Extracting Regular FOV Shots from 360 Event Footage

Anh Truong[†], Sara Chen[*], Ersin Yumer[†], David Salesin[±], Wilmot Li[†]

[†]Adobe Research
San Francisco, CA, USA
{truong, yumer, wilmotli}@adobe.com

[*]Durham University
Durham, UK
sara.h.chen@durham.ac.uk

[±]Snap Research
Venice, CA, USA
salesin@snap.com

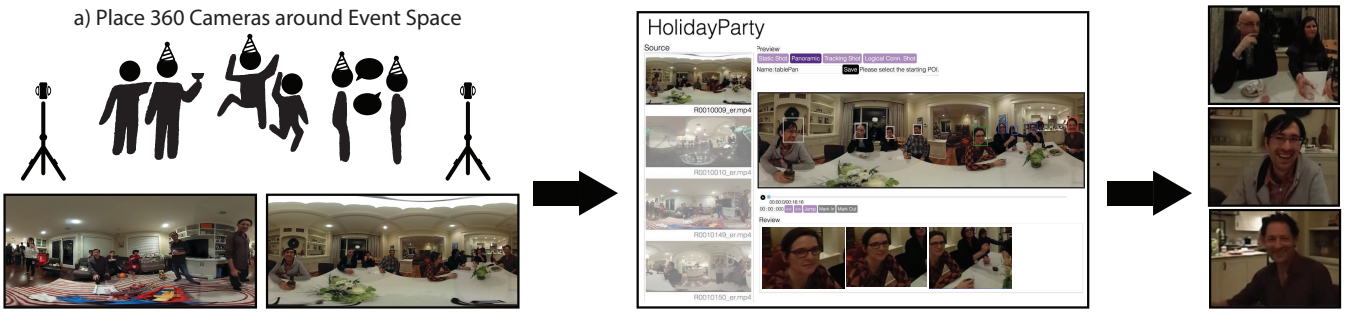


Figure 1. In our approach, party hosts set up stationary 360 video cameras (a) that passively record the entire event (b). We provide a user interface for extracting regular field-of-view (RFOV) shots from the raw footage (c). The user selects moments and points-of-interest that they want to depict, and our system automatically generates good RFOV shots using design guidelines motivated by viewer preferences (d).

ABSTRACT

Video summaries are a popular way to share important events, but creating good summaries is hard. It requires expertise in both capturing and editing footage. While hiring a professional videographer is possible, this is too costly for most casual events. An alternative is to place 360 video cameras around an event space to capture footage passively and then extract regular field-of-view (RFOV) shots for the summary. This paper focuses on the problem of extracting such RFOV shots. Since we cannot actively control the cameras or the scene, it is hard to create "ideal" shots that adhere strictly to traditional cinematography rules. To better understand the tradeoffs, we study human preferences for static and moving camera RFOV shots generated from 360 footage. From the findings, we derive design guidelines. As a secondary contribution, we use these guidelines to develop automatic algorithms that we demonstrate in a prototype user interface for extracting RFOV shots from 360 videos.

ACM Classification Keywords

H.5.2. Information Interfaces and Presentation (e.g. HCI): User Interfaces.

Author Keywords

360 Video; Video Editing; Event Video; Video Summaries.

CHI 2018, April 21–26, 2018, Montreal, QC, Canada

 ISBN 978-1-4503-5620-6/18/04…$15.00

DOI: https://doi.org/10.1145/3173574.3173890

INTRODUCTION

Video has become a pervasive medium for communicating ideas and sharing moments with friends and family. In particular, video summaries of important events (also known as highlight reels) are a popular way to relive or share experiences with others. While event organizers sometimes hire professional videographers to create high quality summaries for formal occasions, like weddings or conferences, this option is too costly and invasive for more casual gatherings like birthday celebrations for kids or house parties with friends. For these informal events, organizers or guests will often create their own highlight reels to post on social media.

However, authoring high quality video summaries without relying on a dedicated videographer is challenging. In many cases, the person responsible for authoring the video is an amateur who does not know how to frame shots appropriately or execute effective camera moves. Even for experienced videographers, capturing an event requires them to be more of an observer than a participant, which is not ideal. Finally, creating a high quality edit that combines the best footage requires time and skill. For these reasons, most "home movies" of casual events on social media are poorly shot and edited, with strange framings, awkward camerawork, long boring shots, and jarring cuts.

The emergence of consumer 360 video cameras offers a potential alternative for capturing casual events. While 360 videos are gaining popularity as a new form of audio-visual media, here we consider the possibility of using 360 cameras to generate regular field-of-view (RFOV) video summaries. The idea is to place 360 cameras around the physical location of the gathering, which makes it possible to capture large portions of the event space without actively filming the action. Of course,

as noted above, capturing is only part of the task. In order to create effective video summaries from the 360 video, we need to extract usable shots from the large collection of raw footage and then edit these shots together.

Putting this approach into practice requires addressing a few key challenges. Setting up the right configuration of cameras that covers the relevant action at the event in a non-invasive way is one task. Another challenge is how to extract good RFOV shots from the large amount of raw footage that gets capture in this scenario. There are several reasons why the extraction problem is non-trivial. First, since the idea is to capture the event passively, there is no way to "stage" a given shot (e.g., by rearranging the subjects in the scene). This is in contrast to narrative film, or even situations where a videographer asks people to reposition themselves for a better shot. In addition, since the 360 cameras themselves are stationary, it is not possible to actively capture certain types of shots (e.g., an over the shoulder shot, or anything that requires a dolly). Finally, even though 360 cameras provide good spatial coverage, the nature of the capture setup can produce a lot of unusable footage. For example, people or objects can inadvertently block the camera, or a subject of interest may be standing too close to the camera to produce a distortion-free shot or too far away to create a high resolution shot.

In this work, we focus on the problem of extracting usable RFOV shots from 360 event footage. One potential approach is to use cinematography rules and best practices as a guide to find good footage. However, many of these rules are not prescriptive. For example, Arijon [2] recommends framing subjects with space in the direction of their gaze, but does not prescribe exactly where to place them. A subject looking towards the right of the frame can be placed anywhere from the center to the left edge of the frame while still honoring this rule. Moreover, for the reasons outlined above, the footage may not contain the type of "ideal" shots that filmmakers try to produce. For example, the rule of thirds may suggest that we place the head of a subject in a specific spot within the frame, but doing so may cut off the face of another person on the other side of the frame. For a moving camera shot, the ideal technique may be to execute a smooth dolly that follows the subject into a room, but we cannot move the center of projection of the cameras.

Despite these limitations, it is often possible to extract good (if not ideal) shots from 360 event footage by strategically relaxing some of the cinematography rules and trading off various desirable properties, like maintaining good compositional balance, keeping subjects entirely within the frame, and minimizing distracting background content. Our paper explores how to make such tradeoffs. We describe studies that examine human preferences for two common types of shots in event videos: 1) static camera shots that focus mainly on one subject and 2) moving camera shots that show a larger scene, follow a specific action, or convey a logical connection between two people or actions. From these studies, we distill guidelines for how to generate rectilinear RFOV shots from equirectangular 360 event videos.

Finally, as a secondary contribution, we present preliminary algorithms to operationalize some of these guidelines and demonstrate their usage via a prototype interface to help users extract usable shots from raw 360 footage.

BACKGROUND

Virtual Videography. Several researchers have developed techniques to automate capture and editing of RFOV videos by leveraging cinematography rules. These video creation methods cover a variety of settings, including virtual 3D environments [11], lectures [3, 12], sports [5], theatrical performances [9], and events captured with first-person cameras [1]. Some of these efforts involve multi-camera capture setups that are similar to ours. In particular, Heck et al. [12] automatically cuts together lecture footage from multiple stationary cameras. In contrast to all of these previous methods, we propose an approach that uses 360 (rather than RFOV) footage as input. Moreover, we focus specifically on capturing social events. Our work demonstrates that 360 event footage has unique characteristics that make it difficult to directly follow traditional cinematography rules.

Editing 360 Video. Several recent efforts propose methods for converting 360 to RFOV video: Su and Grauman [20] moves a RFOV frame to the most salient regions in 360 video; Lai et al. [15] proposes a similar saliency-driven approach to create RFOV hyperlapses that summarize 360 content; Hu et al. [13] uses deep learning to automatically extract an RFOV crop from 360 sport videos; and Pavel et al. [18] proposes saliency based approaches for aligning viewpoints in two 360 videos played back to back in a headset. These methods focus on generating one long moving camera video by utilizing saliency features (whether learned or designed) to find an optimal RFOV crop at each frame of an input 360 video. On the other hand, our goal is to generate short shots that can be compiled into an event highlight reel. In our setting, aesthetic quality rather than saliency is the most important factor, and our approach is to derive design guidelines for extracting pleasing RFOV shots.

Video Retargetting. The early work of Liu and Gleicher [17] uses image saliency and optical flow to zoom and pan RFOV videos for better viewability on smaller displays. Deselaers et al. [6] proposes a similar approach to convert 16:9 videos to 4:3 aspect ratio, and vice versa. More recent work by Jain et al. [14] introduces a gaze-driven method that retargets RFOV videos to different aspect ratio displays. While these methods retarget one RFOV video to another, our method focuses on cropping RFOV regions from 360 videos with the aim of conveying key moments and points-of-interest rather than preserving all the salient content.

Automatic Photo Cropping. Creating good RFOV crops from 360 video is related to the problem of automatic photo cropping. Recent methods by Zhang et al. [23, 22] and Fang et al. [7] use machine learning to detect salient regions, encode visual composition rules, and determine where to place crop boundaries. However, unlike photo cropping, composing good shots in video requires additional consideration for the movement within the frame, as well as the potential movement of the camera.

DATASETS

To investigate the specific challenges in extracting RFOV shots from 360 event footage, we collected four different datasets at social events that we hosted or attended. To capture the events, we placed two to four 360 video cameras around the event space. We filmed two events with the Ricoh Theta S camera at 1080p resolution and two with the Kodak PixPro at 4K resolution. Following the guidelines of Arijon [2], we positioned cameras at the height of the average attendee's head to capture "natural" shots of the action. We also tried to place the cameras at areas of interest where we predicted people would gravitate (e.g., at the centers of tables, close to the doors, in view of food, etc.). While one or two authors did attend each of the events, we did not actively monitor what was happening around the cameras during the gatherings or provide any specific guidance to other guests on how to interact with the cameras (e.g., we did not dissuade people from blocking cameras either intentionally or accidentally).

Two datasets are from separate holiday parties, each of which had roughly 20 guests scattered across 3-4 rooms in both seated and standing positions. For these events, we placed Theta cameras (4 for one party, 3 for the other) in different rooms. We also captured a poster presentation session with about fifty standing attendees gathered in a single room and a recurring set of office tea breaks where about fifteen people mingled around a dessert table. We used two PixPro cameras to capture each of these events.

The following sections describe how we used these datasets to investigate and test different strategies for extracting various types of RFOV shots from the raw event footage.

STUDY 1: STATIC CAMERA

Social events typically involve many interactions between people. As a result, a large portion of the content in most video summaries focuses on such interactions. To capture these moments, videographers typically use static camera shots that focus on a primary subject who may be talking, listening or otherwise engaging with others at the event. For such shots, the main design decision is how to frame the main subject.

Cinematic guidelines propose some rules, but as mentioned in the intro, such rules are not always useful for 360 footage. The specific arrangement of people and objects relative to the camera often makes it difficult to achieve "ideal" framings of a given subject without including distractors (e.g., objects, other people) or cutting people off with the edge of the frame.

Setup

To better understand the impact of these tradeoffs on framing choices, we conducted a study that compares human preferences for different framings. From our datasets, we identified 36 "scenes" that have a variety of characteristics:

dialogue. whether the subject is speaking (d_T) or not (d_F)

movement. whether the subject is moving (m_T) or not (m_F)

gaze. whether the subject is looking to the right (g_R), center (g_C), or left (g_L) of the frame; shown in the top row of Figure 3

Code	IRR
Preference for centered content.	.7961
Preference for content to be offset to the side.	.9202
Position of subject in the frame w.r.t. gaze direction.	.9156
Artistic intention.	1
Prefer less busy backgrounds.	.7973
Distracting content at the edge of the frame.	.8765
How much the main subject is cropped out the frame.	.9399
Prefer more context in addition to the subject.	.8908
Personal preference.	.8233

Table 1. Codes from the freeform responses of the pilot static camera study with Cohen's Kappa inter-rater reliability scores.

crop. whether the crop around the subject is close (c_T), medium close (c_M), or medium (c_W); shown in the bottom row of Figure 3

We found these characteristics to be emergent in the professional highlight reels we viewed and the events we captured. For each scene, we generated 5 different static camera shots that position the center of the main subject's face at different locations in the frame: far right (f_{R2}), right third (f_{R1}), centered (f_C), left third (f_{L1}), far left (f_{L2}). We vertically positioned the subject's head, with a small margin of padding, at the top of the frame.

We then compared all 10 pairs of framings for each scene. For each pair, we produced a side-by-side comparison task and used Amazon Mechanical Turk (AMT) to obtain human preference judgements for each task. 180 workers did 20 comparisons each, giving us a total of 10 judgements per pair.

To better understand the rationale behind each judgement, we asked workers to explain the reason for their preference. We identified the set of potential reasons by running a pilot study with 43 people and asking them to explain each judgement with a freeform response. We used an affinity diagram to identify themes in the responses and converged on 9 codes. Two authors then coded the data independently. We report the codes and their inter-rater reliability in Table 1. In the AMT study, we asked workers to choose from the 9 categories, with an option of "Other" to specify a freeform response.

Findings

At a high level, we summarize the overall framing preferences by fitting a Bradley-Terry (B-T) model to our data and using the ability scores to define a ranking for the five different framings. As shown on the right, the central framings were generally preferred over the extreme framings, with f_C the most preferred. Not surprisingly, the most frequently cited reason for choosing f_C over any other framing was "Prefer centered content" (51%).

Framing	Ability
f_C	1.418
f_{R1}	1.059
f_{L1}	0.806
f_{R2}	0.000
f_{L2}	-0.485

While the aggregated preferences are informative, we also wanted to identify situations or scene types where the preferences deviate from the overall trends. In deciding how to analyze the data, we note that the half of our scenes where the subject moves (m_T) is fundamentally different from the other half (m_F). Since we assign the per-

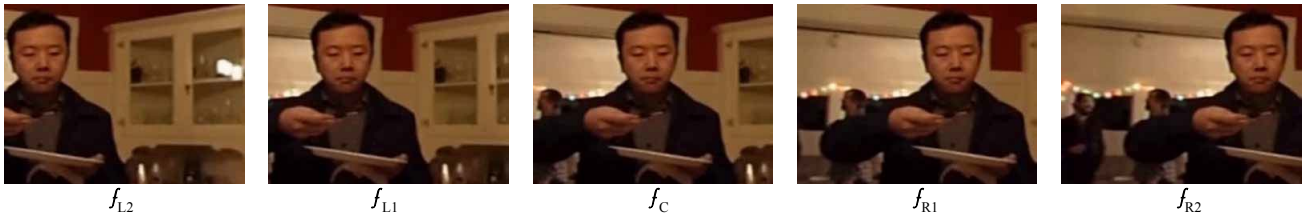


Figure 2. Examples of the five framings used in our studies.

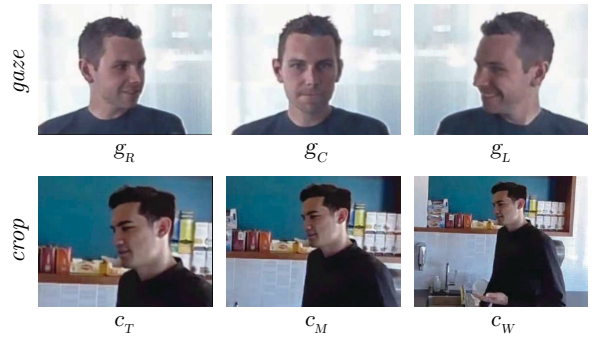


Figure 3. Typical variations for gaze direction (above) and crop (below).

scene gaze value and generate each variation based only on the first frame of the shot, these characteristics are much more consistent and meaningful for the m_F scenes versus the m_T scenes where the subject may move across the frame or change their gaze direction. As a result, we analyze the m_F and m_T scenes independently using different analysis methodologies.

No Movement

To analyze the m_F scenes, we investigate the potential impact of the other scene characteristics on framing preferences. For each target characteristic, we first enumerate subsets of scenes defined by fixing all possible combinations of the non-target characteristics. Then, we split each of these subsets on the target characteristic and analyze whether the framing preferences differ across these splits. For example, if gaze is the target characteristic, one of the subsets would be (m_F, d_T, c_T) and the splits would be (m_F, d_T, c_T, g_R), (m_F, d_T, c_T, g_C), (m_F, d_T, c_T, g_L). Note that m_F persists across all subsets because we are only considering the scenes with no subject movement.

For each set of splits, we look at every pairwise comparison between framings (e.g., f_{R2} vs. f_{R1}) and perform a Chi Squared test to determine whether the preferences vary significantly across the split. Below, we focus on the splits where we found significant effects for comparisons that involve at least one preferred framing.

Viewer Preferences across Crop Levels

We found several notable differences when examining how preferences vary with crop level. If we look at the frequency of the reasons chosen for all preferences across different crop levels, we see that "How much of the main subject is cropped out of the frame" increases from 23% in the widest level (c_W) to 35% in the closest crops (c_T). Conversely, the frequency of "Prefer less busy backgrounds" drops from 21% in c_W to 9% in c_T. These trends align with our expectations that subject cropping is more likely to be a problem in closer crops, while distracting backgrounds are more problematic in wider shots.

Looking at the preferences, we observe that the cinematographic rule to frame subjects with more empty space in the direction of their gaze becomes less aligned with viewer preferences as the crop becomes tighter. In scenes where the subject gazes right (g_R), the most preferred framing changes from f_{L1} to f_{R1} as we move from c_W to c_T(Table 2). In other words, in c_W, the preferred framing aligns with the gaze direction rule because the subject gazes right and is framed on the left. However, in c_T, the preferred framing puts the subject on the right even though they are looking in that direction. There is a similar trend in the preferences for gaze left (g_L) scenes where the preferred framing changes from f_C to f_{L1} as we move from c_W to c_T. For both the g_R and g_L scenes, the effect of crop level on the comparisons between the most preferred framings is statistically significant, as shown in Table 3. For the g_C scenes, f_C is the preferred framing for all crop levels.

c_T		c_M		c_W	
Framing	Ability	Framing	Ability	Framing	Ability
f_{R1}	1.275	f_C	2.197	f_{L1}	0.531
f_C	1.030	f_{R1}	1.867	f_C	0.501
f_{L1}	0.221	f_{L1}	0.693	f_{R2}	0.000
f_{R2}	0.000	f_{R2}	0.000	f_{R1}	-0.359
f_{L2}	-2.263	f_{L2}	-0.606	f_{L2}	-0.688

Table 2. B-T rankings for framings by crop level. g_R is fixed.

Comp. Pair	Results Per Variation		
	c_T	c_M	c_W
f_{R1}, f_C*	f_{R1}(.65)	f_C(.56)	f_C(.83)
f_{R1}, f_{L1}*	f_{R1}(.68)	f_{R1}(.78)	f_C(.76)
f_C, f_{L1}**	f_C(.79)	f_C(.89)	f_C(.5)

Table 3. Pairwise comparisons that differ significantly across crop levels. We report the preferred framing in each cell with the corresponding fraction of votes. g_R is fixed.

Divergence from Centered Framing for Close Crops

We also see significant effects for gaze direction on viewer preferences. In particular, as noted in the analysis of crop levels above, the preferred framings across different gaze directions for close crops appear to break the gaze direction framing rule, with f_{R1} preferred for g_R and f_{L1} preferred for g_L. Looking at the effect of gaze direction in c_T scenes, we find that the differences between these preferred framings are statistically significant, as shown in Table 5.

When examining the reasons for these preferences, we were surprised to find that "Prefer centered content" was chosen most frequently even in comparisons where f_{R1} or f_{L1} were preferred over f_C, as shown in Table 4. To understand these

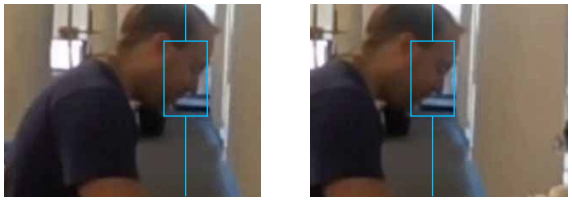
g_L		g_C		g_R	
Framing	Ability	Framing	Ability	Framing	Ability
f_{L1}	2.737	f_C	1.853	f_{R1}	1.275
f_C	1.754	f_{R1}	1.026	f_C	1.030
f_{R1}	1.268	f_{L1}	0.908	f_{L1}	0.221
f_{L2}	0.284	f_{R2}	0.000	f_{R2}	0.000
f_{R2}	0.000	f_{L2}	-0.706	f_{L2}	-2.263

Table 4. B-T rankings by gaze direction. c_T is fixed.

Comp. Pair	Results Per Variation		
	g_L	g_C	g_R
f_{R2} , f_{L1}*	f_{L1} (.94)	f_{L1} (.53)	f_{L1} (.61)
f_{R1}, f_C*	f_C(.81)	f_{R1}(.65)	f_C(.67)
f_C, f_{L1}**	f_{L1}(.88)	f_C(.79)	f_C(.73)

Table 5. Pairwise comparisons that differ significantly across gaze subsets. c_T is fixed.

findings, we manually inspected the videos in each of the g_R, g_C and g_L subsets and found that for many shots, the subjects were visually more centered in f_{R1} or f_{L1} compared to f_C. As described previously, f_C shots are generated by centering the subject's face in the frame without taking the rest of the body into account. However, as shown in Figure 4, when the subject

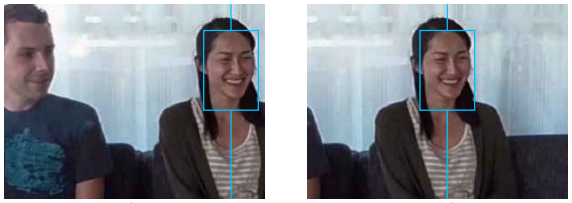


f_{R1} f_C

Figure 4. f_{R1} (left) and f_{L1} (right) framings for a c_T scene.

looks left or right, their shoulders are often positioned on the opposite side of the face with respect to the gaze. Thus, in c_T, where only the head and shoulders are visible, the overall position of the subject may appear off center even if the face is centered in the frame. In these situations, shifting the face towards the gaze direction often makes the subject appear more centered. In other words, for c_T, it may be preferable to center subjects using their entire upper body (head and shoulders) rather than just their face. Similar observations are not present for c_M and c_W, where the f_C was preferred either equally to or more than all other framings across all gaze conditions.

Exception for Extreme Framings



f_{R2} f_C

Figure 5. f_{R2}(left) and f_C framings for a g_L, d_T, c_W scene.

Overall, more central framings (f_{R1}, f_C, f_{L1}) were almost always preferred over the extreme framings (f_{R2}, f_{L2}) in our dataset. This aligns with the general preference for centered content. However, there was one m_F scene in which f_{R2} was preferred over the other framings. The scene had the subject looking to the right (g_L), speaking (d_T), and a wide crop (c_W). Given the gaze direction, the preference for f_{R2} is surprising because it strongly violates the gaze direction rule. Looking more closely at this scene, we see that f_{R2} not only shows the main subject but also includes another seated person on the left side of the frame. All the other framings crop out this second person either partially or completely, as shown in Figure 5. Viewers seemed to prefer having the second person in the shot to add context; "Prefer more context in addition to the subject in the frame" was specified as a contributing reason in at least 62% of all comparisons where f_{R2} was preferred over another framing for this scene. While this preference for an extreme framing only arose in this one scene, the data suggests a plausible motivation for using extreme framings when it allows relevant context to fit in the frame.

Movement

As noted above, the gaze direction and framing labels for the movement (m_T) scenes are less reliable than for the m_F scenes due to the motion of the subject. Moreover, across different scenes, subjects may be moving different distances in different directions, for different amounts of time. These variations make it difficult to compare trends in framing preferences across different scenes. Thus, we analyze the framing comparisons for each scene separately and identify all comparison pairs that exhibit a statistically significant preference using a Chi-Square Goodness of Fit test. Finally, we aggregate the contributing reasons for the more preferred framing across all statistically significant comparison pairs.

The most popular reasons in descending order were "Prefer centered content" (55%), "How much of the main subject is cropped out of the frame" (29%), "Position subject in the frame with respect to their gaze direction" (6%), "Prefer more context in addition to the subject in the frame" (6%), "Distracting content at the edge of the frame" (4%), and "Prefer less busy backgrounds" (4%). This distribution of reasons is generally consistent with the preferences from the m_F scenes.

STUDY 2: MOVING CAMERA

Experienced videographers employ moving camera shots to guide the viewer's attention in a specific way. Three specific categories of moving camera shots, described by Arijon [2], that are prevalent in highlight reels are

panoramic shots where the frame moves to reveal additional context in the scene (e.g., panning across a dance floor at a party or around a dinner table),

tracking shots where the frame follows a subject in motion (e.g., following someone walking across a room),

logical connection shots where the frame moves from one subject to another to convey a relationship (e.g., moving from a fireplace to somebody warming their hands nearby).

Videographers use a variety of camera moves (dolly, pedestal, pan, zoom, etc.) to execute such shots, but as mentioned earlier, our approach of using static cameras limits us to pans and zooms. Thus, in this study, we consider how various panning and zooming strategies affect viewer preferences when creating panoramic, tracking and logical connection shots.

Setup

Code	IRR
Camera distortion.	.9717
Smoothness of camera path.	.8041
Smoothness of zooming.	.8102
Distracting background content.	8290
Speed of movement.	.6498
Better framing of main subject.	.8307
Prefer more context of the scene.	.8
Personal preference.	.8343

Table 6. Codes from the freeform responses of the pilot moving camera study with Cohen's Kappa inter-rater reliability scores.

As with Study 1, we selected a set of scenes from our captured datasets for generating various moving camera shots. We chose scenes that are suitable for the three different categories of shots described above. These scenes exhibit many of the common challenges that arise in 360 event footage. For panoramic shots, we selected three scenes (P_1, $P_2$and P_3) where party attendees were distributed in various configurations around the 360 camera, shown in Figure 6. Framing these moving camera shots is difficult because the points of interest (POI), in this case people, are sitting or standing at different heights or distances away from the camera. For tracking shots, we selected two scenes ($T_1$and T_2) in which the POI, a person, walks across a room, either moving towards or away from the camera. However, these walking actions are not consistent or smooth in space for time. Finally, for logical connection shots, we selected two scenes ($L_1$and L_2), each with two POIs such that one is a human subject and the other is some inanimate object that the subject is about to interact with. However, as with the panoramic scenes, the POIs are distributed non-uniformly in front of the 360 camera, such that they require different FOVs to capture.

For each shot category, we generated three different moving camera shots by varying the panning trajectory and FOV (which effectively varies the amount of zoom). For panoramas, we created 1) a *Constant Wide FOV, Smooth Path (CS)* shot that moves in a straight line from the first to last POI with a fixed FOV wide enough to frame all intermediate POIs; 2) a *Constant Tight FOV, Varying Path (CV)* shot that uses a tighter fixed FOV and deviates from the straight line path to frame all intermediate POIs; and 3) a *Varying FOV, Smooth Path (VS)* shot that moves in a straight line but varies the FOV to frame the intermediate POIs more tightly than the *CS* shot. For tracking shots, we created analogous shot variations where the first and last POIs are defined by starting and ending position of the main subject and the intermediate POIs are defined by the motion of the subject during the shot. Finally, for logical connection shots, there is no reason to vary the path since there are no intermediate POIs. Thus, we generate three different smooth path shots: 1) a *Constant Start FOV, Smooth Path (C_sS)* shot with a fixed FOV based on the first POI; 2) a *Constant End FOV, Smooth Path (C_eS)* shot with a fixed FOV based on the second POI; and 3) a *Varying FOV, Smooth Path (V_S)* shot that varies the FOV from the first to last POI.

We adopt the same approach as Study 1 to compare shot variations. For every pair of shot variations within each scene, we use AMT to gather human preferences. Since Study 2 involves far fewer scenes in total, each worker did all 21 comparisons (3 pairs for each of the 7 scenes). We recruited 30 workers, giving us 30 judgements per pair. Similar to Study 1, we ran a pilot study with 40 participants to obtain freeform explanations of their preferences and coded the responses to obtain the 8 categories listed in Table 6. In the AMT study, we asked workers to choose a reason for each preference from these categories, with an "Other" option for freeform responses.

Findings

For each shot category, we aggregate the data across all scenes in that category and fit a Bradley-Terry model to derive an overall ranking. We also analyze the variation comparisons to identify all comparison pairs that exhibit a statistically significant preference using a Chi-Square Goodness of Fit test. We further apply this analysis separately to each scene within a category. Here, we focus on aggregated data and the comparison sets for which there are significant findings that differ from the trends in the aggregated findings.

Panoramic Shot

P_1 P_2

Figure 6. We show P_1 and P_2 for the panoramic shot type. In these scenes, there are multiple POIs who are distributed non-uniformly throughout.

Var.	Ability
CS	0.000
CV	-0.095
VS	-0.949

	CS	*VS*	*CV*
CS	-	.7222**	.5222
VS	.2778**	-	.3000**
CV	.4778	.7000**	-

Table 7. B-T rankings (left) and preference matrix (right) for the aggregated panoramic shot data. Each matrix entry gives the percentage of preferences for the variation in the corresponding row over the variation in the column. Entries marked * are significant at p $<$ .05 and ** are highly significant at p $<$ .001.

As shown in Table 7, the constant FOV pans, *CS* and *CV*, were generally preferred over the varying FOV pans, *VS*, which suggests that it may be preferable to keep the FOV fixed during pans. While the general trends show no significant preferences between *CS* and *CV*, the isolated preferences for two of the panoramic scenes, P_1 and P_2 suggest that there may be a tradeoff between the two techniques.

In P_1, *CS* is significantly preferred over *CV* while the opposite is true in P_2. Looking more closely at these scenes, the spatial distribution of POIs relative to the camera seems to be an important factor in what type of constant FOV shot is preferable. In P_1 (Figure 6, left), we observe the POIs standing at very different distances away from the physical camera, such that the subject on the left is very close to the camera and the fourth subject from the left is on the far side of the room. In general, subjects standing at very different distances from the physical camera require very different FOVs to be framed tightly by the virtual camera, which makes it hard to find a single tight FOV that works well for the entire shot. Moreover, a tight crop will likely result in a very uneven camera trajectory. Thus, it may

be preferable to select a constant wider framing for such shots. Not surprisingly, in this scene, the most cited reasons for favoring *CS* over *CV* are "Better framing of the main subject"(60%) and "Smoothness of the camera motion"(52%).

On the other hand, in P_2, we find the opposite preference for *CV* over *CS*. The most frequently cited reason for the preference is "Amount of stretching or distortion in the frame" (57%). This distortion occurs as a consequence of the rectilinear projection that is commonly applied to extract RFOV crops from 360 video. A key property of this projection is that straight lines in 3D space will appear as straight lines in the 2D image. However, this property also has the effect of causing objects to appear stretched as they near the edge of the frame in wider FOV shots. In P_2 (Figure 6, right), the second subject from the left is standing while the woman next to her is sitting. To capture both of them with a consistent FOV and smooth trajectory requires a very wide angle, which introduces distortion into the scene. In such instances, it may be preferable to tighten the FOV.

Tracking Shot

Var.	Ability
CV	0.338
CS	0.000
VS	-0.135

	CV	*VS*	*CS*
CV	-	.6167	.5833
VS	.3833	-	.4667
CS	.4167	.5333	-

Table 8. B-T rankings (left) and preference matrix (right) for the aggregated tracking shot data.

As shown in Table 8, we found no significant general trends for the preference of one tracking shot variation over another. However, the data for individual scenes suggests similar tradeoffs as the panoramic shots. In particular, we see the same aversion to distortion in T_1, where there was a significant preference for the tighter FOV shot, *CV*, over the wider angle, *CS*, with "Amount of stretching or distortion in the frame" one of the most frequently cited reasons. However, in T_2, where framing the motion of the subject does not require an overly wide, distortion-inducing shot, viewers significantly prefer the wider FOV *CS* shot because of the "Smoothness of camera motion," which was the most cited reason for the preference. These preferences align with the panoramic shot trends; when distortion is not an issue, viewers seemed to prefer smooth camera motions where the path and FOV are not varied.

Logical Connection Shot

For logical connection shots, there was a significant preference for C_sS to C_eS, both of which maintain a constant FOV and smooth path. The individual scene data further supports the takeaways from the panoramic and tracking shots.

For L_1, we found that viewers preferred C_sS over C_eS and V_S. The preference for C_sS over V_S aligns with our findings for the panoramic shots, where viewers favored keeping the FOV constant. Not surprisingly, 48% of viewers cited "Smoothness of the camera motion" as a reason for making this choice. For the preference of C_sS over C_eS, "Prefer to see more of the scene for context" was the most specified reason. Looking at the scene, we find that L_1 starts with a wide shot of a buffet table and ends with a tight shot of the person about to eat the food. Thus, C_sS maintains the wider FOV required for the buffet table, which shows more context in the shot.

In L_2, which starts with a tight shot and ends with a wide shot, we found no significant preferences between any of the shot variations for the scene. However, in the cases where the wider C_eS shot was not preferred, we found that "Amount of stretching or distortion in the frame" was cited frequently, in 50% of comparisons against C_sS and in 37% of comparisons against V_S. This data agrees with the trends that we saw for the other shot categories where distortion was not preferred.

DESIGN GUIDELINES

Based on our two studies, we propose the following design guidelines for extracting RFOV shots from 360 event footage.

Static Camera Shots

Center the subject. The overall preferences argue for keeping the subject roughly in the center of the frame.

Avoid cropping people. In general, people did not like framings that cut off people (especially the main subject) around the edges of the frame.

Gaze matters for wider shots. The traditional gaze direction rule is preferable for wider shots but less so for tight crops.

More context for wider shots. The data also suggests to include more context in wider shots. Even extreme framings of the main subject can be acceptable if they add useful context.

Avoid distractors. While more context is preferable, distracting background objects have a negative impact on the shot.

Center whole subject in tight shots. For tight shots, people prefer the entire subject (i.e., head and shoulders) to be centered rather than just the head.

Moving Camera Shots

Make the shot smooth. Constant, wider FOV shots with smooth trajectories are preferable for most shots.

Avoid distortion. Distorted shots with an overly wide FOV should be avoided.

Avoid varying FOV. For tighter shots, varying the camera trajectory to include POIs is preferable to varying the FOV.

ALGORITHMS

We use our design implications to formulate initial algorithms to extract static and moving camera RFOV shots from 360 equirectangular footage. The input to these algorithms is a temporal segment of the 360 footage from which to extract the shot and a set of one or more POIs (e.g., the main subject in a static camera shot). Our current implementation outputs 4:3 aspect ratio shots, but the core algorithms could easily be extended to handle other form factors. Note that we implement some, but not all of our design guidelines.

Equirectangular and Rectilinear Projections

The coordinate system for equirectangular video is defined by an azimuth that ranges from $-\pi$ to π in the horizontal dimension and an elevation that ranges from $-\pi/2$ to $\pi/2$

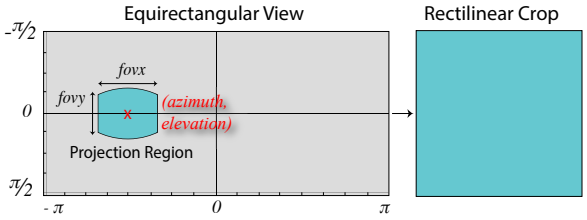


Figure 7. Equirect coordinate system. The projection region becomes the rectilinear crop.

in the vertical dimension. To extract a rectilinear shot, we define a center azimuth and elevation, as well as horizontal and vertical FOVs to specify the shot size. We use the method of Gardner et al. [10] to compute the rectilinear projection.

Preprocessing

Our shot extraction algorithms rely heavily on labeled data about the human subjects in the 360 scene. We extract this data by running each input video through a face tracking, face detection and pose detection pipeline in the preprocessing step.

We first apply the face detection and tracking algorithms of Li et al. [16] and Saragih et al. [19] to obtain a bounding box and 68 facial landmark points for every detected face inside each frame of an input video. We found that these algorithms do not return reliable results when operating directly on the equirectangular video due to distortion. To accommodate this issue, we project the equirectangular video into eight overlapping rectilinear crops at the same elevation and spaced 45 degrees apart in the azimuthal plane with 60° horizontal and 45° vertical FOV. We then apply face detection and tracking to each crop and translate these rectilinear results back into the equirectangular space to obtain a mapping of all the detected faces for each frame of each input scene.

We then extract the pose of each subject by applying the method of Cao et al. [4] and Wei et al. [21] on every input equirectangular video to obtain 18 keypoints around the eyes, shoulders and joints for every detected person in every frame.

Extracting RFOV Static Camera Shots

To extract static camera shots from the pre-processed footage, we first generate candidate shots at three different crop levels (close, medium close, medium) that frame the subject at different horizontal positions without cutting the subject off. To avoid overly wide shots that introduce distortion, we set the maximum horizontal FOV, FOVX_{max}, to 53° and the maximum vertical FOV, FOVY_{max}, to 40° for all of our crops. We then evaluate each candidate shot based on additional design objectives and select the best result for each crop level.

Generating Candidate Shots

First, we compute the *subject bounding box* B_S that contains all of the subject's movement throughout the specified segment of the 360 video. This box defines the portion of the input footage that cannot be cut off by an RFOV shot. To do this, we examine the face tracking landmarks in every frame and determine the leftmost and rightmost azimuth (a_L, a_R) and top and bottom elevation (e_T, e_B) values for the landmarks.

Next, we compute a *crop box* B_C at each crop level that we use to derive the candidate shots for that level. For each crop level, we use the recommended framing heights of Arijon [2] to determine how much of subject to include in the shot. Then, for a given framing height, we determine the target elevation and vertical FOV by computing the distance from the top of B_S (i.e., the top of the subject's head) to the relevant part of the subject's body, as defined by the detected pose in the first frame of the shot. We also align the azimuth of the crop box to B_S. Given the vertical FOV, we compute the horizontal FOV based on the 4:3 output aspect ratio. If the FOVs are larger than FOV_{max}, it means that this crop level is too wide and will produce undesired distortion. On the other hand, if the FOVs are smaller than $B_s.fov$, the crop level is too tight and will crop the subject. In both cases, we mark the crop level invalid.

Finally, for each valid crop level, we shift B_C horizontally to generate five candidate shots, defined by bounding boxes $(B_1, \ldots, B_5)$, that frame the subject at different uniformly spaced horizontal positions.

Evaluating Candidate Shots

The algorithm evaluates each candidate shot according to an objective function

$$E(B) = C(B) + \sum_{i}^{N} F(i) \tag{1}$$

where B is the candidate box, $C(B) = |B_S.azi - B.azi|$ is a centering cost that penalizes B as it deviates from the horizontal center of B_S, and $F(i) = D(i) + 1/P(i)$ is a framing cost that measures the quality of frame i with terms that encode two of our design guidelines: $D(i)$ is the fraction of pixels in the frame belonging to distracting objects as detected by the method of Fried et al. [8], and $P(i)$ is the total number of complete faces within the frame. By penalizing the presence of distractors and encouraging the presence of faces, the framing cost favors shots with little visual noise in the background and more context from other people in the scene.

Extracting RFOV Moving Camera Shots

In our moving camera study, we found that viewers generally preferred smooth camera paths with a constant trajectory and FOV. However, this preference did not hold for scenes with distortion. Thus, our algorithm first computes the minimum FOV required to capture all POIs with a smooth, linear trajectory. If the FOV does not exceed FOV_{max}, we generate a smooth, constant FOV shot. If it does, we use a tighter framing and compute a trajectory that keeps the POIs in the frame.

Determining the Minimum FOV

To determine the minimum FOV_{POI} that still frames all POIs, we compute the crop box (as described previously) at the medium level for every POI and then take the union of these boxes. For tracking shots, we treat the position of the main subject in each frame of the scene as a separate POI. The vertical size of the combined box represents the minimum vertical FOVY_{POI}, and the vertical center of the box determines the elevation for a linear trajectory shot.

Generating the Camera Move

If $FOVY_{POI}$ is less than $FOVY_{max}$, we calculate the corresponding horizontal FOV and create a linear, constant elevation shot that slides horizontally across the POIs. We set the starting and ending azimuths such that the first and last POIs are horizontally centered. However, if $FOVY_{POI}$ is greater than $FOVY_{max}$, we clamp $FOVY_{POI}$ to the maximum allowed value before computing the corresponding horizontal FOV. With this tighter framing, we then compute the azimuth and elevation required to center every POI and construct a B-spline that interpolates these points. For logical connection shots, we cannot predict the effect of constraining the FOV for user specified, non-human POIs. Thus, we vary the shot size by linearly interpolating between the FOVs of the two POIs.

Pan Timing

For the panoramic and logical connection shots, we adjust the timing of the camera motion using a sinusoidal easing function that smoothly increases and decreases the speed of the camera at the start and end of the trajectory. For tracking shots, the timing must take into account the subject's motion. We calculate the subject's displacement at each frame to interpolate a B-spline timing function that follows the subject's pace.

SHOT EXTRACTION INTERFACE

In our initial experiments, we implemented an explicit RFOV shot extraction interface with simple controls to manually edit the shot size and position. However, this interface was difficult to use in situations where either the subject or the virtual camera moved over time. To address these challenges, we developed a new user interface that incorporates our automatic shot extraction algorithms (see Figure 8).

Our interface reads in a set of preprocessed equirectangular videos with tracked faces and poses. To begin, the user can select any of these source videos in the Footage Selector and preview it in the Equirectangular Player, which supports standard timeline navigation controls. Once she identifies a temporal segment of the video (i.e., a scene) from which she wants to extract RFOV shots, she marks that segment with "in" and "out" markers and then selects the type of shot from the Shot Controls to activate the corresponding extraction mode.

In extraction mode, a box appears over all automatically detected faces for the scene at the current frame in the Equirectangular Player. Figure 8 shows the panoramic extraction mode. She then uses these boxes to identify the subjects for her shot. The specific interaction depends on the shot type:

For a *static shot*, she simply clicks on the face of the main subject for the shot.

For a *panoramic shot*, she starts by selecting the first face to appear in the pan, marked in green, followed by the last face in the pan, marked in red. She then selects any other faces that should appear in the shot, marked in blue. She can also draw a freeform box to specify non-face POIs. Our interface checks whether all the intermediate POIs are between the start and end POIs and warns the user if not.

For a *tracking shot*, she selects a person to follow.

For a *logical connection shot*, she selects the start and end POIs for the shot. As with the panoramic shot, she can either select detected faces or draw freeform boxes.

After she specifies the subjects and POIs for the shot, she names it and hits the Extract button. The system then runs our automatic extraction algorithms and displays the generated shots with the lowest three costs at the top of the Crops Viewer on the right. The user can playback the shots in the Crops Viewer and filter them to remove the shots she does not want.

RESULTS

We used our interface to generate a number of static and moving camera RFOV shots, some of which can be seen in Figure 9. These shots aligned well with the design guidelines that we distilled from our studies. For example, as shown for set 1 of Figure 9, our algorithm places the subject on the right of the frame instead of the more commonly preferred center. Whereas the central framing crops off a portion of the woman next to him, creating a distractor, the framing chosen by our algorithm includes both of them in the scene, adding context. A similar situation arises in set 2, where our algorithm positions the subject close to the right edge of the frame despite the fact that he is also looking to the right. As in the previous example, the more centered framing in 2b partially cuts off the man the subject is taking to, only showing his shoulder and part of his face, which serves as a distractor. Finally, in set 3, our system decides to generate a pan with an uneven trajectory to avoid the jarring distortion in the wide shot (3b). We show these and other results in our paper video and website.

LIMITATIONS AND FUTURE WORK

Our work takes an initial step in exploring the use of 360 cameras for event cinematography. However, there are many remaining questions and challenges to consider in future work.

Finding Relevant Footage

Although our approach enables users to capture footage passively and extract usable RFOV from specified "scenes" of interest, scrubbing through videos to find relevant moments is nontrivial. Automating this task for large video datasets remains a challenging problem. As mentioned in our related work, others have used saliency metrics to distinguish potentially interesting moments. However, it is a more challenging problem to consider how to detect people or objects that are not only salient, but also semantically meaningful. We see this as the next step in using our design guidelines to automate RFOV shot extraction from 360 event footage.

Editing

Looking beyond the problem of extracting individual shots, our overall vision is to automatically generate edited summary videos that compose multiple shots in a pleasing way. Editing event footage comes with its own set of challenges and tradeoffs. We believe that we can apply our methodologies to better understand what those tradeoffs are and how to address them.

Scene Generalization

For the studies, we searched through our footage to curate scenes that we commonly observed in professionally edited

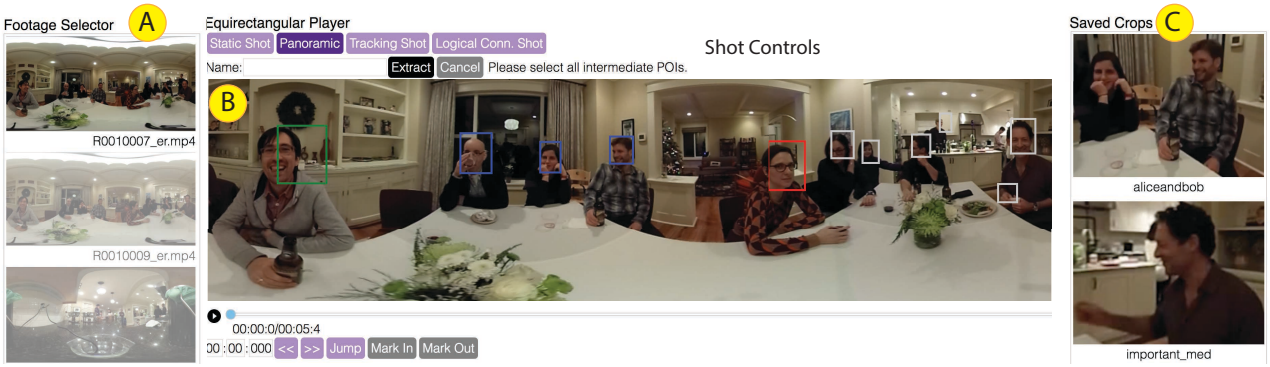


Figure 8. Our shot extraction interface has three main components: a footage selector where users select their source footage (A); an equirectangular player where users specify different types of RFOV camera shots (B), and a saved extractions viewer where users can view the extracted crops (C).

Figure 9. Our results (a) and alternatives (b). These alternatives were not selected by our algorithm, though could have been if we adhered strictly to high level user preferences or cinematography rules.

event highlight reels. We found that people generally converse in a similar way (i.e., sitting or standing in small groups) across different types of social events; thus, we selected dialogue scenes for our static camera and panorama shots. For the non-dialogue static and logical connection shots, we selected scenes of people interacting with familiar components of party environments (e.g., food). For tracking shots, we searched for people moving through the event space. While we believe these scenes are representative for social event highlight reels, considering a broader range of scenes (e.g., "b-roll" that focuses solely on the environment, variations on the scene types) would strengthen and expand the applicability of our findings.

Camera Configuration and Human Camera Interaction

We attempted to set up the cameras at areas of interest where we predicted people would gravitate. However, within these areas, there are still a number of possible camera configurations. Considerations include placing the cameras to allow as much 360 capture as possible (i.e., not against walls), while minimizing intrusiveness and not physically obstructing guests' paths. Based on our capture experiences, we observed some situations where people felt uncomfortable being recorded. One is when guests are not told ahead of time when they will be filmed. In another situation, we did not set up the cameras until after the party had started, at which point they were placed in very conspicuous locations, such as the middle of a dining table. This turned out to be very obtrusive to the guests. We alleviated some of this discomfort at later events by setting up the cameras beforehand.

As noted in the introduction, the question of how to optimally configure these cameras to be effective, yet non-intrusive, is an interesting direction for future work. Considering footage captured using a wider variety of camera configurations could broaden the scope of our findings.

Evaluation

In this iteration of our work, we were not able to run an evaluation of our system. An interesting extension would be to compare the results of our algorithm against those of the saliency based algorithms mentioned in our related work.

CONCLUSION

The improving quality and diversity of commercial cameras makes it possible to explore new capture configurations for a range of different applications. We focus on the task of creating video summaries of social events and propose the use of static 360 video cameras to acquire footage. A key advantage of this setup is that the cameras can capture large portions of the event in a passive way that does not require a dedicated videographer. Our work takes initial steps towards identifying common challenges and tradeoffs in extracting good RFOV shots from 360 event footage. By analyzing viewer preferences, we find that central framings and smooth camera moves are generally preferred, but in many cases, the arrangement of people and objects in the scene require that we take into account other objectives, like avoiding distractors, awkward cropping boundaries, and distortion due overly wide FOVs. In addition, our automated algorithms provide preliminary evidence that these design guidelines can be operationalized to extract good shots from 360 footage with small amounts of user input. Looking ahead, there are still many open questions around the best ways to make use of 360 cameras to produce RFOV video. However, we believe our work represents a valuable stepping stone for future research on this topic.

ACKNOWLEDGMENTS

We thank our party guests and study participants; Jacob Cao for his professional insight; Joy Kim, Daniel Epstein, and Eytan Adar for statistics help; and Jane E for help on figures.

REFERENCES

1. Ido Arev, Hyun Soo Park, Yaser Sheikh, Jessica Hodgins, and Ariel Shamir. 2014. Automatic editing of footage from multiple social cameras. *ACM Transactions on Graphics (TOG)* 33, 4 (2014), 81.
2. D. Arijon. 1976. *Grammar of the Film Language*. Hastings House. https://books.google.com/books?id=Ga0fAQAAIAAJ
3. Michael Bianchi. 2004. Automatic video production of lectures using an intelligent and aware environment. In *Proceedings of the 3rd international conference on Mobile and ubiquitous multimedia*. ACM, 117–123.
4. Zhe Cao, Tomas Simon, Shih-En Wei, and Yaser Sheikh. 2017. Realtime Multi-Person 2D Pose Estimation using Part Affinity Fields. In *CVPR*.
5. Peter Carr, Michael Mistry, and Iain Matthews. 2013. Hybrid robotic/virtual pan-tilt-zom cameras for autonomous event recording. In *Proceedings of the 21st ACM international conference on Multimedia*. ACM, 193–202.
6. Thomas Deselaers, Philippe Dreuw, and Hermann Ney. 2008. Pan, zoom, scanâĂŤtime-coherent, trained automatic video cropping. In *Computer Vision and Pattern Recognition, 2008. CVPR 2008. IEEE Conference on*. IEEE, 1–8.
7. Chen Fang, Zhe Lin, Radomír MËĞech, and Xiaohui Shen. 2014. Automatic image cropping using visual composition, boundary simplicity and content preservation models. In *Proceedings of the 22nd ACM international conference on Multimedia*. ACM, 1105–1108.
8. O. Fried, E. Shechtman, D.B. Goldman, and A. Finkelstein. 2015. Finding distractors in images. In *Computer Vision and Pattern Recognition (CVPR), 2015 IEEE Conference on*. 1703–1712.
9. Vineet Gandhi, Remi Ronfard, and Michael Gleicher. 2014. Multi-clip video editing from a single viewpoint. In *Proceedings of the 11th European Conference on Visual Media Production*. ACM, 9.
10. Marc-André Gardner, Kalyan Sunkavalli, Ersin Yumer, Xiaohui Shen, Emiliano Gambaretto, Christian Gagné, and Jean-François Lalonde. 2017. Learning to Predict Indoor Illumination from a Single Image. *CoRR* abs/1704.00090 (2017). http://arxiv.org/abs/1704.00090
11. Li-wei He, Michael F Cohen, and David H Salesin. 1996. The virtual cinematographer: a paradigm for automatic real-time camera control and directing. In *Proceedings of the 23rd annual conference on Computer graphics and interactive techniques*. ACM, 217–224.
12. Rachel Heck, Michael Wallick, and Michael Gleicher. 2007. Virtual videography. *ACM Transactions on Multimedia Computing, Communications, and Applications (TOMM)* 3, 1 (2007), 4.
13. Hou-Ning Hu, Yen-Chen Lin, Ming-Yu Liu, Hsien-Tzu Cheng, Yung-Ju Chang, and Min Sun. 2017. Deep 360 Pilot: Learning a Deep Agent for Piloting through 360âŰę Sports Videos. *CVPR* (2017).
14. Eakta Jain, Yaser Sheikh, Ariel Shamir, and Jessica Hodgins. 2015. Gaze-driven video re-editing. *ACM Transactions on Graphics (TOG)* 34, 2 (2015), 21.
15. Wei-Sheng Lai, Yujia Huang, Neel Joshi, Chris Buehler, Ming-Hsuan Yang, and Sing Bing Kang. 2017. Semantic-driven Generation of Hyperlapse from 360Âř Video. *CVPR* (2017).
16. Haoxiang Li, Zhe Lin, Xiaohui Shen, Jonathan Brandt, and Gang Hua. 2015. A convolutional neural network cascade for face detection. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 5325–5334.
17. Feng Liu and Michael Gleicher. 2006. Video retargeting: automating pan and scan. In *Proceedings of the 14th ACM international conference on Multimedia*. ACM, 241–250.
18. Amy Pavel, Bjoern Hartmann, and Maneesh Agrawala. 2017. Shot Orientation Controls for Interactive Cinematography with 360 Video. In *To appear, UIST'17*. ACM.
19. Jason M Saragih, Simon Lucey, and Jeffrey F Cohn. 2009. Face alignment through subspace constrained mean-shifts. In *Computer Vision, 2009 IEEE 12th International Conference on*. Ieee, 1034–1041.
20. Yu-Chuan Su and Kristen Grauman. 2017. Making 360Âř Video Watchable in 2D: Learning Videography for Click Free Viewing. *CVPR* (2017).
21. Shih-En Wei, Varun Ramakrishna, Takeo Kanade, and Yaser Sheikh. 2016. Convolutional pose machines. In *CVPR*.
22. Jianming Zhang, Stan Sclaroff, Zhe Lin, Xiaohui Shen, Brian Price, and Radomir Mech. 2016. Unconstrained salient object detection via proposal subset optimization. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 5733–5742.
23. Luming Zhang, Mingli Song, Yi Yang, Qi Zhao, Chen Zhao, and Nicu Sebe. 2014. Weakly supervised photo cropping. *IEEE Transactions on Multimedia* 16, 1 (2014), 94–107.