Preference and Artifact Analysis for Video Transitions of Places

JAMES TOMPKIN, MPI für Informatik and Intel Visual Computing Institute, Germany
MIN H. KIM, KAIST, Korea
KWANG IN KIM, MPI für Informatik, Germany
JAN KAUTZ, University College London, United Kingdom
CHRISTIAN THEOBALT, MPI für Informatik, Germany

Emerging interfaces for video collections of places attempt to link similar content with seamless transitions. However, the automatic computer vision techniques that enable these transitions have many failure cases which lead to artifacts in the final rendered transition. Under these conditions, which transitions are preferred by participants and which artifacts are most objectionable? We perform an experiment with participants comparing seven transition types, from movie cuts and dissolves to image-based warps and virtual camera transitions, across five scenes in a city. This document describes how we condition this experiment on slight and considerable view change cases, and how we analyze the feedback from participants to find their preference for transition types and artifacts. We discover that transition preference varies with view change, that automatic rendered transitions are significantly preferred even with some artifacts, and that dissolve transitions are comparable to less-sophisticated rendered transitions. This leads to insights into what visual features are important to maintain in a rendered transition, and to an artifact ordering within our transitions.

Categories and Subject Descriptors: [**Computer Graphics**]: Image manipulation—*Image-based rendering*; [**Computer Graphics**]: Graphics systems and interfaces—*Perception*

General Terms: Human factors

Additional Key Words and Phrases: Video-based rendering, video transition artifacts.

ACM Reference Format:

James Tompkin, Min H. Kim, Kwang In Kim, Jan Kautz, Christian Theobalt. 2013. Preference and Artifact Analysis for Video Transitions of Places. ACM Trans. Appl. Percept. 10, 3, Article 0 (August 2013), 18 pages.
DOI = 10.1145/0000000.0000000 http://doi.acm.org/10.1145/0000000.0000000

1. INTRODUCTION

For over a hundred years, audiences watching movies have become familiar with the effects of placing video clips in sequence. The switch between clips is called a transition and, while this is most commonly an instant transition or *cut*, various transitions exist to convey information and create effects in the mind of the viewer. Transitions were an artistic introduction which allowed movies to transcend the restrictions of space and time in theater (as well as, though unimportantly, the physical restrictions

DOI 10.1145/0000000.0000000 http://doi.acm.org/10.1145/0000000.0000000

of the amount of film in a reel). Since then, advances in physical and digital visual effects have given movie-makers creative freedom over both discontinuous and continuous or *seamless* transitions.

With the proliferation of video cameras and the rise in video sharing platforms on the Web, new innovative interfaces for video collections are emerging which attempt to link videos of similar content with transitions [McCurdy 2007; Ballan et al. 2010; Tompkin et al. 2012]. One particularly promising subset of videos are those of places: with these, it is possible to automatically reconstruct the geometry in the video and provide a virtual *video-based rendering* transition to seamlessly join two videos. However, it would be incorrect to assume that movie-style transitions are appropriate for this application. For movies, the most significant goals of a transition are to drive emotion and story, whereas the least significant goals are to maintain the two- and three-dimensional *space of action* [Murch 2001, p.18]. We hypothesize that it is important to seamlessly maintain the sense of orientation in the viewer when transitioning between video clips of a place by relating the two- and three-dimensional space of action, else the viewer will become lost in the environment.

Cut transitions represent a change of context and are effective when visual displacement is great [Murch 2001, p.6], but rules for their use do not focus on maintaining the space of action [McCurdy 2007, p.101]. *Dissolve* transitions suggest a change of place or the passage of time [Dmytryk 1984, ch.13] and likewise do not intentionally maintain the space of action. At other times, movie-makers create seamless transitions by employing computer generated visual effects to create physically implausible camera moves, such as simulating a camera moving through the lock in a door [Panic Room, Fincher, 2002]. Recent advances in computer vision and mapping have required new photo transitions not previously seen in movies [Snavely et al. 2006; Snavely et al. 2008; Goesele et al. 2010], and these can be adapted for video. Which transitions are most suitable for video collections of places?

Furthermore, state-of-the-art vision-based methods are able to automatically reconstruct the static geometry of a scene from video. This opens the door for *video-based rendering* transitions; however, the door is not yet wide open. Geometry reconstruction techniques are often brittle, and cannot yet cope with the complex dynamic scenes in daily life. This causes many different kinds of artifacts. Some techniques attempt to overcome these problems with heuristic methods and approximations, but again these can introduce artifacts. That said, some artifacts are more objectionable than others, and investigating which kinds of artifacts are more perceptually off-putting would help direct future work.

This exploratory work attempts to determine suitable video transitions for exploring video collections of places, and an artifact preference order to guide future video transition work. First, we choose and justify a selection of movie and graphics-rendered video transitions to compare in an experiment, and categorize possible artifacts within these transitions (Section 3). Next, we assess transitions for participant preference and present results (Section 4). With these experiment results, we generate heuristics for good transitions, analyze the response to artifacts within transitions, and produce an ordering of artifacts to direct future work (Section 5).

2. BACKGROUND

As vision-based methods have only recently been applied to automatically generate video transitions, there is little work in the literature on their analysis. However, some works exist to analyze the perception of image-based rendering. Morvan and O'Sullivan [2009] look at the effect of occluders in transitions between panoramas, and between panoramas and 3D models. They discover that dissolve transitions are perceptually equivalent to transitions where occluders are explicitly segmented. The authors also assess perceptually the effect of simplifying geometry proxies [Morvan and O'Sullivan 2009]. Vangorp et al. [2011] build on this work to assess the effect of planar proxies on the perception of building facades, and categorize artifacts common in these cases such as static ghosting and scene skewing (parallax distortion). In follow-up work, they experimentally measure the perception of skew

on building facade balconies, and define a perceptual model to guide the location of appropriate cameras for viewing street-level image-based renderings [Vangorp et al. 2013]. Stich at al. [2011] develop a perceptually-motivated image interpolation method for image sequences. Mustafa et al. [2012] follow this up with an electroencephalography method for measuring the presence of and preference for artifacts in image-based rendered images, such as blurring, ghosting, and popping. However, none of these methods address the important differences in video-based rendering.

For video-based rendered transition preference, some work exists: McCurdy [2007] provides evidence to suggest that transitions in low frame-rate multi-camera situations aid scene understanding, and that registered transitions further aid comprehension. Ballan et al. [2010] ask participants for their preference to different transitions, and find that many might often be used but that motion-tracked dots was rarely preferred. To our knowledge, these are the only works which assess preference for video-based transitions, and they do not investigate specific artifact issues within rendered transitions.

3. TRANSITION DESIGN

We wish to study transitions commonly used in both movies and graphics-rendered applications: any example, interactive or otherwise, in which media is digitally transitioned from one image to another. From movies, we include *cut* and *dissolve* (or cross-dissolve) transitions. Cut transitions form a baseline as the simplest way to join two clips. We include dissolve transitions, which commonly represent the passage of time, as there are time differences between clips in our video collections. Other transitions, such as wipes and reveals, are less common; we do not include them as they add nothing over a cut or dissolve to help maintain the space of action.

With computer graphics, we can generate seamless transitions which rely on scene geometry and a virtual camera. We call these *full 3D dynamic* transitions as they require full scene geometry (or a suitable geometry proxy). They maintain the space of action and sense of orientation in the viewer by rendering a perspective-correct view from virtual cameras that join both clips. This transition also maintains as much as possible the motion of dynamic objects by projecting playing video clips onto the scene geometry. We also wish to test seamless spatial transitions which do not maintain the motion of dynamic objects during the transition. In *full 3D static* transitions we render a virtual view using scene geometry as before, but do not keep playing the video as the virtual camera moves. During the transition the world appears as if time has stopped, similar to time-slice photography [Early Work, Macmillan 1980; Frozen Time, Debuchi 1982]. Unlike *full 3D dynamic* transitions, the object motions in the clips do not blend into each other. This will test if and when it is important to maintain dynamic object motions.

Both *full 3D* transitions use accurate scene geometry, but many existing applications employ simple proxy geometry, such as a single plane, to represent scenes [McCurdy 2007; Snavely et al. 2006; Vangorp et al. 2011]. Such *plane* transitions work well for camera rotations, but suffer artifacts if the start and end clips are shot from different positions. This transition type is currently popular among commercial touring and mapping applications, and is a baseline as the simplest registered graphics-rendered approach.

If partial scene geometry is available, *ambient point clouds* (APC) can help fill in gaps in geometry as an alternative to partial planar proxies. Goesele et al. [2010] introduced these transitions to provide visual hints at motion and depth. We include APC transitions as they represent the state of the art in automatic graphics-rendered transitions from community photo collections, where it is often the case that only incomplete geometry is recoverable or available.

Video morph or *warp* transitions are often used as a special effect in movies to transform one object into another, but recent advances in robust feature point correspondence have allowed view change transitions as well [Lowe 2004; Lipski et al. 2010]. Warp transitions provide an alternative both to

transitions that require geometry and to plane transitions: while plane transitions can be classified as a subset of warps with global 2D transformations (4-point correspondence), warps can also be computed from many hundreds of points to exploit more accurate correspondence. We include these many-correspondence warps in our comparison as they maintain the space of action and are visually different from other transitions.

We exclude other transition types which currently require manual work, as might be generated for a feature film. This includes any transition type which requires interactive or background-based segmentation [Horry et al. 1997; Chaurasia et al. 2011]. As we wish to test vision-based reconstruction methods, we exclude transitions which rely on laser-scans [Morvan and O'Sullivan 2009] and hand-modelled geometry [Debevec et al. 1998; Oh et al. 2001], though technically these are full 3D transitions with varying geometry accuracy and fidelity.

3.1 Transition Implementations and Artifacts in Detail

The practical creation of each of the transition types is described concisely in the appendix, with detailed explanation in the supplemental material: For each transition, the explanation contains a historical review of application, our technical method to achieve the transition, and an explanation of artifacts that may appear. For referencing in this section, we collate and categorize all feature and artifact types in each transition in Table I. We will use this table to cross-reference comments from participants in the experiment in Section 4.3.

The identified major artifacts in our transitions and their causes are:

Ghosting. *Static* ghosting on scene objects such as buildings is caused by competing projections from different camera poses onto incorrect proxy geometry. *Dynamic* ghosting on scene objects such as pedestrians occurs as objects must fade in/out across the transition as the video source switches.

Orientation loss. Caused when there is no explicit registration of the two videos and hence no virtual camera transition.

Bad correspondence swirls. Specific to image-based warps; caused by incorrect correspondences between images, e.g., from confusing repeated building features. This creates vortices in the output correspondence field.

Edge flickering. Specific to image-based warps; correspondence fields can vary rapidly at edges as material is revealed/concealed by real camera motion.

Skewed scene. Parallax distortion caused by incorrect planar proxy geometry under wide baselines.

Pepper noise. Specific to APC; caused when individual pixels are not covered by the point cloud proxy and so appear black.

Multiple scene elements. A form of static ghosting; incorrect registration of video frames to scene geometry under camera shake causes the appearance of multiple scene elements.

Recovered geometry failures. Many factors cause holes in geometry reconstructions: specular objects, dynamic objects, and insufficient baseline are the major causes.

Empty black regions. Caused by virtual camera motions which conflict with real camera motions and create areas with no projection, see Section 5.1.

3.2 Clip Choice

Transition preference must be tested across different clips of different scenes, as transition preference may vary between scenes and scene elements. Beyond this, we consider that transition preference may

	Cut	Dissolve	Warp	Plane	APC	Full3DDyn	Full3DSta
Feature							
Registered scene			●	●	●	●	●
3D effect			○[1]		●	●	●
Dynamic objects		●	●	●	●	●	
Smooth virtual camera (Fig. 6)			○[2]	●	●	●	●
Common familiarity	●	●					
Signifies change of time		●					
Explicit motion cues					●		
Frozen time							●
Artifact							
Ghosting (static objects)		●		●[3]			
Ghosting (dynamic objects)		●	●	●	●	●	
Orientation loss	●	●					
Bad correspond. swirls (*a, left*)			●				
Edge flickering (*a, right*)			●				
Skewed scene (*b*)				●		○[4]	○[4]
Pepper noise (*c*)					●		
Multiple scene elements (*d*)					○[5]	●	○[6]
Recovered geom. failures (*d*)					●	●	●
Empty black regions (*e*)				●	○[7]	○[8]	○[8]

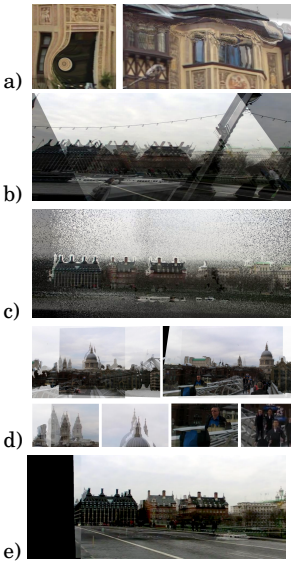
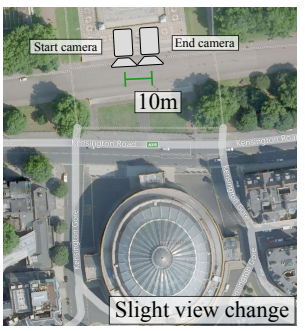


Table I. : *Left:* A table collating all features and artifacts for each transition type. *1*: Partial, only with good regular correspondence and flow correction. *2*: Velocities only from feature-point tracks. *3*: Ghosting is present in almost all transitions because the plane is an inaccurate proxy to the true geometry. *4*: On proxy planes only. *5*: An image forms within the APC as it appears as a noisy plane during slight view changes. *6*: Not as prominent as dynamic case as registration at anchor frames is often better than video-geometry registration. *7*: APC partially reduces empty regions; introduces pepper noise. *8*: Minimized given video-geometry registration. *Right:* Artifacts *a-e* as in table.

vary based on the view change between the start and end video clips in a transition. For instance, two clips with no camera motion shot from the same position would transition with less ghosting in a dissolve than clips shot from different positions. Likewise, a full 3D transition adds very little to two clips shot from the same position and may introduce artifacts from missing geometry, but provides a smooth virtual camera transition that respects the parallax caused by clips shot from different positions.



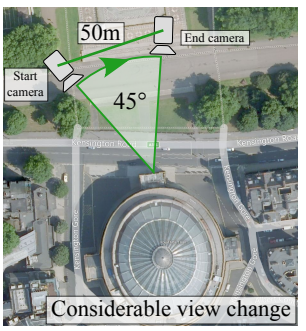


Fig. 1: Scene 4. *Left*: In the slight case, the camera translates 10 meters. *Right:* In the considerable case, the camera translates 50 meters and undergoes a $45°$ yaw rotation. Map data ©2013 Google; see acknowledgments.

To include view change as a variable in our experiment, we create pairs of transitions which show both a slight view change and a considerable view change (Figure 1). A slight view change is a transition from one video clip to another where the visual elements in the scene, such as the buildings and people, approximately maintain their size and position. These slight view changes can still allow considerably different camera positions as we allow zoom to vary. All camera positions in our chosen scenes subtend a $10°$ cone ($9°$ average) with its apex approximately at the depth of the scene in the middle of the video frame. A considerable view change transition occurs between clips with camera positions outside a $10°$ cone, and in our chosen scenes this is maximally $55°$ ($34°$ average).

(a) Scene 1: County Hall, sets 1 & 2.

(b) Scene 2: Palace of Westminster, sets 3 & 4.

(c) Scene 3: Victoria Embankment, sets 5 & 6.

(d) Scene 4: Royal Albert Hall, sets 7 & 8.

(e) Scene 5: Millennium Bridge, sets 9 & 10.

Fig. 2: *Left*: Start frame for slight view changes. *Middle*: Start frame for considerable view changes. *Right:* End frame for both view change conditions.

For each scene, we choose one clip as the reference end clip. Then, we choose two start clips: one each for slight and considerable view changes. The scenes contain buildings in the middle distance (approximately 50-300 meters) along with smaller dynamic objects such as birds, boats, cars, and pedestrians. The 5 scenes were chosen as they each display a potentially difficult situation (Figure 2): Scene 1: dynamic objects at boundary; Scene 2: many dynamic objects with view occlusions and panning camera; Scene 3: panning cameras and dynamic objects; Scene 4: fast moving dynamic objects and shaking camera/rolling shutter; Scene 5: complicated foreground objects and moving, shaking camera. Even in slight view changes, the real camera position may vary substantially due to zooms or motion towards/away from the scene, e.g., in Scene 5, the virtual camera moves 200 meters. Each transition will consist of 2 seconds of video, plus 1 second of transition, followed again by 2 seconds of video. Some clips contain camera shake — if this is the end clip then the shake is present in both slight and considerable view changes. In Scene 4, the shake is so significant as to cause rolling shutter artifacts.

4. EXPERIMENT

We hypothesized that video transition type preference depends on the severity of view change. For instance, if the camera pose changes considerably, then transitions based on 3D scene reconstructions might perform better than those with no geometry; if the camera pose change is slight, then warps and dissolves might perform better.

We conducted a psychophysical experiment by ordinal ranking. We designed our experiment to quantify the perceived preferences of transitions across different scenes. We then analyzed our observations with classical multidimensional scaling [Torgerson 1958; Borg and Groenen 2010].

Stimuli We employed seven transitions: *cut*, *dissolve*, *warp*, *plane*, *ambient point cloud*, *full 3D static*, and *full 3D dynamic*. Each transition was applied to videos of five different scenes with two view conditions. Each stimulus contained two seconds from the start clip, one second of transition, and two seconds from the end clip, all running at 60 Hz. All stimuli can be seen in our supplemental video.

Fig. 3: Screenshot of the interface for the transition ranking experiment.

Procedure For the transition stimuli in each set, observers ranked the transitions by *preference*. Ranking order is an alternative to pairwise comparison that, with analysis, produces relative scores and confidence intervals. While ranking produces greater dispersion than pairwise comparison, we chose this method as, with 70 stimuli, it is more time efficient. This approach might introduce a bias from visual familiarization with artifacts as all participants rank all sets; however, practically avoiding this risk requires many participants and would be extreme for this exploratory work.

As such, first we explained the scenario of video transitions to participants and two transitions were shown as training. Then, participants were asked to rank each set of video transitions given in random order. In each set, transitions were randomly placed into a vertical video list (Figure 3). Participants dragged and dropped videos to reorder the list from most preferred to least preferred. Each of the videos could be played any number of times. Of the 21 participants, twelve were self-described experts with experience in graphics and media production, four were amateurs, and five were novices. It took 52 minutes on average to complete the experiment. Participants were interviewed to describe their response to the transitions.

4.1 Data Analysis

Rescaling We analyzed the data assuming the conditions of Case V in Thurstone's Law of Comparative Judgment [Engeldrum 2000], which yields *psychometric scaling* values in the form of a z-score of the perceived image quality. In particular, this psychometric scaling analysis generates interval scales of image quality by human measurement. We built a proportion-of-preference matrix of all possible comparisons from the ranking data, and then applied a logistic psychometric model [Engeldrum 2000] to convert the observed probability to a *logit* quantity. The logit was assumed to be a linear function of the perceived image quality, of which the gain and offset constants were found by using a linear least squares fit [Cui 2000]. Finally, we calculate the z-score of the logit as a *psychometric scale* by using the inverse cumulative distribution function. See Figure 4 for the rescaled participant responses.

Statistical Significance We tested statistical significance for transition types across all scenes by using the Student's t-test. Following Thurstone's Law, we assume that the probability density function of the perceptual discrimination process follows a normal distribution function in the psychological continuum. We regard the standard deviation as the discrimination dispersion. Therefore, we tested pairwise statistical significance with the Student's t-test parametric approach, rather than with non-parametric multiple comparison approaches like Kruskal-Walis ANOVA. See Table II for pairwise significance tests of psychophysical rescales at the 95% confidence level.

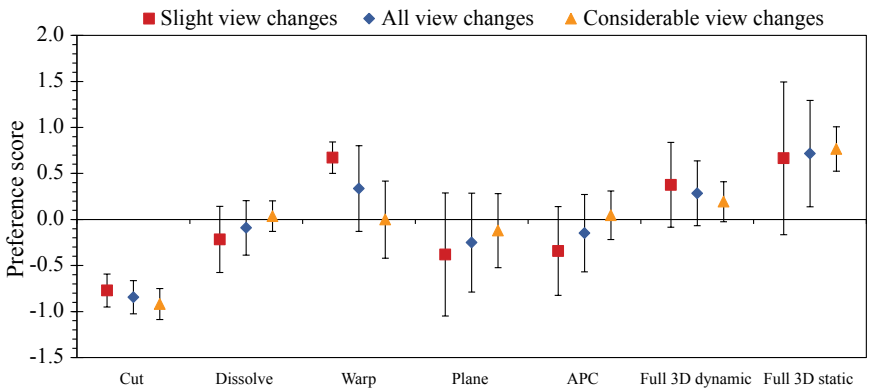


Fig. 4: The result of preference scores for different transition types across all scenes. The scale is in the form of z-score, of which group mean is at 0, the y-value represents multiples of the group standard deviation as discrimination dispersion of the perceived image quality, and higher scores indicates more preference by participants.

Sig.	Cut	Dis.	Warp	Pla.	APC	3D-d	3D-s
Cut		0.00	0.00	0.01	0.00	0.00	0.00
Dis.	0.00		0.06	0.43	0.56	0.06	0.01
Warp	0.00	0.06		0.02	0.08	0.81	0.18
Pla.	0.01	0.43	0.02		0.68	0.07	0.01
APC	0.00	0.56	0.08	0.68		0.03	0.01
3D-d	0.00	0.06	0.81	0.07	0.03		0.03
3D-s	0.00	0.00	0.18	0.01	0.01	0.03	

(a) All scenes (10 sets)

Sig.	Cut	Dis.	Warp	Pla.	APC	3D-d	3D-s
Cut		0.01	0.00	0.27	0.07	0.01	0.03
Dis.	0.01		0.00	0.66	0.48	0.14	0.16
Warp	0.00	0.00		0.01	0.01	0.36	0.99
Pla.	0.27	0.66	0.01		0.93	0.19	0.13
APC	0.07	0.48	0.01	0.93		0.07	0.14
3D-d	0.01	0.14	0.36	0.19	0.07		0.37
3D-s	0.03	0.16	0.99	0.13	0.14	0.37	

(b) Slight view change (5 sets)

Sig.	Cut	Dis.	Warp	Pla.	APC	3D-d	3D-s
Cut		0.00	0.02	0.03	0.00	0.00	0.00
Dis.	0.00		0.85	0.55	0.96	0.31	0.00
Warp	0.02	0.85		0.59	0.88	0.52	0.04
Pla.	0.03	0.55	0.59		0.59	0.26	0.01
APC	0.00	0.96	0.88	0.59		0.07	0.02
3D-d	0.00	0.31	0.52	0.26	0.07		0.02
3D-s	0.00	0.00	0.04	0.01	0.02	0.02	

(c) Considerable view change (5 sets)

Table II. : Pairwise significance tests with the p-values of the preference scores (α = 0.05 each). Green cells denote significantly preferred, and red cells denote significantly less preferred. If column Cut with row APC is red, then Cut is significantly less preferred than APC. If column APC with row Cut is green, then APC is significantly preferred over Cut.

4.2 Results

Figure 4 shows the result of the preference scores across all scenes and view changes, with Tables IIa-IIc showing significance values and whether these cross a positive/negative threshold of p-value < 0.05. Figure 5 plots preference scores into grouped bars. Our perceptual scale variances are computed across scenes only, but after rescaling the participant responses to the scale for fair comparison. Changes in preference scale correspond to percentages of observers as per the standard deviation in a normal distribution. As we only have a small number of scenes, we might have been fortunate to find significances and further experiments to verify these would be necessary. However, our interview responses are consistent with the ranking data and help corroborate our findings.

Comments from the experiment provided by participants are summarized in Tables III and IV. The first table collates comments by transition, noting positive and negative feedback; the second table collates comments by feature/artifact as in Table I.

4.3 Discussion

The results show that there is an overall preference for full 3D static transitions (Figure 4). This is not surprising as the video frames are projected onto actual 3D geometry and this provides the strongest spatial cues of all transitions. From the comments of participants (Table IV), we also know that the 3D transitions not only have smooth camera motion but also provide the effect of being spatially immersed

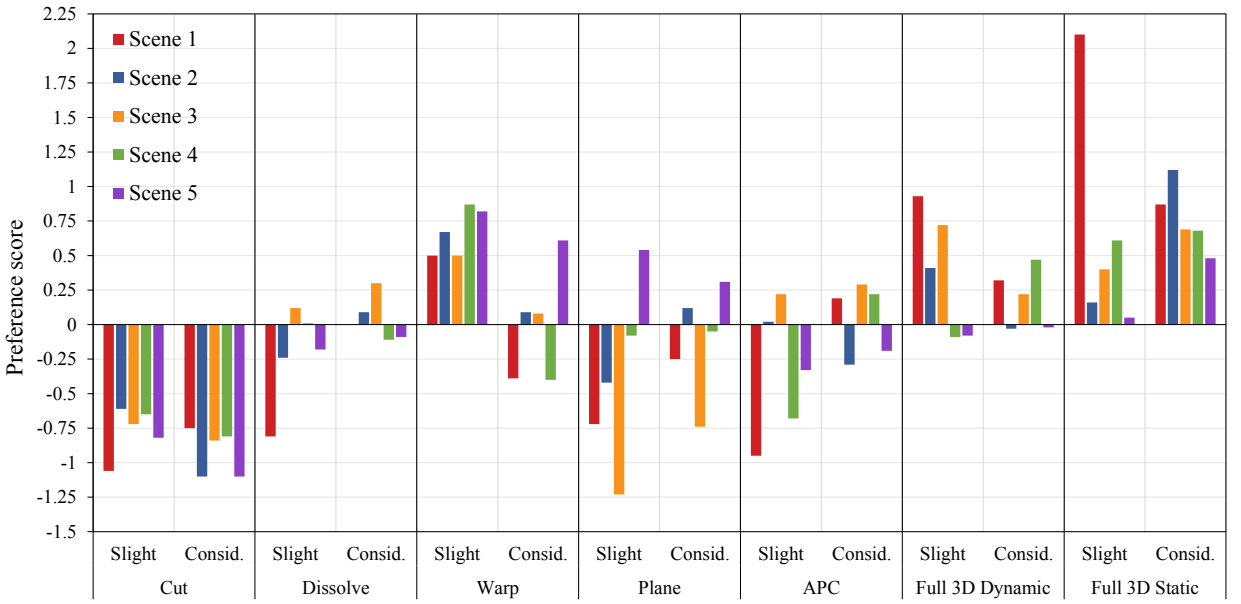


Fig. 5: The result of preference scores for transition types across scenes. Higher scales are better. We can see that cut is disliked, warp is liked for slight cases, full 3D static is generally liked, and that scene-specific artifacts can alter a score significantly.

Transition	# Positive	# Negative	Difference
Cut	6§	2	+4
Dissolve	5§	1	+4
Warp	3	0	+3
Plane	2	1	+1
Ambient Point Clouds	6	6	0
Full 3D Dynamic	5	3	+2
Full 3D Static	20	2	+18

Table III. : Numbers of positive and negative comments from participants for each transition type. §: All of these positive comments referenced when artifacts in other transitions became overwhelming, making the fallback transitions preferred.

in the scene — these features were positively noted most often when compared to other features. Surprisingly, full 3D dynamic transitions where both videos continued playing were preferred less, and this is also reflected in the comments as frozen time was positively noted more often than the presence of dynamic objects. Looking at the per-scene results, we hypothesize that this is due to ghosting which stems from inaccurate camera tracks in the difficult shaky cases.

The warp is preferred for slight view changes, and is significantly better than plane and APC transitions when considering slight view changes only (p-value < 0.05, t-test, Table IIb). While it is not significantly preferred over full 3D transitions, opinion on the warp transition in slight cases was consistent, with a very small variance and the highest mean score of any transition (Figure 4). The static 3D transition is among the top 3 transitions for all sets, and overall is significantly better than all other transitions for considerable view changes (p-value < 0.05, t-test, Table IIc). This justifies the computational cost of reconstructing and rendering such a transition. In almost all cases, cut transitions were significantly unpreferred — only plane and APC slight view cases were similarly so.

	# Liked (# unique participants)	
Feature		
Registered scene	0	(0)
3D effect	6	(5)
Dynamic objects	1	(1)
Smooth virtual camera	5	(5)
Common familiarity	1	(1)
Signifies change of time	0	(0)
Explicit motion cues	2	(2)
Frozen time	3	(2)

	# Disliked (# unique participants)	
Artifact		
Ghosting (static objects)	6	(3)
Ghosting (dynamic objects)	3	(3)
Orientation loss	3	(3)
Bad corresp. swirls	0	(0)
Frame edge flickering	0	(0)
Skewed scene	1	(1)
Pepper noise	3	(2)
Multiple scene elements	2	(1)
Recovered geom. failures	4	(1)
Empty black regions	4	(3)

Table IV. : Collated comments relating to specific features and artifacts identified in Table I.

Beyond these results, it is hard to make strong statements with statistical significance about our scenes and transition types. Rigorously testing for specific features and artifacts, or testing for specific scene objects and effects, would be a much larger experiment and is beyond the scope of this exploratory work. However, it is still worthwhile to discuss these issues based on per-set and per-transition results and on participant comments to deduce as much as possible about transition preference. This is included in supplemental material, and informs the following analysis.

5. OUTCOMES

Many factors may have contributed to the preference of participants, but we find with significance that slight vs. considerable view changes are a key factor. Warp transitions are the perceptually preferred transition type for slight view changes: warps are significantly preferred over all other transitions except the full 3D transitions and, vs. full 3D, warps have a higher perceptual score and a much smaller variance (Figure 4). As such, our results indicate employing warps if the view rotation is slight, i.e., equal to or less than $10°$. Slight view change transitions that have good geometry reconstructions and do not suffer shake (similar to Scene 3, see supplemental video) will also provide high-quality results when using the static or dynamic full 3D transitions. In general, the success of the full 3D transitions is more scene dependent than the warp with the possibility of geometric errors and empty regions caused by matching or conflicting camera motions. The static full 3D transition is significantly preferred for considerable view changes. If video-geometry registration were always accurate then dynamic 3D transitions should be at least competitive, but often small registration errors in the video lead to static ghosting.

Our results also show that a dissolve is preferable to a cut. Should any geometry fail to reconstruct, either from insufficient context or a failure of camera tracking, then it is always preferable to fall back

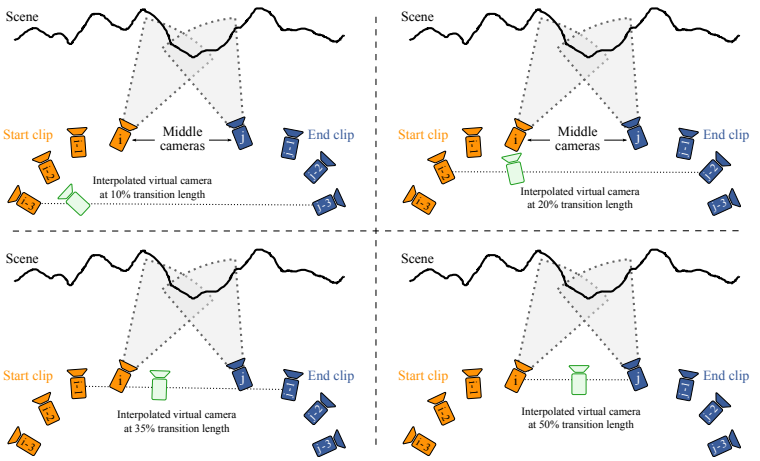


Fig. 6: Progression of interpolated virtual camera pose (green) for start and end clips with contrasting pans, in Western order. Frames i and j are the middle frames of the transition, where we assume visual content similarity. Transition progression beyond frames i and j ($+1$,$+2$,$+3$, etc.) is not shown.

to a dissolve instead of a cut. Further, as there is no significant difference between plane/APC and dissolve transitions, our results suggests that for at least some cases it is not worth the computational effort to perform a video-based rendering transition.

We describe outcomes that were not tested for explicitly, but from the per-transition and per-set analysis merit discussion:

— Participants did not seem to notice or care about the dynamic objects in our scene transitions. If other artifact-causing issues are solved then this might increase in importance, but as it stands it does not appear to be a major factor. This might also change if dynamic objects are the focus of the videos.

— APC works better with considerable view changes and not zooms, that is, large angular view changes or large translations. Slight view changes tend to cause double images as the geometry contrasts with an image formed from a slightly different view within the point cloud.

— Good video registration is imperative but difficult to achieve under camera shake with rolling shutter distortion. Scene 2 (considerable) shows that inaccurate video registration can turn a convincing transition (as full 3D static) into one that is perceptually equivalent to a dissolve (full 3D dynamic).

— Clip content can be a cue as to where a participant expects the camera to be after a transition. In Scene 5 (considerable), the first video pans from a riverbank to view a bridge, then the transition moves the camera to the second clip which was taken on the bridge — first, the bridge is presented as a destination; then, the viewer is taken there. These 'storytelling' cues may affect preference, but their exploration is beyond our scope.

5.1 Camera Motion Effects on Empty Regions

Camera motions in the start and end clips can cause empty areas to appear in the rendered transition due to the difference between the real camera pose and the virtual interpolated camera pose. Figure 6 explains our camera interpolation method, which linearly interpolates positions and spherically linearly interpolates rotations between scene-registered camera poses for every video frame in the transition. Empty regions are a large detrimental factor in transition quality because they break seamlessness. As such, we explain the creation of empty regions in common pan and zoom cases with

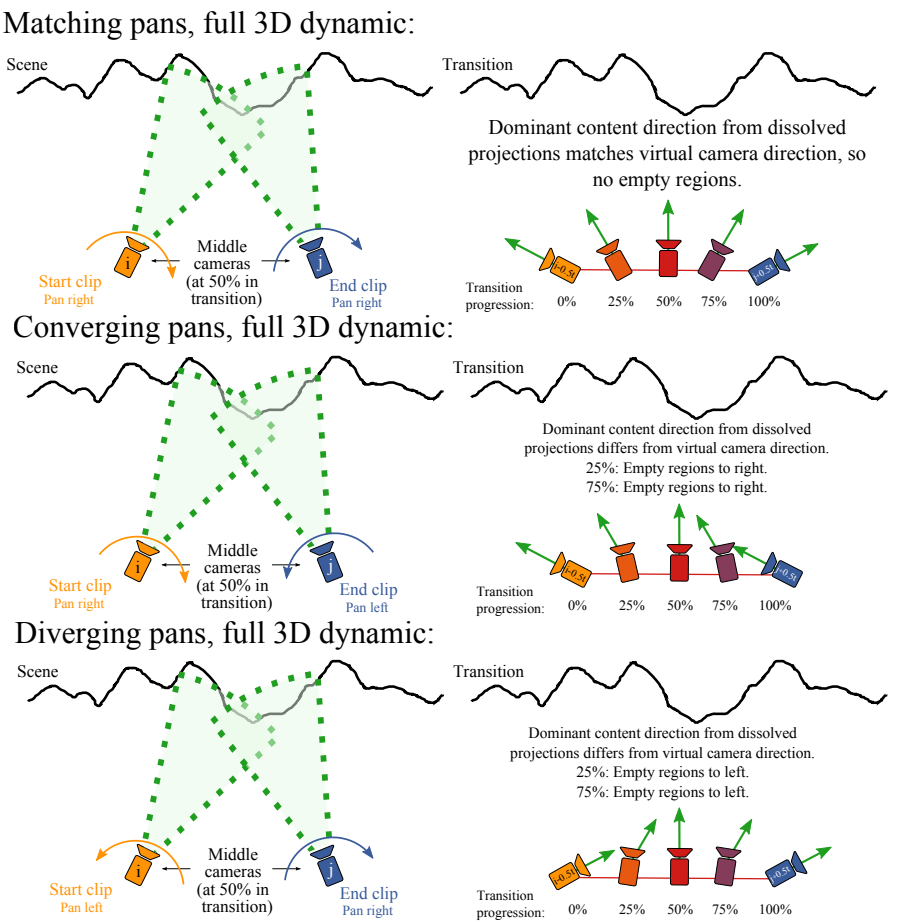


Fig. 7: Horizontal pans in start and end clips affect empty areas in full 3D dynamic transitions. *Right:* Start and end clip camera positions and rotations are interpolated to create the virtual camera (Figure 6). Video content is projected onto the scene, with the *dominant* content direction shown as a green arrow to the *left* (as projections are dissolved across the transition, save the middle of the transition, there is always one video clip which dominates). The difference in angle between the dominant content direction and the virtual camera direction notes the location of empty areas in the view. Relative pan speeds affect the size of the empty area, where larger differences in pan speeds equals larger empty areas.

diagrams. These present ideal results: pans are assumed to move only in the horizontal direction and at equal velocities with no wobble or shake (Figures 7 and 8); zooms are assumed to have constant velocity (Figure 9). Real-world cases are more complicated, but these ideal results can be used to predict areas of empty regions. The diagrams also explain the camera motions which cause full 3D dynamic transitions to be preferred over their static counterparts when other artifacts are not prominent: the static case has larger empty regions (slight cases, Scenes 2 and 3, Figure 5).

Reducing empty regions presents a trade-off: Our current interpolation method provides smooth camera motions that blend between the real camera motions. However, under the conditions outlined in the figures, this leads to empty areas where the dominant video projection in the cross-dissolve is not aligned to the virtual camera motion. One solution is to always strongly weight the camera interpolation factor towards the camera with the dominant projection. This would reduce the size

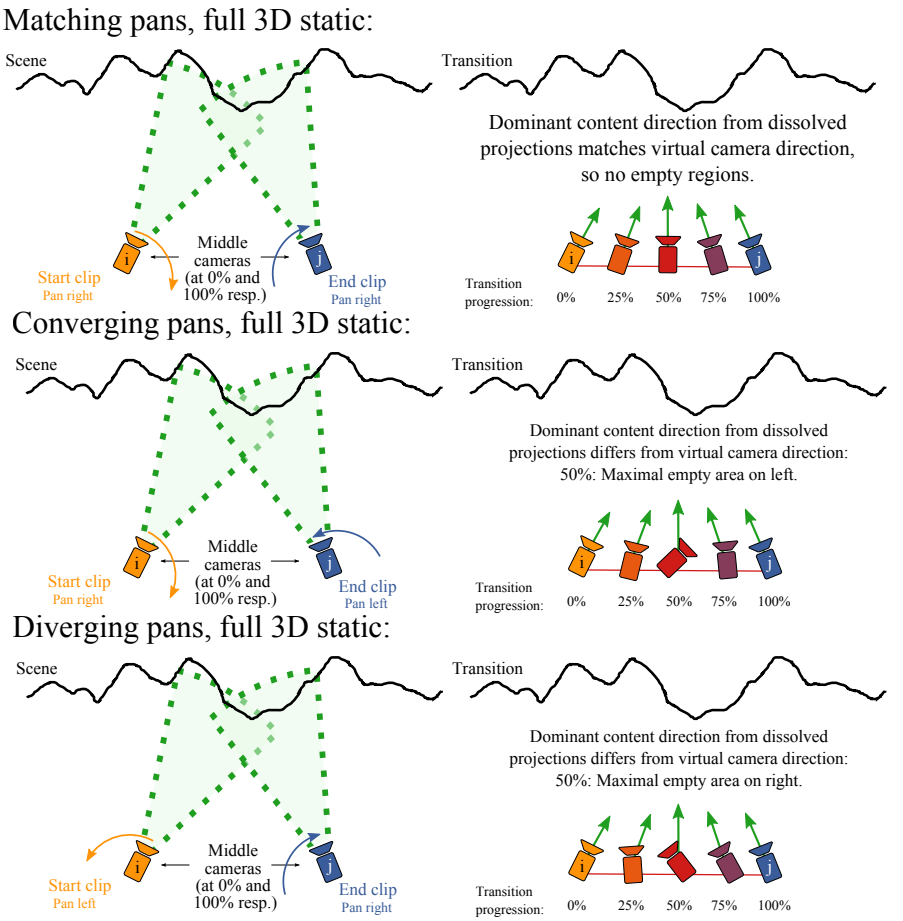


Fig. 8: Horizontal pans in start and end clips affect empty areas in full 3D static transitions. *Right:* Still image content is projected onto the scene from the transition start and end frames, so the *dominant* content direction shown as a green arrow differs from the full 3D dynamic case (Figure 7) and is consistently between frames.

of empty areas by requiring the virtual camera to more closely follow the real camera paths; however, this reduces the smoothness of the virtual camera motion and causes a more abrupt change of direction in the virtual camera. Further, a linear interpolation of real cameras also provides smooth blending between any 'stylistic' motions such as shake (Scenes 4 & 5). In these cases, providing a weighted interpolation would cause a more abrupt change in style. This competition between content is unique to video transitions which involve real and virtual camera motion and, in general, demands future work to explore potential solutions.

5.2 Artifact Ordering

We would like to produce an order of artifacts to resolve their perceptual importance. Our experiment does not directly support the quantitative creation of such an ordering: an experiment which did, across different scene types, is beyond the scope of this exploratory work. Nevertheless, we can present a qualitative ordering for our scenes from our per-transition and per-set analysis (see supplemental material) and corroborate this with the comments of participants.

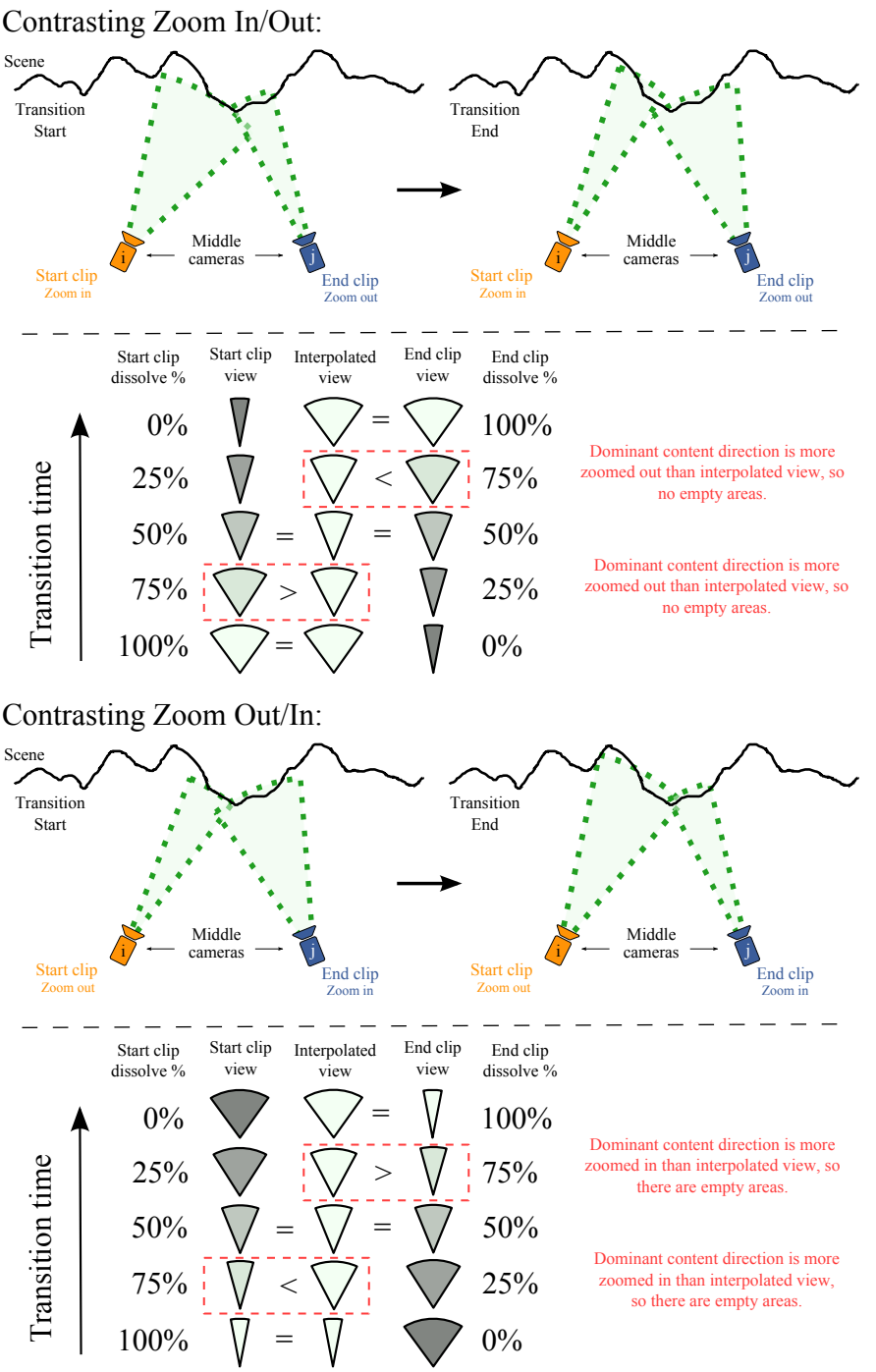


Fig. 9: Zooms in start and end clips affect empty areas. Camera frusta are simplified to circle segments. If the dominant content projection has a larger field of view than the interpolated virtual camera, then there will be no empty areas.

Ghosting on objects which should be static has a large effect on preference. This is exemplified in Scene 2 (considerable), where a large perceptual difference exists between the static and dynamic full 3D transitions which we argue is caused by static ghosting from errors in video registration. We include skewing and geometry errors in this artifact as all have similar causes. Surprising to us was the effect of empty regions. In Scenes 1 and 3, this was the major difference between full 3D static and dynamic transitions, and in each case the transition with smaller empty regions was preferred. APC pepper noise received many negative comments. These artifacts can be reduced by using more points, or by implementing a morphological open post process. Image-plane artifacts in the warp, such as bad correspondence swirls and undulating or flickering parts, seem not to be important in slight view change cases. In considerable view change cases, where the warp is perceptually less preferred and correspondence is harder to find, it is more difficult to judge whether these artifacts have an effect because, relatively, more of the resulting transition image is different.

Given these findings, we order the artifacts. While the evidence supporting this ordering is somewhat anecdotal, it should help direct effort in correcting artifacts in these and other transitions:

$$\begin{aligned}\text{Ghosting on static objects} &\approx \text{empty regions}\\ &> \text{pepper noise}\\ &> \text{swirls} \approx \text{temporal flickering}\\ &> \text{ghosting on dynamic objects.}\end{aligned}$$

Many of these artifacts are caused by failures in computer vision systems to accurately find correspondence or reproduce scene geometry, and these techniques are active research areas.

6. CONCLUSION

We performed an experiment comparing seven video transition types in ten scenes of places over slight and considerable view changes to study which transitions were preferred by participants. We discovered that there was a strong preference for full 3D static transitions in the considerable view change case, and a preference for warp transitions in the slight view change case. We categorized artifacts within the tested transitions, analyzed comments from participants to identify which artifacts caused the most objections, and suggested an ordering of artifact importance. It is clear that in some cases particular transitions are preferred; however, no transition is universally applicable: under difficult conditions, such as shake or incomplete geometry reconstruction, vision-based graphical transitions are preferentially equivalent to dissolves.

Our exploratory work has only a narrow focus and there are many issues to consider such as large dynamic objects, geometry coverage, and other scene categories. One clear direction for future work is how to reduce the effect of video camera motion on empty areas across transitions. We find that pausing in-camera motion can often produce a better result due to reduced artifacts but, with no artifacts, a dynamic video can be just as effective. Finally, camera motions when joined have an implicit semantic meaning, and it may be possible to categorize and exploit this for narrative purposes.

Our complementary website contains all supplemental material and experiment stimuli videos: `http://gvv.mpi-inf.mpg.de/projects/VideoTransitionsOfPlaces/`.

Acknowledgments

Thanks to Christian Kurz, Thorsten Thormählen, and Gunnar Thallin for transition rendering help. Aerial imagery in Figure 1 ©2013 Bluesky, DigitalGlobe, Getmapping plc., Infoterra Ltd. & Bluesky, and The GeoInformation Group.

APPENDIX

We concisely describe how each of the transition types is created; please see the supplemental material for full details. Cut transitions are simply a swap from one video to the other at the required frame. Dissolve transitions pixel-wise linearly interpolate the RGB values of two video clips across the transition length of frames. The other five transition types are more involved.

A.1 Plane

Plane transitions follow the method of Snavely et al. [2006] with adaptions for video. We begin by finding SIFT feature-point correspondences [Lowe 2004] between two anchor frames, one in each video clip, which we assume to have strong visual overlap. From these correspondences, we robustly find a fundamental matrix using RANSAC [Fischler and Bolles 1981] and the eight point algorithm [Hartley and Zisserman 2004]. We estimate the pose of the cameras with bundle adjustment [Lourakis and Argyros 2004], which also produces 3D points for each 2D feature point. A common plane is estimated from the 3D points. This is the best fitting plane in the least-squares sense to the point set observed in both views, and is estimated robustly using RANSAC.

Next, we track each video individually using the Voodoo KLT tracker [Thormählen 2006]. We match SIFT and KLT feature locations and re-optimize a camera pose for each frame of the input videos. This produces a temporally consistent set of pose parameters in the same coordinate space for both video sequences. To render the transition, we project two frames, one from each video, from their respective camera poses onto the plane. As the transition progresses, we linearly interpolate the contributions from each video to dissolve from the start clip to the end clip. The plane is viewed from a virtual camera which is interpolated from the two video camera poses for that timestep (Figure 6).

A.2 Warp

Warp transitions use the same information as in the plane transition: correspondences between anchor frames in each video, and correspondences from each video frame to this anchor frame. The warps themselves are computed using moving least squares image warping [Schaefer et al. 2006]. A warp is computed from the correspondences between anchor frames; each video frame is warped to its corresponding anchor frame; and finally these two warps are accumulated. The transition is created by linearly interpolating these accumulated warps across the transition time.

To improve the 3D effect and to remove some of the minor ghosting due to the sparse nature of the correspondences, we introduce an additional step and use dense optical flow in a similar way to Eisemann et al. [2008]. Flow vectors are computed between the two approximately registered video frames for each timestep. The flow vectors are then interpolated across the transition length such that ghosting on static objects is reduced.

A.3 Full 3D

We begin by recovering whatever geometry possible: first, we estimate camera poses from many views of the scene [Snavely et al. 2006]; next, we compute appropriate multi-view stereo clusters [Furukawa et al. 2010]; then, for each cluster we compute a multi-view stereo point cloud [Furukawa and Ponce 2010]. We compute the union of the point cloud clusters, and form a mesh from the union with Poisson reconstruction [Kazhdan et al. 2006]. We clean the mesh reconstruction by removing faces which are far from any points in the original cloud.

Automatic geometry recovery methods cannot currently recover full scene geometry and certain areas such as the sky will likely always need special treatment. For these regions, we use planes as proxy geometry: we place one sky plane just behind all existing geometry, and one ground plane below all existing geometry. The transition proceeds to project video frames onto this geometry from the

recovered video camera poses (computed as per the plane transition), and dissolve their contribution across the transition time. The only difference between static and dynamic full 3D transitions is that, in the static case, the projection is frozen at the anchor frame.

A.4 Ambient Point Clouds

APC is a recent technique developed by Goesele et al. [Goesele et al. 2010] to fill holes in recovered geometry and provide motion cues during transitions. It starts by computing the minimum and maximum depths of any recovered geometry in the two views between which to transition. For each pixel in each view, APC generates points at random positions between the minimum and maximum depths along the ray through the center of projection of the camera and the pixel. Typically, five points are generated along each ray, with the colour of each point taken from the respective pixel in the image. When the virtual camera interpolates between the two views, the APC is drawn in the empty spaces between the recovered geometry. The points in the cloud splay out in the direction opposite to the camera motion, and so provide strong motion cues to the virtual camera direction of motion.

Points along the ray are perturbed very slightly by random offsets in x and y to reduce aliasing and moiré-like patterns at the beginning and end of the transition (when the point cloud *almost* represents the original image). A plane is rendered at the very beginning and end of the transition to smooth the introduction of the point cloud.

REFERENCES

BALLAN, L., BROSTOW, G. J., PUWEIN, J., AND POLLEFEYS, M. 2010. Unstructured Video-based Rendering: Interactive Exploration of Casually Captured Videos. *ACM Trans. Graph. (Proc. SIGGRAPH) 29,* 4, 1.

BORG, I. AND GROENEN, P. 2010. *Modern Multidimensional Scaling: Theory and Applications*. Springer Series in Statistics.

CHAURASIA, G., SORKINE, O., DRETTAKIS, G., AND INRIA, R. 2011. Silhouette-aware Warping for Image-Based Rendering. In *Eurographics Symposium on Rendering*. Vol. 30.

CUI, C. 2000. Comparison of Two Psychophysical Methods for Image Color Quality Measurement: Paired Comparison and Rank Order. In *Proc. 8th Color Imaging Conference on Color Science and Engineering Systems, Technologies and Applications (CIC-00)*. IS&T, Springfield, Virginia, 222–227.

DEBEVEC, P., YU, Y., AND BORSHUKOV, G. 1998. Efficient View-dependent Image-based Rendering with Projective Texture-mapping. In *Rendering Techniques 98: Proceedings of the Eurographics Workshop*. Number CSD-98-1003. Vienna, Austria, 14.

DMYTRYK, E. 1984. *On Film Editing*.

EISEMANN, M., DECKER, B. D., MAGNOR, M., BEKAERT, P., DE AGUIAR, E., AHMED, N., THEOBALT, C., AND SELLENT, A. 2008. Floating Textures. *Computer Graphics Forum 27,* 2, 409–418.

ENGELDRUM, P. G. 2000. Psychometric Scaling: A Toolkit for Imaging Systems Development. Imcotek Press, Winchester, MA.

FISCHLER, M. A. AND BOLLES, R. C. 1981. Random Sample Consensus: A Paradigm for Model Fitting with Applications to Image Analysis and Automated Cartography. *Communications of the ACM 24,* 6, 381–395.

FURUKAWA, Y., CURLESS, B., SEITZ, S. M., AND SZELISKI, R. 2010. Towards Internet-scale Multi-view Stereo. In *Proc. IEEE CVPR*. 1434–1441.

FURUKAWA, Y. AND PONCE, J. 2010. Accurate, Dense, and Robust Multi-view Stereopsis. *IEEE TPAMI 32,* 8, 1362–1376.

GOESELE, M., ACKERMANN, J., FUHRMANN, S., HAUBOLD, C., AND KLOWSKY, R. 2010. Ambient Point Clouds for View Interpolation. *ACM Trans. Graph. (TOG) 29,* 4, 1–6.

HARTLEY, R. AND ZISSERMAN, A. 2004. *Multiple View Geometry in Computer Vision* Second Ed. Cambridge University Press, ISBN: 0521540518.

HORRY, Y., ANJYO, K.-I. A., AND ARAI, K. 1997. Tour Into The Picture: Using a Spidery Mesh Interface to make Animation from a Single Image. In *Proc. SIGGRAPH*. 225–232.

KAZHDAN, M., BOLITHO, M., AND HOPPE, H. 2006. Poisson Surface Reconstruction. In *Proc. Eurographics Symposium on Geometry Processing*. Eurographics Association, New York, NY, USA, 61–70.

LIPSKI, C., LINZ, C., NEUMANN, T., WACKER, M., AND MAGNOR, M. 2010. High Resolution Image Correspondences for Video Post-Production. In *Proc. European Conference on Visual Media Production (CVMP)*. IEEE, 33–39.

LOURAKIS, M. I. A. AND ARGYROS, A. A. 2004. The Design and Implementation of a Generic Sparse Bundle Adjustment Software Package based on the Levenberg-Marquardt Algorithm. *ICSFORTH Technical Report TR 340,* 340.

LOWE, D. G. 2004. Distinctive Image Features from Scale-Invariant Keypoints. *International Journal of Computer Vision 60,* 2, 91–110.

MCCURDY, N. J. 2007. RealityFlythrough: A System for Ubiquitous Video. Ph.D. thesis, University of California, San Diego.

MORVAN, Y. AND O'SULLIVAN, C. 2009. A Perceptual Approach to Trimming and Tuning Unstructured Lumigraphs. *ACM Trans. Appl. Percept. 5,* 4, 19:1–19:24.

MORVAN, Y. AND O'SULLIVAN, C. 2009. Handling Occluders in Transitions from Panoramic Images: A Perceptual Study. *ACM Trans. Appl. Percept. 6,* 4, 1–15.

MURCH, W. 2001. *In The Blink Of An Eye*. Silman-James Press.

MUSTAFA, M., GUTHE, S., AND MAGNOR, M. 2012. Single-trial EEG Classification of Artifacts in Videos. *ACM Trans. Appl. Percept. 9,* 3, 12:1–12:15.

OH, B. M., CHEN, M., DORSEY, J., AND DURAND, F. 2001. Image-based Modeling and Photo Editing. In *Proc. SIGGRAPH*. ACM Press, New York, NY, USA, 433–442.

SCHAEFER, S., MCPHAIL, T., AND WARREN, J. 2006. Image Deformation using Moving Least Squares. *ACM Transactions on Graphics 25,* 3, 533.

SNAVELY, N., GARG, R., SEITZ, S. M., AND SZELISKI, R. 2008. Finding Paths Through the World's Photos. *ACM Trans. Graph. (Proc. SIGGRAPH) 27,* 3, 1.

SNAVELY, N., SEITZ, S. M., AND SZELISKI, R. 2006. Photo Tourism: Exploring Photo Collections in 3D. In *ACM Trans. Graph.* Vol. 25. ACM, 835–846.

STICH, T., LINZ, C., WALLRAVEN, C., CUNNINGHAM, D., AND MAGNOR, M. 2011. Perception-motivated interpolation of image sequences. *ACM Trans. Appl. Percept. 8,* 2, 11:1–11:25.

THORMÄHLEN, T. 2006. Zuverlässige Schätzung der Kamerabewegung aus einer Bildfolge. Ph.D. thesis, Universität Hannover.

TOMPKIN, J., KIM, K. I., KAUTZ, J., AND THEOBALT, C. 2012. Videoscapes: Exploring Sparse, Unstructured Video Collections. *ACM Trans. Graph. (Proc. SIGGRAPH) 31,* 4.

TORGERSON, W. S. 1958. *Theory and Methods of Scaling*. Wiley, New York, NY, USA.

VANGORP, P., CHAURASIA, G., LAFFONT, P.-Y., FLEMING, R. W., AND DRETTAKIS, G. 2011. Perception of Visual Artifacts in Image-based Rendering of Façades. In *Eurographics Symposium on Rendering*. Vol. 30.

VANGORP, P., RICHARDT, C., COOPER, E. A., CHAURASIA, G., BANKS, M. S., AND DRETTAKIS, G. 2013. Perception of Perspective Distortions in Image-Based Rendering. *ACM Trans. Graph. (Proc. SIGGRAPH) 32,* 4.