next}} - x_{\text{last}}) \times \beta, 15\right) \tag{12}$$

$$\Delta y = \min\left((y_{\text{next}} - y_{\text{last}}) \times \beta, 15\right) \tag{13}$$

Here, β is the dynamic speed factor, θ is the angle between vectors of consecutive positions, c and c_{last} are speed constants, $x_{\text{next}}, y_{\text{next}}$ are the next known POI, and $x_{\text{last}}, y_{\text{last}}$ are the last known POI.

F. Cropping Module

Finally, cropping is used to isolate regions of interest within the images or videos, thereby reducing computational complexity and improving focus on key areas. This module crops each frame based on the *Aspect Ratio* parameter in the pipeline configuration giving the requested output format, using ffmpeg. To select the center of the cropping area, we have tested two basic options: 1) **frame-centered cropping** where the cropping center is statically selected in the middle of the frame; and 2) **ball-centered cropping** where the cropping center is based on the coordinates of the ball as detected and calculated by the detection and interpolation modules.

G. Post-Processing Module

This module creates an `.mp4` file from the cropped frames, and returns this file as the output of the pipeline. It also has additional functionality to prepare processed data for visualization, summarization, or further analysis. This involves restructuring the data and generating plots.

IV. Objective Evaluation

The goal of the objective evaluation is to investigate the performance of each module in the SmartCrop pipeline, and determine the best models/methods/techniques for each module to use in the final version (to be deployed in production).

A. Evaluation Metrics

We evaluate performance using Root Mean Square Error (RMSE) for interpolation and outlier detection, and the confusion matrix and Precision-Recall (PR) curve for object detection models. RMSE provides a direct accuracy measure, while the latter metrics offer comprehensive performance assessments. Together, these metrics facilitate a thorough analysis of the models and techniques.

B. Object Detection Performance

Table II compares the object detection performance of various YOLO configurations. Notably, `Y8_sc11` outperforms all other models, especially the default `YOLOv8-Medium`, in detecting balls, players, and logos. The confusion matrix for `Y8_sc11` further substantiates this point. `Y8_sc11` boasts an 82% true positive (TP) rate for balls and a remarkable 99% TP rate for both players and logos. Consequently, `Y8_sc11` emerges as the optimal choice for object detection within the smart crop pipeline.

The Precision-Recall curve in Figure 4 is an essential diagnostic tool for understanding the performance of our model over its training epochs. The y-axis represents 'Precision,' which gives us an idea of how many of the positively labeled samples are indeed positive. The x-axis represents 'Recall,' informing us how many of the actual positive samples were captured by our model. Our aim is to move towards the top-right corner of the plot where both precision and recall are high, indicating a well-performing model. In our experiment, we observed a generally positive trend in both precision and recall as the training epochs progressed. Starting from epoch 1 with a low precision of 0.71 and recall of 0.046, we noticed a gradual increase, reaching up to a precision of approximately 0.969 and a recall rate that also improved significantly by epoch 109. This indicates that our model is learning effectively, being able to correctly identify more

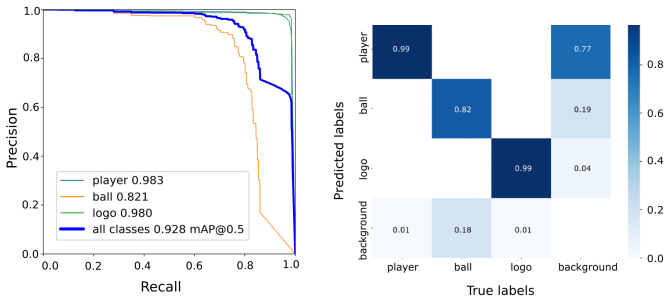


Fig. 4: Precision-Recall (PR) curve and normalized confusion matrix for the Y8_sc11 model.

positive samples (high precision) while also capturing a larger proportion of the total positive samples (high recall). However, it is important to note that there are fluctuations in the curve, signifying that some epochs result in a decrease in either or both metrics. This could be attributed to various factors such as the model's learning rate, the quality of the training data, or other hyperparameters. The curve serves as a guide for further fine-tuning, suggesting that while our model has learned effectively up to this point, there may still be room for improvement.

Model Name	Ball TP	Player TP	Logo TP
Y8_sc01	55%	97%	100%
Y8_sc02	61%	98%	100%
Y8_sc03	61%	98%	99%
Y8_sc04	61%	98%	100%
Y8_sc05	52%	97%	99%
Y8_sc06	63%	98%	99%
Y8_sc07	68%	98%	100%
Y8_sc08	73%	98%	99%
Y8_sc09	69%	98%	99%
Y8_sc10	81%	97%	95%
Y8_sc11	**82%**	**99%**	**99%**

TABLE II: Performance metrics for different YOLO models.

C. Outlier Detection Performance

In the conducted validation, balls retained as correct by each outlier detection method were assessed. These retained balls were subsequently compared with the ground truth to evaluate the efficacy of the respective outlier detection methods in eliminating incorrect ball detections.

Based on the validation results the RMSE for the *Average* outlier detection method was 150.74, for the *Z-score* method it was 120.82, and for the *IQR* method it was 120.20. The IQR method emerged as the most accurate for detecting outliers and has since been integrated into the smart crop pipeline.

D. Interpolation Performance

Method	RMSE
Linear Interpolation	163.33
Polynomial Interpolation	167.82
Ease-in-out Interpolation	172.72
Heuristic Interpolation	**137.56**

TABLE III: RMSE values for the interpolation methods.

Among the tested interpolation methods, Heuristic Interpolation outperformed others with the lowest RMSE value of 137.56 (Table III). Therefore, it can be concluded that the Heuristic Interpolation is the most accurate method for use in our smart crop pipeline. Figure 5a shows that the Heuristic Interpolation trend aligns most closely with the ground truth, visually confirming its superior accuracy.

E. System Performance

Configurations for local deployment: The local deployment was tested on a system with an NVIDIA GeForce GTX 1050 GPU, Driver version 535.54.03, CUDA version 12.2, and a memory size of 4096 MiB. Tests were conducted with an input video of 30 second duration, at 25 frames per second (FPS), and video resolution of 1280×720. Two different active configurations were tested to study the impact of Skip Frame on the pipeline execution time (`None` and 13), with detection confidence set to 0.4, target aspect ratio set to 9:16, and output format set to `MP4`.

Execution time analysis: Execution time for each module in the pipeline was measured under both configurations, and the results are represented in Figure 6. *Constant execution time:* The scene detection, outlier detection, interpolation, cropping, and post-processing modules exhibited consistent execution times across both configurations. This indicates that these modules are not significantly affected by the changes in the tested configurations. *Variable execution time:* Interestingly, the object detection module showed a considerable decrease in execution time from Configuration 1 to 2, dropping from $51.30s$ to $14.99s$. This is likely due to the Skip Frame parameter, set to 13 in Configuration 2, allowing the object detection algorithm to process fewer frames and thereby reducing execution time.

V. Subjective Evaluation

What is the best possible video clip is often subjective, and we have therefore performed a subjective user study where the participants assess alternative cropping methods. The goal of the user evaluation is to investigate the performance of the *overall* pipeline (as opposed to individual modules), in comparison to a number of benchmarks.

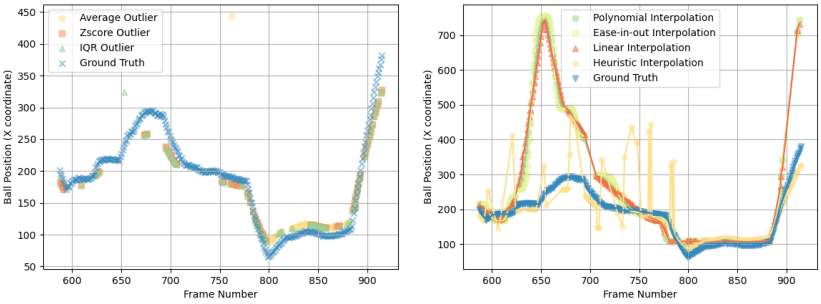
A. Cropping Method and Case Selection

We selected two representative videos to evaluate different cropping methods under varying conditions specific to soccer games. As shown in Table IV, four cases were carefully constructed to scrutinize the performance of different cropping methods, particularly in terms of keeping focus on the main action and objects under different speeds of motion and cropping aspect ratios.

Case	Video #	Event	Crop	Features
1	Video 1	Normal gameplay	1:1	Fast motion, edge field
2	Video 2	Goal	1:1	Varying motion, ball occlusion
3	Video 1	Normal gameplay	9:16	Fast motion, edge field
4	Video 2	Goal	9:16	Varying motion, ball occlusion

TABLE IV: Cases used in the subjective evaluation.

We selected six cropping methods in total, with four alternative versions of SmartCrop and two static benchmarks. The different methods (see Table V) crop an input video of original aspect ratio 16:9 to an aspect ratio of 1:1 (similar to certain features on Instagram post) or 9:16 (similar to TikTok and reels in Instagram). To limit the number of alternatives, we have selected the "best" performing algorithms in the pipeline (e.g., best interpolation method and best outlier detection method) for options $4-6$, based on our objective results described in Section IV. Type 6 corresponds to the full SmartCrop.



(a) Outlier detection methods.

(b) Interpolation methods.

Fig. 5: Outlier detection and interpolation validation against ground truth.

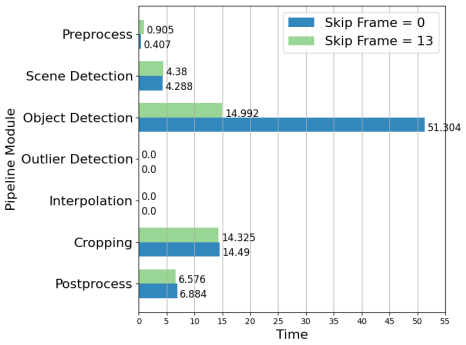


Fig. 6: Runtime per module in local deployment, with and without Skip Frame.

Type	Centering	Description	Outl.	Interp.
1	frame-centered	static no padding	✖	✖
2	frame-centered	static w/black padding to 16:9	✖	✖
3	ball-centered	use last detected ball position	✖	✖
4	ball-centered	w/interpolation	✖	✔
5	ball-centered	w/outlier detection	✔	✖
6	ball-centered	w/interpolation & outlier detection	✔	✔

TABLE V: Cropping types used in the subjective evaluation.

B. Experimental Setup

We used a Google Forms-based online survey to retrieve responses from participants in a crowdsourced fashion. The survey consists of 6 pages: (1) Introduction and pre-questionnaire, (2-5) One questionnaire page per case, and (6) Post-questionnaire. In the Introduction, participants were asked to complete the survey on a mobile phone, as this provides a more realistic setting for our evaluation [23].

Participants first viewed the original video in a 16:9 aspect ratio before evaluating each cropping alternative using a 5-point Absolute Category Rating (ACR) scale, as recommended by ITU-T P.800. Three questions were posed to assess various aspects: overall QoE, the smoothness of the video, and the video's ability to capture the essence of the original content.

C. Participant Details

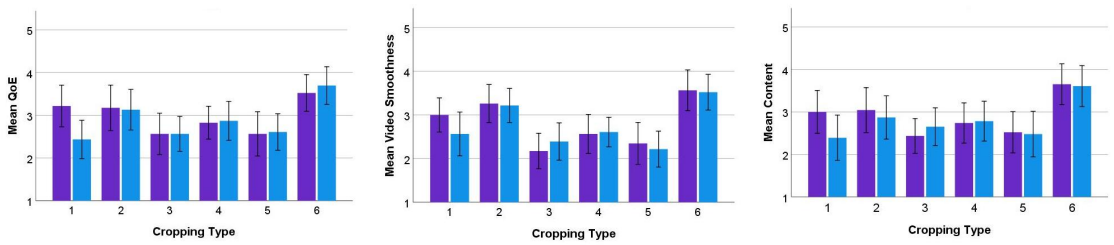
In total, we recruited 35 participants: 11 females, 23 males, and 1 identifying as 'other'. Out of these, 23 participants (9 females and 14 males) met the eligibility criteria. Those who were excluded were using devices other than iPhones, and their data was consequently discarded. The age of the participants ranged from 18 to 63, with a mean age of 30.95 and a standard deviation of 9.96. All participants were active on social media. Regarding daily usage, 26% reported spending 1-2 hours, 30.4% reported 2-4 hours, 17.4% reported 30 minutes to 1 hour, 17.4% reported less than 30 minutes, and 8.6% reported more than 4 hours. On a scale of 1 to 5, participants indicated that they enjoy watching football videos, averaging a score of 3.4. Additionally, they reported actively following soccer content on social media, with an average score of 3.1.

D. Results

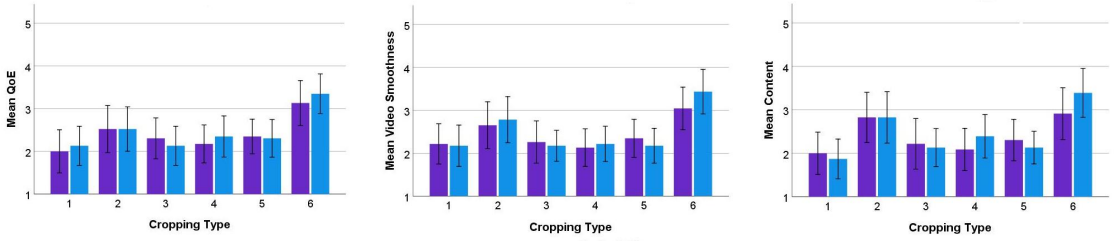
A one-way repeated measures ANOVA was conducted to assess the impact of six different cropping methods, as detailed in Table V, on QoE across four scenarios (two aspect ratios $\times$ two videos). Figures 7a illustrate the Mean Opinion Score (MOS) for QoE with a 1:1 aspect ratio. For Video 1 in this aspect ratio, cropping type had a significant impact on user QoE ($F(5, 110) = 4.73, p < .01$). Post hoc tests using the LSD method revealed that cropping type 6 significantly enhanced QoE compared to cropping types 3, 4, and 5, which shows the importance of outlier detection and interpolation. Notably, there was no significant difference between smart cropping type 6 and the static videos in cropping types 1 and 2. For Video 2 with a 1:1 aspect ratio, the results were similar; it showed that the smart crop does not create a good experience without outlier detection and interpolation (cropping type 3, 4, and 5) in comparison to type 6. Unlike the tests with Video 1, the users preferred the smart crop over the static crop even in a 1:1 aspect ratio for Video 2.

Figure 7d shows the MOS rating of the QoE in the videos with a 9:16 aspect ratio. Unlike the 1:1 videos where static cropping (type 1 and 2) could maintain a reasonable level of QoE, it fell short in the 9:16 format due to the limited cropping window. Our results indicate that the main effect of the cropping type exists on QoE ($F(5, 110) = 4.08, p < .01$). The post hoc LSD test on Video 1, shows cropping type 6 created a significantly higher QoE for the users compared to all the other cropping types than 2, and for Video 2 in 9:16 format, the cropping type 6 was better than all the other cropping types which show that cropping type 6 is the best solution in creating the user QoE.

Regarding the quality aspect of video smoothness, a one-way repeated measures ANOVA confirmed a significant main effect of cropping type in all four scenarios. Figure 7b shows the quality ratings for video smoothness in videos with a 1:1 aspect ratio. In Video 1 ($F(5, 110) = 9.03, p < .001$), the post hoc test shows that without interpolation and outlier detection, the video would lose the smoothness. However, cropping type 6, which features frequent window movement, produced smoothness levels comparable to static videos (types 1 and 2). The results hold true for Video 2 ($F(5, 110) = 9.36, p = .00$), where post hoc LSD tests revealed that cropping type 6 had significantly better video smoothness than non-static types (3, 4, 5).



(a) Question 1: QoE (b) Question 2: smoothness (c) Question 3: content



(d) Question 1: QoE (e) Question 2: smoothness (f) Question 3: content

Fig. 7: Bar plots showing different quality ratings for target aspect ratio 1:1 (a-c) and 9:16 (d-f), with 95% confidence intervals. Purple bars represent cases with Video 1, while blue bars represent Video 2.

In the 9:16 videos, the smoothness ratings for static videos decreased by one MOS point, resulting in inferior quality across all types except for cropping type 6 (Figure 7e). For Video 1 ($F(5, 110) = 3.46, p = .022$) and for Video 2 ($F(5, 110) = 8.45, p = .000$), post hoc tests indicated that cropping type 6 produced significantly higher video smoothness than all other types, both static and non-static.

For the aspect of capturing the essence of the original video (content), a one-way repeated measures ANOVA was utilized to evaluate the performance of the six cropping methods. Figure 7c shows the user ratings for videos with a 1:1 aspect ratio. In videos with a 1:1 aspect ratio, the results indicated that there is a significant main effect of cropping type on capturing the essence of the video for both Video 1 ($F(5, 110) = 6.00, p < .001$) and Video 2 ($F(5, 110) = 5.08, p < .01$). Post hoc tests for video 1 showed that cropping type 6 was significantly better than types 3, 4, and 5. Furthermore, cropping types without interpolation and outlier detection performed badly compared to types 1 and 2. For video 2, type 6 outperformed all the other types except type 2; however, the difference between types 6 and 2 was marginally insignificant (p = .054).

In the 9:16 aspect ratio videos, similar patterns emerged. Figure 7f presents the user ratings. For Video 1 ($F(5, 110) = 3.71, p < .05$), cropping type 6 showed superior performance to all other types, although the difference was marginally insignificant when compared to types 5 (p = .064) and 3 (p = .057). For Video 2, a very significant main effect was observed ($F(5, 110) = 10.55, p < .001$). The post hoc tests revealed that type 6 was significantly better than all other types.

VI. Discussion

This paper presents a framework that adapts soccer videos for mobile displays, incorporating scene detection, object detection, outlier detection, and interpolation, optimized through objective evaluations and subjective studies. SmartCrop significantly enhances user QoE, especially for the 9:16 aspect ratio scenario.

The focus on maintaining the ball in the frame aligns with our post-survey findings, which interestingly revealed that users prioritize content over video quality. The importance of being able to see the soccer ball and players was rated at an average of 4.03 over 5, whereas video quality and smoothness received a lower rating of 3.80. These findings underscore the essential role of content prioritization, validating our focus on maintaining the ball in the frame while also delivering a smooth video playback experience.

There are a number of potential technical enhancements to the SmartCrop pipeline. The user feedback for cropping type 2 suggests that an algorithmic adjustment for **adaptive zoom** could be beneficial. An AI-based approach could be employed to detect important focal points in the video and adjust the zoom level accordingly. This could potentially be integrated into the existing outlier detection and interpolation methods. For cropping type 4, the feedback indicated that the algorithm exhibited unnecessary camera movements. Advanced **motion suppression** algorithms could be incorporated to mitigate jitter or erratic movements, thereby enhancing the QoE. Although cropping type 6 performed well in terms of QoE and video smoothness, users noted limitations in content capture. An extension to the algorithm could involve **semantic segmentation** to prioritize essential game elements like players and the ball, potentially utilizing advanced neural networks. High-resolution videos are increasingly popular, and research could explore how **super-resolution techniques** may be applied to improve the video quality when cropped portions are expanded, especially for wide-angle views. Continuous monitoring and **automated fine-tuning** could be performed

using real-time performance metrics. Reinforcement learning could be employed for performance optimization of cropping algorithms.

Our pipeline evaluation can be expanded by **benchmarking against manual cropping** using metrics such as Mean Opinion Score (MOS). A **broader subjective evaluation** with diverse demographics could further validate our findings. This study, focusing on men's soccer from three Scandinavian leagues, can be adapted to different leagues or **sports** like futsal or beach soccer, considering their unique requirements.

Gender-specific algorithms could be developed for women's soccer by training machine learning models on women's league data. Additionally, incorporating emerging technologies like HuggingFace's Segment Anything (SAM) and algorithmic improvements in segmentation can be tested for performance, speed, and cost-effectiveness, enhancing cropping efficiency and QoE.

Our pipeline could also integrate **different modalities** such as audio, text, and structured data [24], and explore different visual perspectives using novel models such as DINOv2 or SAM, which extract visual information such as depth estimation, matching points, and semantic segmentations. These perspectives can improve cropping accuracy and focus on important frame segments. Similarly, in soccer videos, crucial moments often occur where the main action does not revolve around the ball. For example, off-side situations require focusing on players positioned away from the ball to properly assess the scene.

Currently, our framework primarily focuses on keeping the ball within the frame during cropping, potentially missing out on important actions happening away from the ball, such as off-side situations or crucial player reactions. To address this, we can explore incorporating semantic understanding of soccer game dynamics into our SmartCrop approach. This would involve detecting and understanding soccer-specific events and adjusting the cropping window accordingly. Possible methods include employing advanced object detection algorithms capable of recognizing not just the ball and players, but also other important game elements such as the goalposts, referee, and off-side lines. Alternatively, machine learning models trained on soccer game data could recognize specific game events like goals, fouls, and offsides, adjusting the cropping window to ensure that the most relevant parts of the scene are included in the frame.

VII. Conclusion

We addressed the optimization of media representations across different aspect ratios using the SmartCrop pipeline, which integrates object detection, scene and outlier detection, and interpolation to identify POIs in videos. Our fine-tuned object detection model, trained on 1,500 images, achieved an 82% true positive rate for ball detection, as validated by Confusion Matrix and Precision-Recall curve methods. The IQR and Heuristic Interpolation methods showed the highest accuracy for outlier detection and interpolation, confirmed by RMSE calculations. Furthermore, a subjective evaluation demonstrated significant improvements in user QoE when using SmartCrop (cropping type 6), which incorporates both outlier detection and interpolation. This was especially true in a 9:16 aspect ratio, where smart cropping was the only effective solution. In conclusion, the SmartCrop pipeline offers a robust and user-validated approach for video retargeting across varying conditions.

VIII. Acknowledgements

The authors would like to thank the Norwegian Professional Football League ("Norsk Toppfotball") for making videos available for the research.

References

[1] M. H. Sarkhoosh *et al.*, "Soccer on social media," *arXiv preprint arXiv:2310.12328*, 2023.
[2] O. A. Norgård Rongved *et al.*, "Real-time detection of events in soccer videos using 3d convolutional neural networks," in *Proc. of IEEE ISM*, 2020, pp. 135–144.
[3] J. O. Valand *et al.*, "Automated clipping of soccer events using machine learning," in *Proc. of IEEE ISM*, 2021, pp. 210–214.
[4] A. Husa *et al.*, "Automatic thumbnail selection for soccer videos using machine learning," in *Proc. of ACM MMSys*, 2022, p. 73–85.
[5] S. Gautam *et al.*, "Soccer game summarization using audio commentary, metadata, and captions," in *Proc. of NarSUM*, 2022, p. 13–22.
[6] H. Nam *et al.*, "Jitter-robust video re-targeting with kalman filter and attention saliency fusion network," in *Proc. of IEEE ICIP*, 2020, pp. 858–862.
[7] H.-S. Lee *et al.*, "Smartgrid: Video retargeting with spatiotemporal grid optimization," *IEEE Access*, vol. 7, pp. 127 564–127 579, 2019.
[8] K.-K. Rachavarapu *et al.*, "Watch to edit: Video retargeting using gaze," *Computer Graphics Forum*, vol. 37, pp. 205–215, 2018.
[9] E. Jain *et al.*, "Gaze-driven video re-editing," *ACM TOG*, vol. 34, no. 2, pp. 1–12, 2015.
[10] T. Deselaers *et al.*, "Pan zoom scan – time-coherent trained automatic video cropping," in *Proc. of IEEE CVPR*, 2008, pp. 1–8.
[11] F. Liu *et al.*, "Video retargeting: automating pan and scan," in *Proc. of ACM MM*, 2006, pp. 241–250.
[12] H. Kaur *et al.*, "Video retargeting through spatio-temporal seam carving using kalman filter," *IET Image Processing*, vol. 13, no. 11, pp. 1862–1871, 2019.
[13] S. Wang *et al.*, "Multi-operator video retargeting method based on improved seam carving," in *Proc. of IEEE ITOEC*, 2020, pp. 1609–1614.
[14] Y.-S. Wang *et al.*, "Motion-based video retargeting with optimized crop-and-warp," in *Proc of ACM SIGGRAPH*, 2010, pp. 1–9.
[15] S. Kopf *et al.*, "Algorithms for video retargeting," *Multimedia Tools Appl*, vol. 51, no. 2, pp. 819–861, 2011.
[16] K. Apostolidis *et al.*, "A fast smart-cropping method and dataset for video retargeting," in *Proc. of IEEE ICIP*, 2021, pp. 2618–2622.
[17] W. Liu *et al.*, "Ssd: Single shot multibox detector," in *Proc. of ECCV*, 2016, pp. 21–37.
[18] T. Y. Lin *et al.*, "Focal loss for dense object detection," in *Proc. of IEEE ICCV*, 2017, pp. 2980–2988.
[19] S. Ren *et al.*, "Faster r-cnn: Towards real-time object detection with region proposal networks," *Advances in neural information processing systems*, vol. 28, 2015.
[20] J. Redmon *et al.*, "You only look once: Unified, real-time object detection," in *Proc. of IEEE CVPR*, 2016, pp. 779–788.
[21] T. Soucek *et al.*, "Transnet V2: an effective deep network architecture for fast shot transition detection," *CoRR*, 2020.
[22] Roboflow, "football-players-detection dataset," 2023. [Online]. Available: https://universe.roboflow.com/roboflow-jvuqo/football-players-detection-3zvbc
[23] S. Kemp, "Digital 2020: Global digital overview," 2020. [Online]. Available: https://datareportal.com/reports/digital-2020-global-digital-overview
[24] S. Gautam, "Bridging multimedia modalities: Enhanced multimodal ai understanding and intelligent agents," in *Proc. of International Conference on Multimodal Interaction*, 2023, pp. 695–699.