



Real-time Computational Cinematographic Editing for Broadcasting of Volumetric-captured events: an Application to Ultimate Fighting

Francois Bourel, Xi Wang, Ervin Teng, Valerio Ortenzi, Adam Myhill, Marc
Christie

► To cite this version:

Francois Bourel, Xi Wang, Ervin Teng, Valerio Ortenzi, Adam Myhill, et al.. Real-time Computational Cinematographic Editing for Broadcasting of Volumetric-captured events: an Application to Ultimate Fighting. 16th ACM SIGGRAPH Conference on Motion, Interaction and Games (MIG '23), Nov 2023, Rennes, France. pp.6, 10.1145/3623264.3624468 . hal-04394781

HAL Id: hal-04394781

<https://hal.science/hal-04394781v1>

Submitted on 4 Sep 2025

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.



HAL Authorization

Real-time Computational Cinematographic Editing for Broadcasting of Volumetric-captured events: an Application to Ultimate Fighting

Francois Bourel*
francois.bourel35@gmail.com
Unity
France

Valerio Ortenzi
valerio.ortenzi@unity3d.com
Unity
United States

Xi Wang[†]
xi.wang@inria.fr
LIX, Ecole Polytechnique, IP Paris
France

Adam Myhill
adam.myhill@unity3d.com
Unity
United States

Ervin Teng
ervin@unity3d.com
Unity
United States

Marc Christie
marc.christie@irisa.fr
University Rennes, Inria, CNRS, IRISA
France

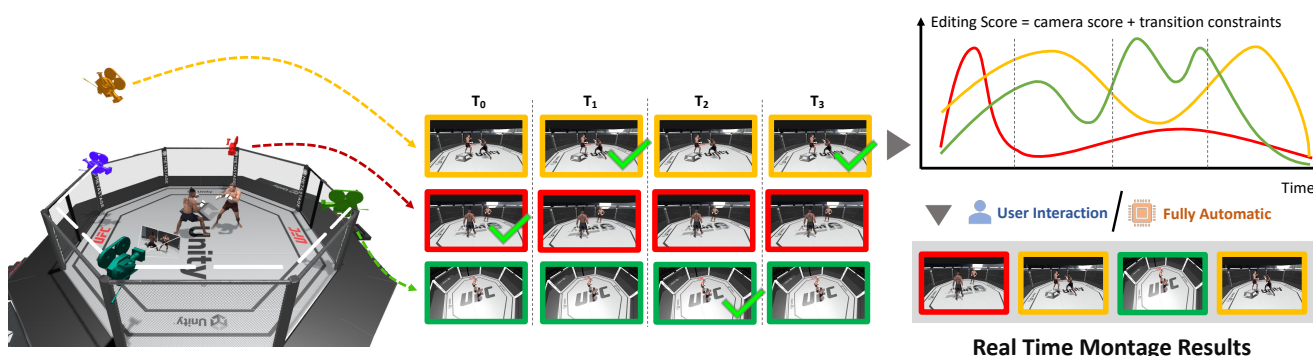


Figure 1: This paper presents a computational virtual cinematography system for volumetric-captured fighting events. Our approach extracts key cinematic features from a given number of virtual cameras, to compose a viewpoint interest score, and proposes a system to interactively or automatically generate edits of the sequence that replicates broadcast shooting style, but also has the capacity to generalise to novel camera angles.

ABSTRACT

The capacity to capture and broadcast sports events with close to real-time volumetric reconstruction techniques opens exciting perspectives in how audiences can consume and interact with these contents. In this work, we propose the design of a real-time cinematography system that is capable of generating qualitative framing and editing of volumetric-captured content, by mimicking real broadcast footage. To illustrate our approach, we focus on the specific problem of cinematography for ring-based events such as Ultimate Fighting Championships (UFC). We start by extracting statistical features from hours of real footage to understand the

specific framing and cutting behaviors of real broadcast directors. We then exploit these features in a real-time editing system, not only to replicate the behaviors in existing broadcasting, but also to generalize to novel camera layouts. We demonstrate our approach on the volumetric reconstruction of UFC fights, compare our results with baseline methods, and report different qualitative and quantitative evaluations.

CCS CONCEPTS

• Computing methodologies → Procedural animation.

KEYWORDS

Virtual Cinematography, Camera Behaviors

ACM Reference Format:

Francois Bourel, Xi Wang, Ervin Teng, Valerio Ortenzi, Adam Myhill, and Marc Christie. 2023. Real-time Computational Cinematographic Editing for Broadcasting of Volumetric-captured events: an Application to Ultimate Fighting. In *ACM SIGGRAPH Conference on Motion, Interaction and Games (MIG '23)*, November 15–17, 2023, Rennes, France. ACM, New York, NY, USA, 8 pages. <https://doi.org/10.1145/3623264.3624468>

*Performed this work while working at Unity.

[†]corresponding author

Permission to make digital or hard copies of all or part of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. Copyrights for components of this work owned by others than the author(s) must be honored. Abstracting with credit is permitted. To copy otherwise, or republish, to post on servers or to redistribute to lists, requires prior specific permission and/or a fee. Request permissions from permissions@acm.org.

MIG '23, November 15–17, 2023, Rennes, France

© 2023 Copyright held by the owner/author(s). Publication rights licensed to ACM.

ACM ISBN 979-8-4007-0393-5/23/11...\$15.00

<https://doi.org/10.1145/3623264.3624468>

1 INTRODUCTION

Thanks to the recent technological advances in volumetric reconstruction techniques of human characters, sports events can now be captured and streamed with very short delays. The availability of such volumetric reconstruction techniques lays the foundation for entirely new ways to consume and interact with sports content. In particular, audiences will now have the power to explore and choose from a variety of camera angles, camera edits, slow motions and replays, while broadcasters will be offered endless possibilities to create novel ways of shooting, cutting and portraying sports events. Yet, the user-tailored conveying of such volumetric contents however requires understanding a number of sport-dependent cinematographic techniques related to how performances are framed by cameras, and how edits are performed in relation to player positions, game states or events occurring. This requires providing the user, or a broadcaster, with a number of high-level means to control and replicate the framing and editing techniques, while still ensuring qualitative framing and cutting behaviors.

Existing approaches that portray sports events essentially focus on real-time footage analysis to provide automated broadcasting systems, achieving the so-called viewpoint selection task. Solutions generally encode hand-designed rules for showing events [Daniyal and Cavallaro 2011; Lefevre et al. 2018; Wang et al. 2008, 2016], rely on learning techniques to estimate the viewpoint [Chen et al. 2019], or preferred event-dependent viewing angle [Wang et al. 2016]. Recently, more complex approaches were exploited, using the three-camera patterns where three different cameras portray the event-begin, event in-progress and event-end [Pan et al. 2021]. Such contributions have a strong focus on event-detection mechanisms, and generally rely on trained view classifiers to select the best viewpoint. As a result, such systems cannot adapt to new camera layouts without re-training, nor control the framing of the cameras. Indeed the composing of shots and editing of sequences together is an intrinsically complex problem: given the same footage, different broadcast directors may produce fairly different edits which would still all be considered good choices depending on the aesthetics they are aiming for. And while some sports-dependent idioms – typical ways of shooting events and editing them – have been developed over years, there are considerable variations in how these can be used and composed, making it difficult to propose a general rule-based framework for generating qualitative sequences or evaluating their quality.

In this paper, we propose a virtual cinematography system to first understand the key features which influence cinematography shooting and editing decisions, and then replicate cinematographic sequences in CG-reconstructed sports events, by relying on volumetric reconstruction technologies. We focus specifically on Ultimate Fighting Championship for which we have access to broadcast data and volumetric 3D reconstruction. The offer of such high-level control from users on the sports cinematography requires addressing the following two problems: 1) extracting sports-specific visual/3D features to understand the underlying key visual properties and their relation with cinematography techniques, and 2) delivering a means to automatically or interactively generate qualitative shootings and edits based on those features and on user inputs.

To comply with these requirements, and by relying on information extracted from the pose of the players, our method automatically shoots the event from a given range of cameras by respecting the UFC camera framing behavior, and generates an edit of these cameras, respecting the UFC broadcast directing behavior. Our method works in real-time and can adapt to any specific camera layout that may not match the positions of the real UFC setup. As a result, it provides novel ways of viewing these sports events without disturbing the viewing experience.

The contributions of this paper are:

- (1) a prior statistical analysis of key visual framing and directing behaviors used in live broadcast of UFC fights;
- (2) a real-time framing and editing algorithm that complies with broadcast framing and directing behaviors, validated through Time Warp Edit Distance (TWED) metrics and subjective evaluations.
- (3) a framework that seamlessly adapts to new camera layouts while ensuring qualitative visual results validated through subjective evaluations.

2 RELATED WORK

The computational editing of real or synthetic footage is a well-known problem at the intersection of multimedia and computer graphics fields. We therefore split the analysis of existing work according to these fields. In multimedia, the keys issues consist in extracting relevant information (e.g. event, motions, states) from a number of existing footage or real-time feeds, and to propose computational editing modalities that would choose what are the best moments to show, from which angles, and with which cuts. Given its great application potential, a number of solutions have been proposed for automated broadcast directing of sport events, mostly football [Wang et al. 2008, 2016] and basketball [Lefevre et al. 2018], which is a camera selection process. Most approaches build on empirical and hand-coded shooting and editing rules that ensure visibility of the actions [Daniyal and Cavallaro 2011; Wang et al. 2008, 2016]. Such approaches require a expert understanding and a manual encoding of rules and weights that generally limits its applicability. Different learning-based approaches have been explored [Chen et al. 2013, 2019, 2018], using techniques such as Random Survival Forest to handle partially missing data. More recently, Pan *et al.* proposed an event-driven directing system [Pan et al. 2021; Yang et al. 2019] which rather focuses on learning the rules from the sport events by identifying explicit sport events rather than proposing a single general viewpoint quality function. We follow a similar approach in our work, identifying two different states of the games and measuring correlations between visual features, and shooting+editing strategies. Approaches have also been proposing more personalized experiences by training viewer preferences to achieve personal viewpoint recommendation systems [Wang et al. 2016] by learning the relationship between personal viewpoints and extracted information from the content (in this specific case, a soccer game). Here we rather propose a system that is able to adapt to any random or specific camera layout crafted by a designer. At last, there is a recent and increased effort to automatically edit filmic [Leake et al. 2017] sequences or learn automated editing techniques from movies [Huang et al. 2021], yet

the problem remains quite specific to dialogue driven scenes and needs to consider a very large range of events and cinematographic conventions.

In computer graphics, the information extraction level is less challenging since most data is available beforehand. Yet the proposed systems are made more complex since not only should they perform computational edits of the sequence, but they have infinite possibilities of placing the cameras. Most contributions address the problem of automated editing of CG filmic contents, using hand encoded rules on viewpoint quality and transitions [Galvane et al. 2015], or more elaborate editing patterns [Wu et al. 2018]. The work we propose both proposes a computational editing system to automatically cut between already placed cameras in a volumetric captured scene, but also adapts to hand-crafted camera layouts.

3 OVERVIEW

This paper aims at analyzing the behavior of cameras in UFC-like fighting broadcasting data, from which we statistically infer the cinematographic characteristics that influence the shooting and editing decisions of broadcast directors, and finally compute real-time automated broadcasts from 3D volumetric UFC data. Our system operates in two stages: First, an offline stage consists of analysing real footage from 2D live broadcasting UFC data, from which we extract a number of visual features and build correlations between these features to help the framing+cutting decisions (see Sec. 4). The system takes as input the raw footage recorded from all the camera angles and the ground truth edit broadcast by UFC, and selects as output the most relevant visual features. The second stage is online (see Sec. 5). Our system takes as input a streamlined volumetric captured UFC performance, and relies on the visual characteristics extracted from the first phase to perform real-time framing and editing. We evaluate the results by showing that the process can easily reproduce qualitative sequences on real footage, as well as on volumetric contents for standard as well as for specific camera layouts (see Sec. 7).

4 STATISTICAL ANALYSIS OF UFC DATA

Our UFC data contains a total of 2 hours of raw data footage, shot from 6 different camera angles, and with the professional director's live editing decisions that were broadcast on mainstream media. In a typical UFC setup, there are 6 camera angles used by the broadcasting team (see the Fig. 2 on the left). Four of the cameras are evenly distributed at the top of four ring poles to cover the fighting area. These cameras have fixed positions but can freely rotate and zoom. The fifth camera is at a fixed position above the ring and also has the possibility to freely rotate and zoom. The sixth camera is placed outside of the ring and shows the opponents from outside. Given that the sixth camera films through the fence of the ring, therefore, leads to detection failure on bodies or faces. We decided to discard camera 6 during our analysis whereas this choice does not impact our work as this camera only represents 4.2% of the broadcast footage.

4.1 Statistical Analysis of Visual Features

In order to understand the rationale driving director's live editing decisions, we propose to perform a number of statistical analyses

on the raw video information from all viewpoints, and from the corresponding professional edited sequences. Our investigation focuses on examining specific video features that we hypothesize may significantly impact framing and editing decisions. By evaluating the degree of correlation between these video features and the framing and editing choices made by directors, we aim to uncover their practical influence in the editing process.

- **Body visibility:** this feature is an obvious cue that drives editing decisions as with many camera viewing/editing systems. In the UFC context, visibility issues occur when a fighter is fully or partially occluded in view (by another fighter or the referee) or when the fighter is partially or fully out of the frame.
- **Body Orientation:** this feature is important in classical editing decisions. Here we propose to understand whether there is a correlation between the body orientation of UFC players and main shooting and editing decisions.
- **Distance between Players:** the on-screen distance between two players is also considered a non-negligible factor in cinematography, and is a key factor of the on-screen relationship between characters.
- **Facial Visibility:** when two fighters are grappling each other, pose estimation techniques often fail, on the only reliable feature here is the facial visibility in which the information is relatively well conserved.
- **Framing:** framing refers to the way characters are positioned in the frame. Here we consider how the joints of the poses of each of the fighters are framed on the screen.
- **Shots:** shot types represent one of the basic lexicons in the cinematographic language system. Directors use different shot types to express semantic and emotional information through the screen and often rely on alternating strategies between shot types to introduce variety.

In the following subsections, we will formalize these features and propose practical means to extract them from audio-visual sequences. We denote the set of all cameras as C , the set of fighters as P , the set of joints for a pose as J , the set of all frames I and the set of frames on which the pose of the fighter p has been detected on the camera c as I_p^c .

For all $c \in C$, $p \in P$, $i \in I_i^c$, $j \in J$: $x_{p,i,j}^c \in \mathbb{R}^3$ represents the position of the joint j of fighter p at frame i of camera c in the 3d pose referential.

4.1.1 Visibility. The visibility of one single fighter p can be modeled by $\omega_p^c : I \rightarrow \{0, 1\}$ s.t.:

$$\omega_p^c(i) = \begin{cases} 1 & \text{if } i \in I_i^c \\ 0 & \text{otherwise} \end{cases}$$

The visibility of both fighters is then defined by taking the average of their visibility:

$$\omega_{OR}^c(i) = 0.5 \cdot (\omega_0^c(j) + \omega_1^c(i)) \quad (1)$$

4.1.2 Body orientation. The body orientation feature quantifies the alignment of the fighters with respect to the camera. We define a unit vector $u \in S^2$, orthogonal to the plane formed by the torso of the player and pointing toward the front of the player's body. We denote ls , rs , lh , lh the joint corresponding to respectively the

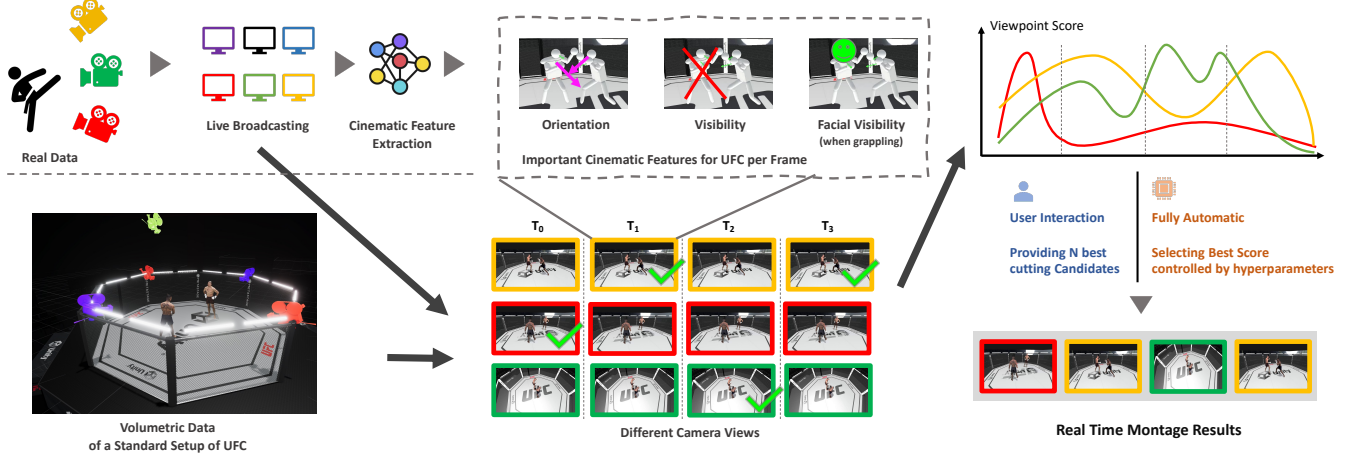


Figure 2: Overview: we propose an algorithm inspired by the statistical analysis of UFC broadcasting data, which can automatically perform well-framed and well-edited sequences from a volumetric captured performance, given an arbitrary camera setup. By relying on prior dedicated feature extraction techniques, our system generates relevant framings of the performance and outputs smoothed per-frame scores for each different camera view angle. The fully automated viewpoint selection process then generates *on-the-fly* editing montage results by selecting novel camera views if their scores are significantly better.

left shoulder, the right shoulder, the left hip and the right hip. We compute u as

$$u_{p,i}^c = 0.5[((x_{p,i,rh}^c - x_{p,i,rs}^c) \wedge (x_{p,i,ls}^c - x_{p,i,rs}^c)) + ((x_{p,i,rs}^c - x_{p,i,ls}^c) \wedge (x_{p,i,lh}^c - x_{p,i,ls}^c))] \quad (2)$$

The body orientation score $\theta_p^c : I \rightarrow [0, 1]$ is then computed using the normalized cosine between $u_{p,i}^c$ and the backward axis of the camera c at frame i .

4.1.3 On-screen distance between fighters. To estimate the on-screen distance between fighters, we calculate the distance between the centre of mass of each of them. The centre of mass of a human body can be roughly estimated as the middle point on the line connecting the hips (and this approximation only loosely impacts the distance estimation). The middle points are respectively computed from the 3D poses of both fighters and projected on the camera plane to measure the on-screen distance.



Figure 3: Character pose estimation results and computation of 2D on-screen distance between fighters.

4.1.4 Framing and Shots. From the camera behaviour aspect, framing refers to the way elements are arranged in the frame. The elements considered are the on-screen pose joints J of each of the

fighters. Usually, the framing used by each camera is constant during a fight. It can be manipulated by different zooming levels of each camera *i.e.* camera shots. Classically, three types of shots are utilised when broadcasting UFC games: full shot, cowboy shot and large shot. The full shot and large shot frame the whole fighters on the screen, while the cowboy shot only frames the characters above the hips. In a typical UFC setup (see Fig. 2), two cameras are using full shots, two for cowboy shots and the last one is the large shot.

4.2 Feature extraction

In practice, we rely on off-the-shelf pose estimation networks (Dope [Weinzaepfel et al. 2020] and LCRNET++ [Rogez et al. 2019]) to obtain 2D poses in the image space and 3D estimated poses in the ring space. When players are close together and they are grappling, there are many visual obstructions and proposed estimation techniques frequently fail in reconstructing a valid pose. Since these situations are challenging to solve in these grappling cases, we propose to rely on the face detector mechanism provided by FaceNet [Schroff et al. 2015] to extract the faces of the fighters. The grappling and non-grappling states of the games were manually annotated in the real footage.

4.3 Analysis Methodology

Given the identification of our different visual features, we now propose to investigate whether they are able to influence the final editing decision *e.g.* if a frame I_i^c at camera c and frame number i is selected by a selection function $s(I)$ in Eq. 3 (we ignore the c and i for the sake of simplicity).

$$s(I) = \begin{cases} 1 & \text{if selected by director} \\ 0 & \text{if abandoned} \end{cases} \quad (3)$$

Table 1: We report the probability of a shot being selected by the broadcast director, conditioned by the maximum of different feature scores in order to demonstrate their statistical significance for editing decisions.

Feature	$p(s(I) = 1 \max(I_c) = 1)$	$p(s(I) = 1 \max(I_c) = 0)$
Orientation	0.467	0.133
Visibility	0.418	0.144
Distance	0.435	0.141
Facial Visibility*	0.415	0.105

We rely on two different methods to analyse our collected features on director-labelled real UFC data: i) the conditional probability s.t. $P(s(I) = 1 | \max(I_c) = 1)$ and $P(s(I) = 1 | \max(I_c) = 0)$ for each feature conditioned on if the investigated score of I_c is the maximum among all cameras at this given frame ($C = 5$ for a UFC setup). This consists in understanding if a feature influences the probability of frame selection; ii) the histogram distribution of each feature between the selected frames $H(\cdot | s(i) = 1)$ and abandoned ones $H(\cdot | s(i) = 0)$.

4.3.1 Pre-processing and moving average. Due to the instability of pose estimation methods and noise influence, a common pre-processing is performed to average the collected features along the temporal direction. For this purpose, we investigated how the size T of this temporal window affected the statistical metrics and their influence on the framing selection.

4.3.2 Conditional Probability. We rely on conditional probabilities to understand whether a high score on a certain feature has an influence on the final decision by computing conditional probability $P(s(I) = 1 | \max(I_c) = 1)$ and $P(s(I) = 1 | \max(I_c) = 0)$ of director-selected frames conditioned on maximum feature score among all camera view-angles given a certain time. The results can be found in Tab. 1 and Fig. 4 (first row) for best found metric and the influence of window size respectively. The term $P(s(I) = 1 | \max(I_c) = 1)$ shows that when the frame holds maximum feature score, the chance of this camera view being selected varies from 0.41 to 0.47 for different metrics showing significance to a random selection probability among 5 cameras equal to 0.2 (also showed as the horizontal black dotted line in Fig. 4). Vice versa, when a frame is not the maximum feature, the $P(s(I) = 1 | \max(I_c) = 0)$ only shows inferior probability around 0.1. The reason why the positive examples $P(s(I) = 1 | \max(I_c) = 1)$ is not very high is due to the **many-solutions** nature of film editing, in other words, with a given time it could exist multiple equally good candidates, therefore eliminates the uniqueness of the solution.

Another observation is that the probabilities tend to be sensitive to window size T . The optimal windows size are labelled in Fig. 4 with a green dotted vertical line. It can be seen that for visibility and orientation features, the optimal window sizes are around 200 whereas for facial visibility used only when grappling, this window size tends to be bigger. Hypothetically we think it might be due to the slower cutting frequency and small motion dynamic of characters when grappling happens.

4.3.3 Histogram distribution. In Fig. 4 second row, histograms show that orientation and visibility seem to be the most *sensitive*

features to the director’s selection due to the difference of the shape of distributions between the director-selected positive (red) and abandoned negative (blue) frames. For distance and facial visibility features, a sizable quantity of frames is cumulated around zero suggesting that the negative data points suffer heavy occlusion owing to the grappling moves in UFC fighting.

4.4 Framing behaviors and shot types

The analysis of sequences from real footage shows that framing behaviors and shot types are constant for a given camera throughout a fighting sequence. We will follow this constraint. The framing behavior of a camera can be described as the way the camera decides to visually place the players on the screen. We measure this framing behavior by reporting the spatial distribution of fighters’ pose joints on the screen. UFC broadcaster relies on three typical types of shots: Full shot, Cowboy shot and Large shot, where the Full shot and Large shot take into account all the joints and Cowboy shot only capture the ones that are above the hips. In Fig. 5 we show the averaged joints distribution (not differentiating the types of joints) of all three shot types respectively. It can be seen that the cowboy and full shots have a clear cinematic pattern of two-shots (i.e. a shot where two characters positioning side by side) as a result of the dual symmetric density kernel, and large shot distribution showing less the pattern due to the grappling view. This information will be exploited later in the algorithmic part to replicate the framing behavior on volumetric contents. Our work also builds implicitly on the assumption that the cameramen in real always film the best way they can, even if their camera is not currently used by the broadcast. Raw footage confirmed this assumption and we considered all the frames of all cameras for the study.

4.5 Shot transition probabilities

Another crucial factor in understanding camera editing is related to the shot transition matrix. They represent constraints and patterns when cutting from one camera to another. These are often encoded by rules [Galvane et al. 2015], learnt transition probabilities [Merabti et al. 2016], or hard-coded transition values [Leake et al. 2017] in Markov chains.

By exploiting the raw and director-edited data of the UFC, we compute the probability transition matrix between the cameras in a classical UFC camera layout (see Tab.2). We observe that cameras 1 and 3 do not cut to each other, and similarly for cameras 2 and 4 which only cut to each other one time in our whole dataset. Weak transition probabilities from any camera to camera 5 are caused by the small number of shots of this camera throughout our dataset (5 is the top-view camera mostly used in grappling). Implicitly these figures, together with the UFC camera layout, encode the classical 180-degree rule (avoid cutting to a camera on the other side of the line composed by the two fighters). The "Change of shot size" rule (that requires a significant change in shot size during a cut) is also implicitly encoded since neighbour cameras are of different shot sizes (cowboy vs full-shot).

5 AUTOMATED EDITING ALGORITHM

By relying on image features analysed in previous sections, we propose a lightweight yet effective *real-time* editing algorithm for

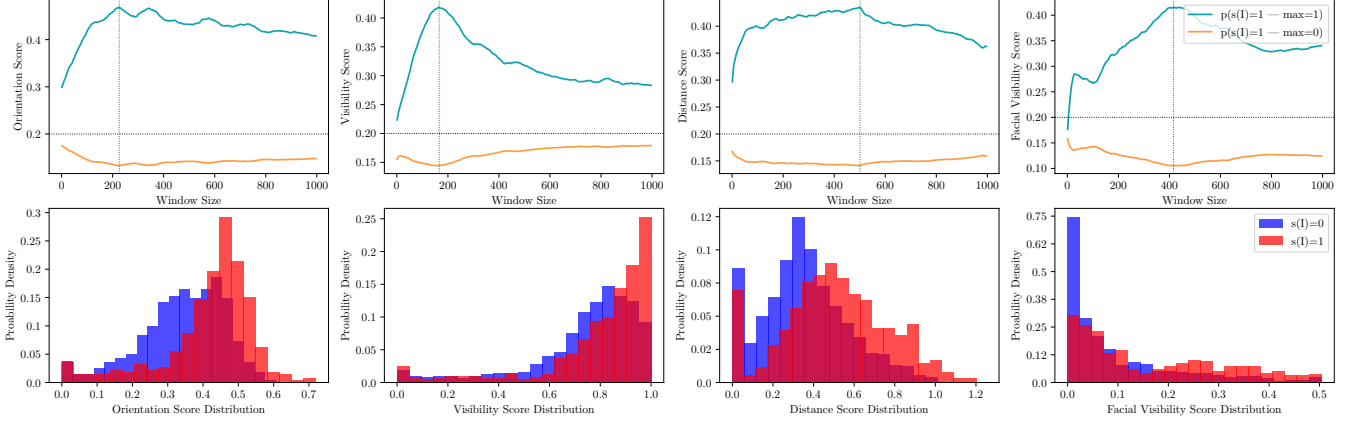


Figure 4: Results of the proposed statistical metrics

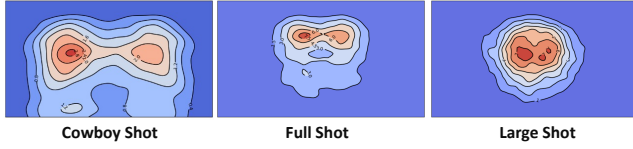


Figure 5: Cowboy / Full / Large Shot on-screen joints density distributions

Table 2: Transition Probabilities between cameras. Full Shot (1 and 3), Cowboy Shot (2 and 4), Large Shot (5).

Camera	1	2	3	4	5
1	-	0.625	0	0.375	0
2	0.292	-	0.59	0.042	0.084
3	0	0.44	-	0.56	0
4	0.8	0	0.2	-	0
5	0	1	0	0	-

both volumetric captured data (*i.e.* one point-cloud per frame) and raw video data of multiple cameras. The algorithm is classically based on i) a camera score computation and ii) on a camera selection technique as with most live editing systems [Chen et al. 2013; Wang et al. 2008].

5.1 Camera Score

We define a viewpoint score ϕ using only previous frames. As shown by the feature analysis, we decide to rely on the visibility ω , body orientation θ and fighter's distance d to compute our score. As such, for all camera c and frame i , we define a frame score ϕ_i^c as:

$$\phi_i^c = 0.25(\omega_0^c(i) \cdot \theta_0^c(i) + \omega_1^c(i) \cdot \theta_1^c(i)) + 0.5(\omega_0^c(i) \cdot \omega_1^c(i) \cdot d_{1,2}^c) \quad (4)$$

A weighted average is then performed using as input the T past frame scores to determine the camera c final score ϕ^c .

$$\phi^c = \frac{\sum_{k=0}^T (T-k) \phi_{i-k}^c}{\sum_{k=0}^T (T-k)} \quad (5)$$

5.2 Camera Selection

The selection of the appropriate camera relies on three aspects: i) the smoothed camera score ϕ^c computed from each view angle; ii) transition constraints between cameras, and iii) a improvement factor in camera score.

The transition constraint is the *180-degree rule* implicitly encoded with stand UFC camera layouts as displayed by zero probability transitions in the shot transition probability table (see 2), and we rely on this table to filter out the impossible camera transitions. However, when the two fighters are spatially close to each other, *e.g.* when grappling, such a spatial line continuity was not maintained in the real data, and therefore not considered. Indeed, grappling fighters can practically be seen as one single subject for editing tasks.

Finally we introduce an improvement factor σ on the score to trigger the cut. Implicitly, this prevents a too fast cutting pace, or avoid cutting between cameras with very similar scores.

6 IMPLEMENTATION ON VOLUMETRIC CONTENTS

To perform human pose estimation on the volumetric data, we rendered the scene from four viewpoints and applied LCRNet++ [Rogez et al. 2019] on the different views and used weighted averaging using the confidence scores provided LCRNET++.

6.1 Broadcast-like real-time shooting and editing

Any virtual camera layout can be used as an input to the system. The designer only needs to define the shot type of each camera (cowboy shot, full shot, or large shot). Framing behaviors are then applied to all the cameras to animate them in real-time and camera selection is then performed in the same way as with real footage.

6.1.1 Achieving camera framing behavior. We draw on the statistical framing information derived from the broadcast to feed the Cinemachine camera composer in Unity¹. More specifically, we rely on the group composer of Cinemachine that automated the framing

¹<https://docs.unity3d.com/Packages/com.unity.cinemachine@2.3/manual/index.html>

of a number of isolated with a given on-screen radius that should be maintained in the screen. We derived different group composers for cowboy shots, full shots and large shots (e.g. using upper body joints of both characters for cowboy shorts, and larger on-screen radius for large shots) to replicate UFC broadcasting data.

6.1.2 Camera Selection. For 3D volumetric data, the visibility feature (including facial visibility) is computed by ray casting towards the center of segments linking joints of the skeleton) and aggregating the results. The orientation feature can be computed directly from the pose skeleton information. The state of the fight (grappling or non-grappling) is then determined using a distance threshold between the two fighters.

7 RESULTS AND EXPERIMENTS

Finding a good metric to compare two different edits remains a complex problem. Most approaches have retained simple metrics for comparison, e.g. measuring the percentage of generated shots that were also selected by the human director within a 1-second time interval [Pan et al. 2021]. Despite the different ways that such accuracy metrics can be computed (e.g. number of similar frames, number of similar shots, duration of shots), these remain insufficient to properly compare two edits. Indeed, the use of very short-timed shorts is not accounted for, as well as small shifts in time between two cuts which obviously impact the resulting edit.

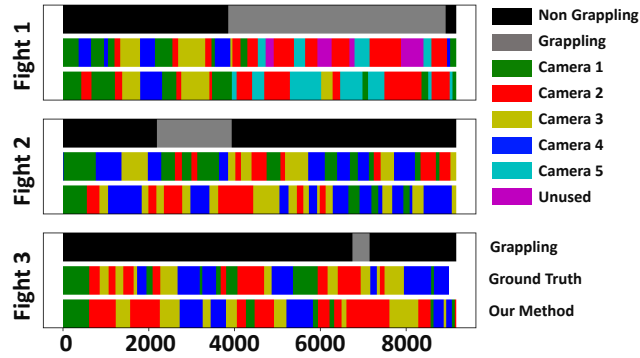


Figure 6: Visualization of the edit results generated by V+O feature combination on three fights, for each fight the grappling information is shown in the first row. The ground truth and our method are presented in the second and the third row respectively with different colors representing different camera index selections at each frame.

More evolved distance metrics such as string comparison edit distances can be considered if we encode the fight edit as a string sequence where each camera is encoded with a different character [Miller et al. 2009]. We propose to rely on the Time Warp Edit Distance (TWED) [Marteau 2008] that accounts well for time elasticity between edits (same cameras at slightly different times).

7.1 Results on the UFC Dataset

The editing results can be compared by using the TWED metric between the ground truth (i.e. a director-selected sequence of cameras) and the ones generated from our proposed system. A random

comparison baseline is also given by exploiting the Poisson distribution for frame duration (i.e. when to cut) where λ is defined by the mean duration of all the UFC shots, and the transition matrix from Sec. 4.5 is used as a Markov random process to select the next cutting candidate (i.e. which camera to cut) based on the current status. As a self-ablation, we test all principal combinations of different features during non-grappling scenarios: *visibility*, *orientation* and *distance* (see Eq. 4) and when grappling happens, we use the face visibility score as the others tend to be heavily occluded. From the experiment results demonstrated in Tab. 3, three fights are tested with a duration of around 9000 frames (5 mins under 30 fps) for each fight. We experimentally found that the combination of visibility and orientation features gives the best performance, despite that the combination of all three features outperformed others in Fight 2. An interpretation of the low contribution of the distance feature to the performance could be due to the fact that the distance should be regarded as a consequence rather than a cause of the edit decision i.e. the main reason for cutting is usually the lost visibility. Here we show the edits produced with the visibility and orientation feature in Fig. 6.

Table 3: TWED Distance between our Generated Edit and GT.

	V ↓	V + O ↓	V + D ↓	V + O + D ↓	baseline ↓
Fight 1	6980.24	5064.66	5816.64	5507.43	9066.7 ± 1569.99
Fight 2	9472.35	9595.7	9702.43	9216.90	11354.64 ± 1349.93
Fight 3	9836.27	7079.53	9112.15	8505.28	11782.8 ± 792.32
Avg.	8762.95	7246.63	8210.41	7742.21	10734.71 ± 1749.28

7.2 User evaluation

To assess the effectiveness of our system, we performed a user study. The group was composed of 30 participants (age range 22-49) with 9 females and 21 males and had normal or corrected-to-normal vision. Two different evaluations were performed (#1, #2) in that specific order. Participants were split into two groups.

The evaluation #1 consisted in pairwise comparisons between different real footage videos presented to the participant. Three sequences were proposed (one grappling and two non-grappling) and for each sequence, three versions of the edit were designed: A-the original broadcast edit (ground truth), B-our computation edit, and C-baseline edit. Side-by-side comparisons were A/C, A/B and B/C. For each video, the participant was asked which sequence portrayed the best players' performance. Table 4 reports the average preference for the participants and shows our results are close to the original edits.

The evaluation #2 also consisted in pairwise comparisons, but with videos from the volumetric performance captured. In a similar way, three sequences were proposed (one grappling and two non-grappling). For each sequence, four versions of the edit were proposed: A-an expert edit (ground truth, which took **20 minutes** to produce), B-our computation edit (generated in **real-time**), C-baseline edit, and D-default framing. Side-by-side comparisons were B/A, B/C and B/D. For each video, the participant was asked three questions: which was the sequence that portrayed the best event, which was the best-edited sequence, and which was the best-framed sequence. Results are displayed in Table 5, and show that our method competes with the expert edit.

Table 4: User evaluation of real UFC broadcasting data among our method, ground truth (i.e. director selected version for broadcasting) and baseline random (see sec. 7.1 for baseline method). Our method only shows around 6% difference to the UFC director cutting while being far better than the baseline.

broadcasting	B - ours	A - gt	C - baseline
	0,372	0,439	0,189

Table 5: User evaluation of volumetric rendering among our method, professional expert edited, random baseline and default results (i.e. default means the camera view is only tracking to the fighter with no cutting presented). Our method's result competes with professional editing, especially on cutting and framing, whereas the professional editor conveys better sports performance. The generated results outperform the default and random results in the volumetric data as well.

	ours vs. exp.		ours vs. random		ours vs. default	
	ours	exp.	ours	random	ours	default
sport perf	0,417	0,583	0,733	0,267	0,817	0,183
cutting	0,467	0,533	0,667	0,333	0,867	0,133
framing	0,467	0,533	0,633	0,367	0,800	0,200

7.3 Discussion and Perspective

Through our experiments and evaluations, we demonstrate the potential of selectively incorporating informative features for computational framing and editing of volumetric sports events, specifically focusing on the Ultimate Fighting Championship in our case study. Our approach in this research leans on statistical methodologies, primarily influenced by the constraints of limited data and the need for real-time processing. Looking ahead, there are several promising avenues to explore. There's potential in amassing larger datasets, utilizing generative models for content creation, and adopting more derivative or implicit approaches for scene representation or camera control [Wang et al. 2023]. Additionally, tailoring the computational processes to match a director's unique style [Jiang et al. 2021] or adapting to different sports genres could further enhance the system's customization level. Recent breakthroughs in decision-making algorithms, particularly the successes achieved by Deep Reinforcement Learning [Kaufmann et al. 2023] in real-time tasks, indicate a promising trajectory, rivaling the capabilities of human operators.

8 CONCLUSION

This paper has tackled the problem of real-time computational framing and editing of volumetric sports events with an application to Ultimate Fighting Championship performances. We first showed how key visual features from real sports footage could be extracted to better understand framing and editing camera behaviors, and composed to create a relevant camera score. We then demonstrated how to exploit this score to re-edit real broadcast footage. And we finally showed how shooting and editing behaviors could be applied to fully synthetic content. Results overall demonstrate the perceived quality of our generated results.

REFERENCES

- Christine Chen, Oliver Wang, Simon Heinze, Peter Carr, Aljoscha Smolic, and Markus Gross. 2013. Computational sports broadcasting: Automated director assistance for live sports. In *2013 IEEE International Conference on Multimedia and Expo (ICME)*. 1–6. <https://doi.org/10.1109/ICME.2013.6607445>
- Jianhui Chen, Keyu Lu, Sijia Tian, and Jim Little. 2019. Learning sports camera selection from internet videos. In *2019 IEEE Winter Conference on Applications of Computer Vision (WACV)*. IEEE, 1682–1691.
- Jianhui Chen, Lili Meng, and James J Little. 2018. Camera selection for broadcasting soccer games. In *2018 IEEE Winter Conference on Applications of Computer Vision (WACV)*. IEEE, 427–435.
- Fahad Daniyal and Andrea Cavallaro. 2011. Multi-camera scheduling for video production. In *2011 Conference for Visual Media Production*. IEEE, 11–20.
- Quentin Galvane, Rémi Ronfard, Christophe Lino, and Marc Christie. 2015. Continuity editing for 3D animation. In *Proceedings of the AAAI Conference on Artificial Intelligence*, Vol. 29.
- Yuzhong Huang, Xue Bai, Oliver Wang, Fabian Caba, and Aseem Agarwala. 2021. Learning Where to Cut from Edited Videos. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*. 3215–3223.
- Hongda Jiang, Marc Christie, Xi Wang, Libin Liu, Bin Wang, and Baoquan Chen. 2021. Camera Keyframing with Style and Control. *ACM Trans. Graph.* 40, 6, Article 209 (dec 2021), 13 pages. <https://doi.org/10.1145/3478513.3480533>
- Elia Kaufmann, Leonard Bauersfeld, Antonio Loquercio, Matthias Müller, Vladlen Koltun, and Davide Scaramuzza. 2023. Champion-level drone racing using deep reinforcement learning. *Nature* 620, 7976 (2023), 982–987.
- Mackenzie Leake, Abe Davis, Anh Truong, and Maneesh Agrawala. 2017. Computational Video Editing for Dialogue-driven Scenes. *ACM Trans. Graph.* 36, 4, Article 130 (July 2017), 14 pages. <https://doi.org/10.1145/3072959.3073653>
- Florent Lefevre, Vincent Bombardier, Patrick Charpentier, Nicolas Krommenacker, and Bertrand Petat. 2018. Automatic camera selection in the context of basketball game. In *International Conference on Image and Signal Processing*. Springer, 72–79.
- Pierre-François Marteau. 2008. Time warp edit distance with stiffness adjustment for time series matching. *IEEE transactions on pattern analysis and machine intelligence* 31, 2 (2008), 306–318.
- Billal Merabti, Marc Christie, and Kadi Bouatouch. 2016. A virtual director using hidden markov models. In *Computer Graphics Forum*, Vol. 35. Wiley Online Library, 51–67.
- Frederic P. Miller, Agnes F. Vandome, and John McBrewhster. 2009. *Levenshtein Distance: Information Theory, Computer Science, String (Computer Science), String Metric, Damerau-Levenshtein Distance, Spell Checker, Hamming Distance*. Alpha Press.
- Yingwei Pan, Yue Chen, Qian Bao, Ning Zhang, Ting Yao, Jingen Liu, and Tao Mei. 2021. Smart Director: An Event-Driven Directing System for Live Broadcasting. *ACM Transactions on Multimedia Computing, Communications, and Applications (TOMM)* 17, 4 (2021), 1–18.
- Gregory Rogez, Philippe Weinzaepfel, and Cordelia Schmid. 2019. Lcr-net++: Multi-person 2d and 3d pose detection in natural images. *IEEE transactions on pattern analysis and machine intelligence* 42, 5 (2019), 1146–1161.
- Florian Schroff, Dmitry Kalenichenko, and James Philbin. 2015. Facenet: A unified embedding for face recognition and clustering. In *Proceedings of the IEEE conference on computer vision and pattern recognition*. 815–823.
- Jinjun Wang, Changsheng Xu, Engsiong Chng, Hanqing Lu, and Qi Tian. 2008. Automatic composition of broadcast sports video. *Multimedia Systems* 14, 4 (2008), 179–193.
- Xi Wang, Robin Courant, Jinglei Shi, Eric Marchand, and Marc Christie. 2023. JAWS: Just a Wild Shot for Cinematic Transfer in Neural Radiance Fields. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. 16933–16942.
- Xueting Wang, Kensho Hara, Yu Enokibori, Takatsugu Hirayama, and Kenji Mase. 2016. Personal multi-view viewpoint recommendation based on trajectory distribution of the viewing target. In *Proceedings of the 24th ACM international conference on Multimedia*. 471–475.
- Philippe Weinzaepfel, Romain Brégier, Hadrien Combaluzier, Vincent Leroy, and Grégory Rogez. 2020. Dope: Distillation of part experts for whole-body 3d pose estimation in the wild. In *European Conference on Computer Vision*. Springer, 380–397.
- Hui-Yin Wu, Francesca Palù, Roberto Ranon, and Marc Christie. 2018. Thinking like a director: Film editing patterns for virtual cinematographic storytelling. *ACM Transactions on Multimedia Computing, Communications, and Applications (TOMM)* 14, 4 (2018), 1–22.
- Danqing Yang, Longfei Zhang, Yufeng Wu, Shugang Li, Dong Liang, and Gangyi Ding. 2019. Computable Framework For Live Sport Broadcast Directing. In *2019 IEEE International Symposium on Multimedia (ISM)*. IEEE, 239–2391.