Interactively Stylizing Camera Motion

Neel Joshi, Dan Morris, and Michael F. Cohen

Microsoft Research
{neel, dan, mcohen}@microsoft.com

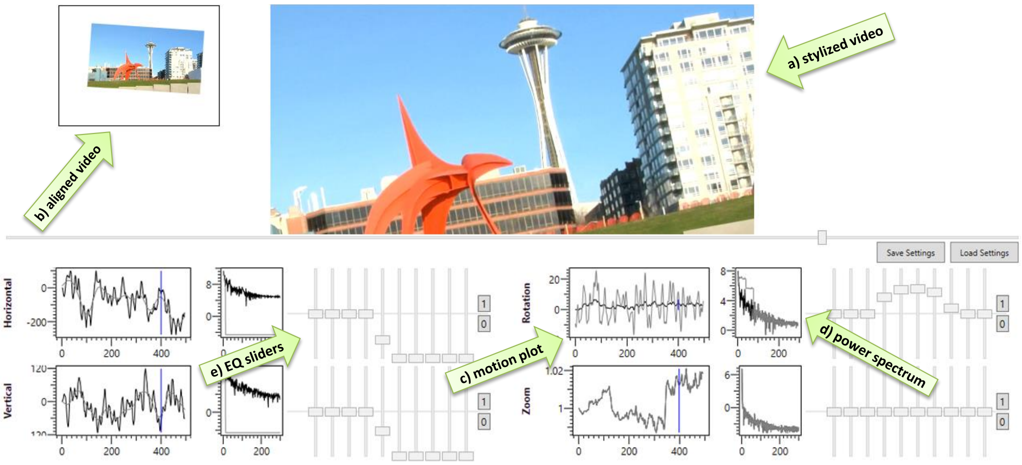

Figure 1. A user can modify camera motion style in real time by manipulating sliders which modify the power spectra for each camera motion parameter. The UI shows a) the input video with the current stylized motion applied and b) the video aligned to the global canvas. There are sub-panels for the motion parameters: x and y translation, rotation, and scale. Each sub-panel has c) a plot of the original and stylized motion and d) power spectrum and e) a 10-band equalizer (EQ). In the motion plot, the bold line is the original path and the finer gray line is the stylized path. Here we have dampened the mid and low frequencies for the x and y motion and amplified the mid frequencies for rotation.

ABSTRACT

Movie directors and cinematographers impart style onto video using techniques that are learned through years of experience: camera movement, framing, color, lighting, etc. Without this experience and expensive equipment, it is very difficult to control stylistic aspects of a video. We introduce a novel approach for post-hoc editing of one specific aspect of cinematography – camera motion style – via an equalizer-like set of controls that manipulates the power spectra of a video's apparent motion path. We explore free manipulation of apparent camera motion as well as the transfer of motion styles from an example video to a new video to create a wide range of stylistic variations. We report on a user study confirming the ability of non-expert users to create motion styles.

Author Keywords

Camera motion editing; video stylization

ACM Classification Keywords

H.5.2 [Information Interfaces and Presentation]: User Interfaces - Interaction styles;

INTRODUCTION

Our emotional responses to a film are guided by a composition of acting performances, staging, sound effects, scoring, and cinematography.

CHI 2014, April 26–May 1, 2014, Toronto, Ontario, Canada.

ACM 978-1-4503-2473-1/14/04 ...$15.00.
http://dx.doi.org/10.1145/2556288.2556966

One of the more visceral aspects of a film is *camera movement*. Camera movement is a powerful tool used by filmmakers to establish pace, point of view, or rhythm in a scene – it can be used to draw the viewer into the action or to convey a disconnect between audience and characters. Camera motion, as a stylistic choice, is often so powerful that it can be the primary memory of a film or video – movies such as "The Blair Witch Project" and "Cloverfield" evoke feelings of tension and chaos, while Alfred Hitchcock's "The Rope" is remembered for the "continuous take", and "The West Wing" for its long "walk-and-talk" shots.

While camera movement is well-understood, its artistic control is challenging. This ability is typically restricted to professional cinematographers and directors, because it requires experience, careful planning, and often costly equipment (Steadicams, tracks, etc.). For example, the director of "Cloverfield" created the movie's signature shakes and swoops with a custom rig that would be prohibitively expensive for amateur videographers. Furthermore, even if one is able to film with a desired style, there are few tools for changing the apparent motion *after* the fact without reshooting.

Video stabilization is a well-known methodology that manipulates a moving crop window to remove much of the apparent motion of the camera. We extend the stabilization paradigm to impart many other motion styles. Specifically, we present an application that allows one to adjust the relative frequency characteristics of the camera motion using a paradigm similar to graphic equalization (EQ) in audio editing. Just as an EQ on a mixing panel can be used to change the prominence of certain tones to balance instruments or give a song a "concert feel" or "studio sound", our tool can amplify or dampen the frequency components of camera motion to change its perceived style. Our contributions include 1) an interactive tool for editing apparent camera motion by manipulating frequency content and 2) a method for automatically setting the editing parameters from another video clip.

To the best of our knowledge, this is the first time that an equalization-like interface has been used for controlling camera motion. We demonstrate the creation of a variety of camera motion styles using our interactive app. We also report on a user study in which participants were asked to create four specific styles of motion. Most users were able to quickly impart the requested styles, and the quality of their created styles was confirmed by a separate set of scorers who correctly labeled the resulting videos with the four styles.

PREVIOUS WORK

Video stabilization is the removal of undesirable high-frequency motion, often resulting from the instability of handheld cameras. This is a special case of motion stylization that has received significant attention [2]. Current stabilization systems minimize user interaction by design and do not generalize to other forms of motion stylization. We leverage previous work in stabilization to re-synthesize a video given a modified camera path, but we provide novel, interactive approaches for generating that camera path.

The notion of interactive motion stylization has been addressed in other areas of computer graphics with similar goals: to enhance a non-expert's creative control over a video artifact. However, that work primarily addresses stylization of human motion animation and does not address video camera motion. Neff et al. [4], for example, provide an intuitive parameterization of kinematic chains for interactive motion editing in animation. Our work aims to provide a similar parameterization of *camera paths* for video motion.

The transfer of styles from one artifact to another has also been addressed in the computer animation literature: Hsu et al. [3] present an approach to learning and transferring stylistic components of an animation sequence to a new animation, without disrupting the non-stylistic content of the target. This in part inspired our "Stylizing Camera Motion by Example" approach to capturing the stylistic dynamics of an existing video. This aspect of our system also draws on the audio domain, where the multi-band equalizer is common and has been exploited for stylistic transfer (e.g., the Match EQ feature of Apple's Logic Studio[1], which transfers equalizer settings from a reference recording).

STYLIZATION FRAMEWORK

The goal of our work is to modify stylistic elements of camera motion interactively. It has two components: 1) inferring camera motion from a video clip and 2) modifying the motion in a stylistically meaningful way. The first step is essential for stylizing camera motion, however it is not our focus, and we use existing methods. Our contributions address the second component: stylistically meaningful motion editing.

Modeling Camera Motion

We build on video stabilization methods for recovering camera motion. Though in reality the camera may have moved and rotated in three dimensions, we use 2D transformations (translation, rotation, and scale) for modeling the camera motion. This simplification assumes that the effect of camera motion on the scene can be modeled as a time-varying set of rigid image transformations. It does not model depth or perspective changes. As was noted in [2], despite these assumptions, the model works quite well for most videos.

As our goal is to interactively manipulate stylistic elements of camera motion, we have the additional constraint that the parameters of the motion model should be concise and understandable to a user. Thus we use a *similarity* motion model, which models camera motion as a time-varying set of transformations, S_t, that decomposes to four values: $[x_t, y_t, \theta_t, s_t]$, representing x and y translation, in-plane rotation, and global image scale. In cinematographic terms, these map to pan, roll, and zoom (or forward/backward dolly).

Recovering the apparent camera motion from a video amounts to inferring the sequence of transformation for each frame that best maps that frame to a base frame. The alignment is computed by extracting image features for each frame and performing a search between frames to find matching features. A feature is determined to be a match if the descriptor distance of the best match is sufficiently different from that of the second-best match (a.k.a. the ratio test). To avoid locking onto scene motion, the tracks are analyzed to distinguish foreground motion from background static features by using a RANSAC (RANdom SAmple Consensus) method to find the largest set of inlier tracks such that a single temporal sequence of similarity transforms can map all background features to their positions in the base frame. The transforms are then decomposed into x and y translation, rotation, and scale.

Modifying Camera Motion

The next section, the focus of our work, will address generating a new, stylized camera motion path based on our interface. First, we briefly discuss how we will apply that new path to render a stylized video. We assume that the stylization engine has output a desired motion path $S'[t, 0]$, which can be decomposed into $[x'_t, y'_t, \theta'_t, s'_t]$.

Given $S[t, 0]$ as the transform in the original video between any frame (t) and the first frame (0), represented by $[x_t, y_t, \theta_t, s_t]$, and the modified transform $S'[t, 0]$ from the stylized motion path $[x'_t, y'_t, \theta'_t, s'_t]$, a frame in the stylized sequence can be computed by warping the frame based on the difference between $S[t, 0]$ and $S'[t, 0]$. This is equivalent to warping the image to the base (first) frame's coordinate system and then applying the inverse transformation to map to the stylized sequence.[2]

Any transformation other than the identity will create unknown regions around the video border. The final result is thus cropped to eliminate the unknown regions (we use a fixed crop to 80% of the original video size). This also sets bounds on how much the stylized motion can vary from the original.

Stylizing Camera Motion Interactively

[1]http://apple.com/logicpro

[2]Any frame can serve as the reference frame; we use the first frame.

The question of how to modify and/or transfer camera motion style is really a question of what aspects of camera motion are "story vs. atmosphere" or "substance vs. style". Editing operations, such as stabilization, do affect camera movement style in some sense, but provide no artistic control for achieving a particular style.

While there is an informal understanding in cinematography of what aspects of camera motion are stylistic, there is no mathematical description of camera motion style that we are aware of. Thus, drawing on inspiration from audio and image editing, we use a frequency-based equalization approach to edit camera motion style. Our approach is analogous to using a graphic equalizer in audio editing. Just as on a mixing panel an EQ can be used to change the prominence of certain tones or instruments to change the "feel" of a song, our camera motion equalizer is an interactive method to amplify, dampen, or transfer the frequency components of camera motion to create a desired look and feel.

Just as an audio equalizer can operate independently on the left and right channels of a stereo audio signal, we break our motion signal into the four channels of x (horizontal) translation, y (vertical) translation, in-plane rotation θ, and scale s. For each channel, we compute a frequency-space representation using an FFT (Fast-Fourier Transform) and then multiply or add power to bins in this frequency representation as a function of user-driven sliders corresponding to frequency bands. An inverse FFT is then used to generate the output motion paths. Our bins are logarithmically spaced in frequency, with the bin size increasing as the frequency increases; this is also common practice with audio equalizers. Our EQ does not modify the DC component of the signal; in other words, we aim to manipulate the style of movement, not to completely re-position the camera post-hoc. Just as in audio, where other tools are used to change the fundamental notes or melody, our work would complement other tools for changing the fundamental camera motion path, such as motion transformations and key-framing in Adobe Premiere[3].

The user-provided values are between 0 and 2 for each bin, where a value of 1 returns the original motion. From 0 to 1, we treat the value as a simple multiplier in the frequency domain. Thus these values dampen frequencies in the original signal. For example, setting all values to zero creates a stabilized video. Values above 1 result in an additive operation instead of a multiplicative operation. The system switches from multiplicative to additive because multiplication would have little to no effect when the original magnitude of a frequency in the motion path is at or near zero. By adding to the magnitude, we can add frequency content that was not originally present. This allows us to stylize stationary videos, such as those filmed on a tripod, in addition to hand-held videos.

Stylizing Camera Motion by Example

In some cases, a user may have another example video whose style they want to match. For this case, we provide an automated approach to set the EQ sliders from an example. The user loads the example into the application, and we calculate the EQ values that will scale or add (as appropriate) to the power in each band so the input video has the same average power in each band as is present in the example.

[3]http://www.adobe.com/products/premiere.html

INTERACTIVE SYSTEM

The app (Figure 1) is designed to be simple yet expressive to help users quickly change camera motion style. It reflects three primary design decisions: 1) camera motion is parameterized by x and y translation, in-plane rotation, and scale; 2) stylization of motion is done by manipulating a multi-band equalizer independently for each motion parameter; 3) edits occur in real time and the effect is seen immediately.

The main panel shows the input video with the current stylized motion applied. When a video is loaded, the equalizer sliders are set to 1; thus when the video is first played, the video motion appears unchanged. In the top-left, we show the video aligned to the global canvas, to show how the current frame relates to the others in space.

The bottom panel has four sub-panels for each motion parameter (x, y, θ, s). Each panel has a plot of the original and stylized motion, a plot of the original and stylized power spectrum, and a 10-band equalizer (Figure 1). The equalizer control has shortcut buttons to set all sliders to "1", which yields the original, unmodified motion, and "0", for no motion at all. The effects are applied in real time, at 30 fps, when running on a 2.67 GHz PC.

USER STUDY

We conducted a preliminary study to assess both the efficacy and usability of our system for modifying apparent camera motion. Ten participants (5 female and 5 male) used the system to modify existing videos. Participants were first asked whether they have used video editing software "never" (n=1), "a few times" (n=6), or "regularly" (n=3). Participants were also asked whether they have used graphic equalizers (as found in most music player apps) "never" (n=3), "a few times" (n=5), or "regularly" (n=2). Thus, although participants all had some technical background, the range of familiarity with video editing and with equalizers varied widely.

A session began with approximately three minutes of training. We then presented each participant with one of six 10- to 20-second videos to modify. They were asked to create four modified videos that exhibited each of four styles: (1) Bouncy: as if the videographer had a bounce in their step, (2) Earthquake: as if the videographer was filming during an earthquake, (3) Boat: as if the videographer was on a boat in stormy seas, and (4) Steadicam: as if the camera was moving in a controlled, stable way. Participants were given 12 minutes total to complete the four videos and were allowed to complete the styles in any order and revisit them as desired.

In a second 12-minute task, participants were given a new video and were asked to create four versions, but this time they were asked to *match* four stylized "target" versions of the video. As before, participants could work in any order and revisit their styles or watch the target videos as desired.

The studies aimed to validate two hypotheses: (1) non-expert users can create multiple styles that exhibit the meaning of descriptive labels, and (2) non-expert users can re-create styles based on visual examples.

In a second phase of the study, we recruited seven scorers to match the four videos created by the each of the initial participants in their first task (40 videos total) to the text labels. Each scorer was presented with a panel of four videos, randomly arranged into a 2x2 grid, representing the four videos from one of the participants. They were asked to assign the four labels (Bouncy, Earthquake, Boat, Steadicam) to the videos. They then went on to four more videos from another participant and continued this task until all ten sets had been labeled. Thus we had seven labels for each of the 40 videos.

User Study Results

We use the labels assigned by our scorers to assess whether the study participants were able to create meaningful styles. Given that sets of four videos were presented simultaneously, the probability of randomly getting all 4 in a set correct is 1 in 24, or 4.17%. Of the 70 sets labeled, 53 (or 76%) were all labeled correctly. One can also assess accuracy at the individual-video level; the odds of labeling any video correctly on its own is 1 in 4, or 25%, and the scorers correctly identified 86% of the videos. These results support hypothesis (1): non-experts were able to create meaningful styles using our tool.

In the second task, participants matched a pre-recorded video created with our system. In this case, we automatically classify the sets by numerically comparing the power spectra of the pre-recorded videos with those created by our participants. For example, if user video U1 is matched to target video T1, we compute the L_2 error between the power spectra of U1 and T1. Four user-generated videos are matched to the set of target videos to minimize the sum of L_2 errors across the four matches. Even this simple approach results in 8 in 10 (80%) of the participants' sets being labeled correctly, vs. a 4.17% random chance. Looking individually at each video in the automatically labeled sets, 90% are correctly labeled, vs. a 25% random chance. This supports the hypothesis (2): non-expert users can re-create styles based on visual examples using our tool.

General Study Observations

Based on survey comments, the second task was somewhat frustrating for some. Although most converged quickly to a video similar to the target, many were still not satisfied. It became clear there were two different mental models employed. For some, who adopted a "frequency space" view, getting the stylistic aspects correct was enough. Others, who were more frustrated, were trying to also match the "phase", i.e., they wanted the apparent camera motion to go up or down exactly when the target video's motion went up or down. This is clearly a more difficult task and not as well-supported by our tool. As our primary goal is to allow exploration of motion "style" and not perform the matching problem, we do not see this as a flaw, but it does inform the way one would communicate the equalization metaphor to a broad audience.

Another issue relates to the ease of immediately understanding changes in the higher-frequency vs. lower-frequency sliders. Not surprisingly, changing a high-frequency slider results in an almost immediate response. However, changes in low-frequency sliders are reflected with some inherent latency; for example, it takes at least a second to see the addition of 0.5Hz sinusoidal motion to the video, and in practice the perceived latency is higher, as it is compounded with other low-frequency motion. This led to what we observed to be some users "thrashing" when setting the low-frequency values. Here, the analogy to audio EQ breaks down, since even low-frequency changes are immediately noticeable (the lowest relevant frequency for audio EQ is typically 20Hz).

Yet, even with some latency, the effect is still much more immediately perceived than with current alternatives, and we believe users could master the interface quickly via real-time trial and error. However, further study is warranted. Admittedly, the number of sliders we used (40) may be intimidating to some users and perhaps fewer sliders would work just as well. We focused on evaluating the overall efficacy of our approach; however, there are open questions such as how many sliders are ideal and how efficient are users with these sliders.

DISCUSSION

We have used our work to create many different styles as seen in our supplementary video. The video also includes results from our method to automatically set the equalizer controls using example videos and some results from our user study.

We have presented an interface for manipulating apparent camera motion style based on an *equalizer* (EQ) interface inspired by those used to balance audio frequency components. The EQ values can be set manually or automatically. We believe having a familiar metaphor for doing *anything* to camera motion is novel and exciting, and our approach is easier and faster than the current approach of key-framing individual motion changes. For example, consider creating our shaky "Earthquake" style for a 20-second video. The user would need to create a key frame for each change in direction, so vertical shaking alone at 10 Hz would require creating 400 key frames. Our system requires moving just one slider.

For future work, it would be interesting to consider other interfaces for global editing of camera motion, e.g., "direct manipulation" [1] approaches could be be used for camera motion editing, as they would allow a quick, broad specification of a motion path. It would be interesting to combine our method with such an approach or with key-framing interfaces.

REFERENCES

1. Dragicevic, P., Ramos, G., Bibliowitcz, J., Nowrouzezahrai, D., Balakrishnan, R., and Singh, K. Video browsing by direct manipulation. In *Proceedings of CHI '08* (2008), 237–246.
2. Grundmann, M., Kwatra, V., and Essa, I. Auto-directed video stabilization with robust l1 optimal camera paths. In *Proc IEEE CVPR 2011* (2011).
3. Hsu, E., Pulli, K., and Popović, J. Style translation for human motion. *ACM Trans. Graph. 24*, 3 (July 2005), 1082–1089.
4. Neff, M., and Kim, Y. Interactive editing of motion style using drives and correlations. In *Proceedings of the 2009 ACM SIGGRAPH/Eurographics SCA* (2009).