Can Motion Features Inform Video Aesthetic Preferences?

Scott Chung
Jonathan Sammartino
Jiamin Bai
Brian A. Barsky

Electrical Engineering and Computer Sciences
University of California at Berkeley

Technical Report No. UCB/EECS-2012-172
http://www.eecs.berkeley.edu/Pubs/TechRpts/2012/EECS-2012-172.html

June 29, 2012

Can Motion Features Inform Video Aesthetic Preferences?

Scott Chung Jonathan Sammartino Jiamin Bai Brian A. Barsky

Computer Science Division
Electrical Engineering and Computer Sciences Department
University of California
Berkeley, CA 94720-1776

June 28, 2012

Abstract

We explore a novel approach to evaluate aesthetic qualities of video clips by analyzing key motion features. With a growing number of videos and more interest in sorting through media, evaluating aesthetics is increasingly important. Our approach examines high-level motion-based features in video that account for distribution, trajectory, fluidity, completeness, as well as smooth camera and subject motion. We introduce six fundamental motion features which are derived from basic guidelines in film and then conduct a comprehensive viewer study to examine the correlations between the features and viewer preference.

1 Introduction

Relevance-based searching of video through video analysis (or in general, any media) is difficult. Searching for videos with aesthetics that contribute to the overall satisfaction poses another question. There is a wide range in the quality of photographers/videographers, from amateurs to semi-professionals to professionals. The same scene shot by different videographers can appear quite different due to variations in camera angles, camera and subject movements, lighting, focus, zooming, etc. The aesthetic differences from these parameters has been studied qualitatively by videographers and a body of knowledge exists among professional videographers for shooting high quality videos; however, this knowledge remains subjective.

Measuring visual aesthetics is challenging because differences can be due to variations in the quality of videographers, cameras, lenses and the shooting environment. There are also other parameters such as frame rate, resolution, color balance, camera's post processing, etc. that affect people's preferences [15] To study the *visual* aesthetic preferences of humans, we must control as many of these extraneous parameters as

possible, to focus on only the visual information using computer vision. To the best of our knowledge there is currently no work that does this and no data set exists that controls for these variables. To address these shortcomings, we assembled a data set of video sequences - all of which are shot by the same camera model at the same resolution and frame rate (thereby also removing the effects of different post processing, sensor color balance etc). Since our work focuses on helping amateur videographers assess the aesthetics of their videos, we used a portable handheld camera with a fixed lens and we did not do any additional processing of the videos (other than suppression of audio and downsizing). The videos were also of the same length. More details on the data set are presented in Section 4.

Visually, a significant amount of information in videos in conveyed through motion. Optical Flow is the most important metric for measuring local motion in video sequences. In recent years, there has been a strong interest in the development and implementation of highly accurate optical flow algorithms on fast Graphics Processing Units (GPUs) [17, 21, 19] that has increased their applicability on large video data sets. We predominantly use features derived from optical flow for our aesthetics analysis. Some of the features used include spatial and temporal motion distribution, subject trajectory, completeness of motion, and smoothness of both camera and subject motion. Detailed justification for the features used and their particulars are described in Section 3.

Our database consisted of 120 videos arranged into 30 sets with four videos each. For each set, we recorded the same scene with four different viewpoints and motion. We asked participates of Amazon Mechanical Turk to rate the videos in the order of aesthetic preference. For each scene, we filmed the scene with four different viewpoints and motion. The users' task was to rank the videos in each set in order of aesthetic appeal. We use these rankings as ground truth to validate our data.

2 Previous Work

Although there has been interest in aesthetics evaluation for photographs, there has been very little such research for videos. Photographic aesthetic evaluation began with extracting simple low-level features. In [18], low-level features such as color, energy, texture, and shape were extracted and used to build a classifier that could sort photographs according to low or high aesthetic appeal. The work of [5, 11, 9] introduced the notion of high-level features which incorporates photographic rules of thumbs like color harmony, rule-of-thirds, shallow depth-of-field, region composition, simplicity, and image familiarity.

An algorithm that evaluates the aesthetics of videos was presented in [11]. The authors used the percentage of shaky frames and the average moving distance of the (in focus) subject for the clip as low-level features that represent motions for the classifier. The features are calculated for 1 frame per second and are averaged across the entire length of the video. A subsequent effort by Moorthy et al. [15] hypothesizes that the ratio of motion magnitudes between the subject and the background has a direct impact on video aesthetic appeal. To that end, they compute optical flow and segregate the motion vectors into two classes using k-means. The mode of each class is selected

as the representative vector. They compute two features: the ratio of the magnitude of the representative vectors and the ratio of the number of pixels in each class. The features are computed for each frame and they are combined every second.

Although Moorthy et al. [15] provides a good start in terms of evaluating aesthetics in videos, their use of motion features was severely hampered by their choice of dataset. They used a variety of videos from Youtube with widely varying resolutions, frame rate etc. Such extreme variability meant that features like frame rate, resolution, etc. provide most of the aesthetic information. In fact, Moorthy et al. [15] show that almost no motion features were needed for their best performing classifier (after feature selection to prevent over-fitting).

Understanding video aesthetics involves understanding the role of motion in aesthetics. This has been subjectively studied and several rules of thumb exist; such as:

- Composition in videos changes from frame to frame and thus should be considered as a whole with moving elements and dynamic on-screen forces rather than on a per-frame basis
- Directionality of an object on screen is determined primarily by the direction in which it is moving; that is, the motion vector provides a better indication of directionality than shape, color, etc.
- It is generally desirable to have a good amount of headroom to make the frame seem "balanced"
- The foreground object should be kept in the bottom one-third or two-thirds of the screen for best presentation
- Viewers tend to look at a screen from left to right and pay more attention to objects on the right than on the left
- Objects should have lead room in the direction of their motion vectors (e.g., an object moving to the right should start on the left)
- Objects that occupy a large portion of the frame should have small changes in their velocity and should move slower than small objects
- Motion should be temporally distributed throughout the entire shot, not just limited to a few frames
- Camera motion in itself is distracting and should be avoided unless it is used for a specific purpose like following action, revealing action, revealing landscapes, relating actions, or inducing action

To implement these rules as feature vectors, we need objective measurements of motion in videos. Optical flow is the classic way to measure short term motion. In recent years, many techniques have been proposed to improve accuracy and decrease runtime costs through implementations on GPUs [17, 21, 19]. In particular, Large Displacement Optical Flow (LDOF) [4, 17] provides good quality flows for real world videos and can be run efficiently on GPUs. LDOF combines both feature matching and optical flow in a single mathematical setting that captures both large and small motion efficiently.

2.1 Shortcomings

There are two shortcomings in evaluations of video aesthetics, both of which will be addressed. The first is the absence of a dataset which controls for extraneous parameters like camera model, quality, environment etc., so that analysis can be performed on a truly visual basis. Secondly, although it is widely believed that motion is a key parameter for determining video aesthetics [1, 3, 2, 7], there is little quantitative evidence for this. We address both shortcomings in this paper by proposing our own dataset. We also objectively measure the benefits of using motion features on this dataset using high quality optical flow.

3 Algorithm

Motion plays a large role in evaluating aesthetics of a video. Spatial and temporal motion distribution, subject trajectory, completeness of motion, and smoothness of both camera and subject motion are features that play a large role in perceiving aesthetically pleasing videos. These proposed features of video can be broadly classified into two families: per frame features and entire video sequence features. Frame features are those that can be computed independently for each pair of frames, such as the motion in headroom feature. We design a feature that quantifies the absence or presence of motion in the top one third of the video. This information requires just two neighboring frames. Video sequence features such as characteristic of motion requires the analysis of the subject motion across all frames within the video. We will now introduce our proposed features.

Frame analysis features:

- Motion in headroom
- Directional of motion
- Magnitude of motion versus subject size

Video clip analysis features:

- Characteristic of subject motion
- Characteristic of camera motion
- Locality of motion

Frame analysis features are computed at every frame. They provide an independent measure of motion within the frame and can be aggregated over a number of frames. Video clip analysis features are computed over the entire sequence of frames that constitute the video clip. These features provide a measure of motion throughout video clip.

It should be noted that these features are designed to describe motion in videos. We assume that lighting changes and motion blur are minimal to ensure the reliability of the motion estimates. Further discussion will be provided later in the paper on how to address each of these aspects.

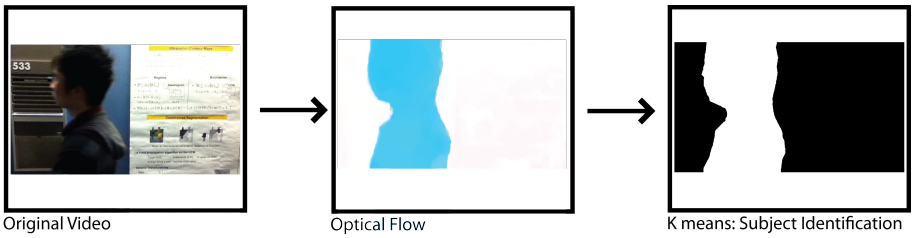

Figure 1: Pre-Computation of our pipeline. After a reliable estimate of the optical flow is computed for the input video, we use segmentation and clustering to estimate foreground and background regions.

3.1 Pre-Computation

Computing optical flow is computationally expensive for large amounts of data. To make the problem tractable, we down-sample the video and use a GPU accelerated large-displacement optical flow [17] to compute the motion field for each pair of neighboring frames in videos. The accelerated optical flow takes w seconds for a 853×480 resolution video per frame, which is a 60 times speed up over a CPU implementation. The output of the optical flow are forward and backward vectors that describe the motion from frame i to $i+1$ and motion from frame $i+1$ to i. To reduce the ambiguity of the flow estimates, we compute both the forward flow and backward flow and check if the results are similar; if they differ beyond a small threshold, we regard the flow as unreliable.

We assume there is one subject in the video, and will discuss relaxing this constraint later. We use k-means clustering with two clusters to segment the motion flows to those belonging to the foreground and background [15]. The cluster with the most frame edge pixels is chosen to be the foreground cluster and the cluster with the least edge pixels is chosen as the background cluster. We calculate the average subject and background flow vectors and number of pixels of each cluster, represented by $\mu_{foreground_n}$, $\mu_{background_n}$, $A_{foreground_n}$, and $A_{background_n}$ respectively. The representative position of the cluster is the average point (x, y) of all the points within the cluster, which is represented by $X_{foreground_n}$ and $X_{background_n}$. We also compute the effective subject motion flow, which is the motion of the subject compensated with the motion of the camera. If we were to assume that the background is mostly stationary, then $\mu_{background_n}$ would be the estimated camera motion. Thus, our compensated subject motion flow would be the average flow computed from the GPU accelerated large-displacement optical flow subtracted with $\mu_{background_n}$. This compensated subject motion flow will be used for feature extraction as we are interested in analyzing the motion of the subject independent of camera motion.

3.2 Frame analysis features

As mentioned previously, frame analysis features are computed at each frame for the video clip. Therefore, we will obtain N feature values per feature per video for a video

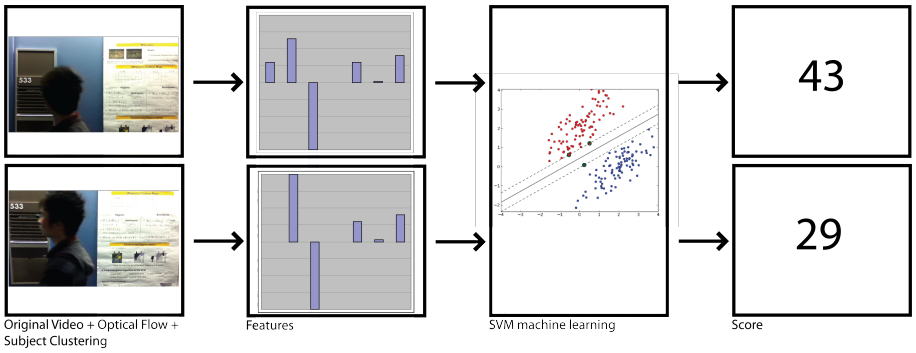

Figure 2: Features for each video is computed based on its optical flow and foreground background segmentation. If the features are highly correlated with the perceived aesthetics of the video, we can use a linear classifier such as a SVM to predict the scores.

with N flow frames. It is not entirely clear how these N feature values can be intuitively combined to give a lower dimension vector that would still encapsulate the behavior of the feature for the video. Previous work [15] uses statistical quantities averaged over small windows, as mean, variance and quartile values. Instead, we propose to represent the N features by a histogram as well as by the mean values over small windows. The reason for using both is that although the histogram is robust to outliers, it discards temporal information. Hence, we also compute the mean and variance over small fixed windows and have the statistical quantities ordered in a vector.

3.2.1 Motion in headroom

One observation from the video-graphic community is that viewers often find motion that is larger in the top third than in the bottom two thirds to be distracting. Therefore, this feature is designed to capture the relative motion in the top third of the video frame with respect to the bottom two thirds of the video frame. To that end, we compute the average motion of the compensated subject motion in the top third and the bottom two thirds. A simple ratio is used as the feature with addition of unity in both the numerator and denominator to ensure stability [15].

$$f_1 = \frac{1 + \mu_{top}}{1 + \mu_{bot}}$$

3.2.2 Direction of motion

There is a substantial body of literature in perception that suggests that viewers may have an overall bias for leftward or rightward directionality – both for motion and for the facing direction of objects in static images. This work originated with a claim by [20], later cited by [8], which postulated that aesthetically pleasing images tend to have the content of interest located on the right side of the image. These claims were later tested, and though some authors [10, 14, 13] found this effect to be modulated by

handedness, the most recent explorations [16, 12, 6] demonstrated that reading direction accounted for the preponderance of the bias. With this research in mind, we design a feature that describes the general motion within a particular frame. This is done by taking the ratio of the average motion to the left versus the average motion to the right of the subject.

$$f_2 = \frac{1 + \mu_{left}}{1 + \mu_right}$$

.

3.2.3 Magnitude of motion versus subject size

One interesting rule of thumb that videographers use that has not been explored is that large objects should have relatively small motion to avoid overwhelming the viewer so that the viewer will not be overwhelmed whereas small objects should have larger motions to prevent the scene from being visually boring. Therefore, we propose a feature that expresses this relationship as the ratio of the motion of the subject relative to the area of the subject.

$$f_3 = \frac{1 + |\mu_{foreground_n}|}{1 + A_{foreground_n}}$$

.

3.3 Video clip analysis features

Video clip analysis features are computed using all the frames in the video. These can be thought of as global features that gives a bigger picture to the motion behavior in the video. We believe that these features will provide high level motion descriptors that were absent in previous work [15, 11].

3.3.1 Characteristic of subject motion

One rule of thumb in videography is that viewers often prefer subject motion to be smooth and predictable. One way to characterize the motion of the subject across the video sequence is to fit a polynomial to the 2-D trajectory of the subject. First, we compute the subject's motion relative to the background for each frame using $\mu_{relative} = \mu_{foreground} - \mu_{background}$. Integrating the subject's relative motion provides an estimate for a 2-D path of the subject. We fit a linear and a quadratic polynomial to the path and use the coefficients and the residues as features for the subject motion. Intuitively, the polynomial coefficients will describe the nature of the motion while the residues will describe the jerkiness of the motion.

3.3.2 Characteristic of camera motion

Although it is not uncommon to see videos that are intentionally shot with shake, most viewers prefer the camera to move smoothly, if at all. Again, we fit a linear and a quadratic polynomial and use the coefficients and residues as features for the camera

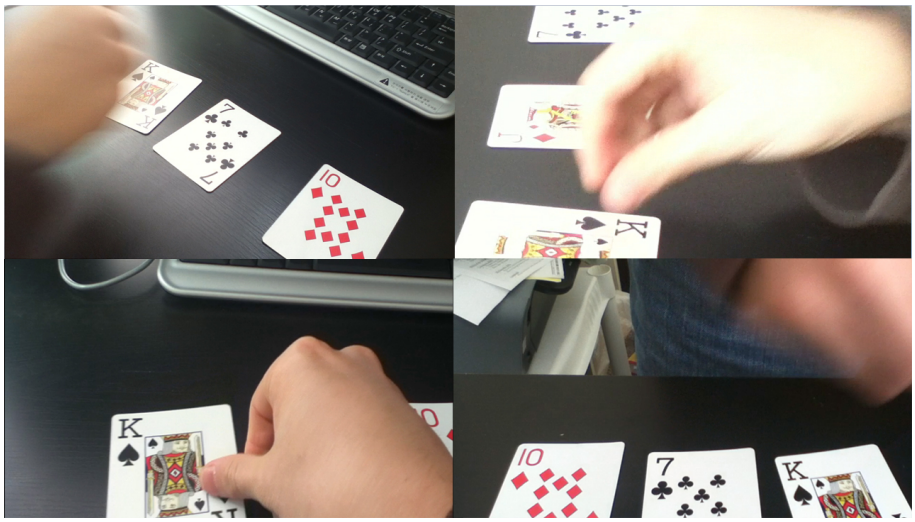

Figure 3: A set of videos contain 4 shots of the same scene with varying aesthetic qualities.

motion. However, we integrate just the background motion $\mu_{background}$ to obtain a 2-D path of the camera motion.

3.3.3 Locality of motion

This feature is designed to describe the location of the subject throughout the video. To achieve this, we compute a histogram as well as the mean and variance of the centroid, $X_{foreground}$, of the subject over small windows similar to the technique used to aggregate frame analysis features. This yields a distribution of both vertical and horizontal position of the subject over the video sequence and the statistics of the position of the subject across time.

4 Dataset and Evaluation

In our dataset, we removed the inconsistent hardware and post-processing involved in capturing video. We used multiple identical Apple iPod Touches, taking videos at 720P and 30 frames per second. All video clips have been through only the iPod's native post-processing and not any additional post-processing. We felt these are limitations to the hardware and, therefore, should not be accounted for in the dataset to evaluate aesthetics.

The data contains 64 sets, each comprising four video clips. We provided multiple scenarios such as panning, dollies, subject-focusing and rotations. The videos are shot through multiple scenarios to present a better range. We also took the images under multiple lighting conditions such as indoor, night, and day.

To find ground-truth data, we asked participants to evaluate the videos in our dataset. The participants were recruited using Amazon's Mechanical Turk in conjunction with Qualtrics Labs, Inc. software of the Qualtrics Research Suite. We collected data from 401 participants and grouped them to have between 70 to 90 people per group. Each group was shown one set of four different videos of the same scene for each of three scenes. Each set of videos was presented on a single screen, and participants could watch them in any order and multiple times if they chose to do so. After viewing each set, participants were prompted to rank order the four videos in that set from most aesthetically pleasing to least aesthetically pleasing. They were then presented with a free-response field in which they were asked to explain in a few sentences why they had ranked the videos in the order that they did. To ensure that participants watched each video completely at least once, there was a randomly-generated character that was presented as the final frame of each video, and participants were asked to type in these characters before proceeding to the next set.

5 Analysis and Discussion

If there is a strong correlation between our proposed features and user preferences, we would expect to observe a monotonic (or some functional) relationship between our feature scores and user scores. However, it is unfortunate that such a relationship does not exist for the features.

One reason is the unreliability of the foreground/background estimates. For example, panning shots would result in similar motions for background and foreground subjects. Also, if the video has multiple moving objects, the current clustering technique will not work.

6 Discussion and Future Work

Accurate segmentation and clustering of the foreground and background is crucial for our features to represent motion in the video. Therefore, inaccuracies in these steps is likely the cause of the poor performance of the features. With inaccurate subject detection, our heuristics provide unreliable results, creating unpredictable rankings. Work can be extended in this regard to make the subject detection more reliable, giving better results.

Another aspect for future work is to extend the idea of analyzing motion to more heuristics.

7 Conclusion

We hypothesized that motion features can inform us on user video preferences. However, with our current study we have found limited correlation between the two. This is likely due to the poor foreground and background segmentation. Also, it is difficult to eliminate semantic variations between the videos.

8 Acknowledgements

We would like to thank Narayanan Sundaram for his help with optical flow and the insightful discussions.

References

[1] R. Arnheim. *Film as Art*. University of California Press, 1957. 4

[2] R. Arnheim. *Visual Thinking*. University of California Press, Apr. 1969. 4

[3] R. Arnheim. *Art and Visual Perception: A Psychology of the Creative Eye*. University of California Press, Nov. 1974. 4

[4] T. Brox and J. Malik. Large displacement optical flow:descriptor matching in variational motion estimation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, August 2010. 3

[5] R. Datta, D. Joshi, J. Li, and J. Wang. Studying aesthetics in photographic images using a computational approach. In A. Leonardis, H. Bischof, and A. Pinz, editors, *Computer Vision ECCV 2006*, volume 3953 of *Lecture Notes in Computer Science*, pages 288–301. Springer Berlin / Heidelberg, 2006. 2

[6] M. De Agostini, S. Kazandjian, C. Cavézian, J. Lellouch, and S. Chokron. Visual aesthetic preference: Effects of handedness, sex, and age-related reading/writing directional scanning experience. *Writing Systems Research*, 2011. 7

[7] S. Eisenstein. *Film Form: Essays in Film Theory*. Harvest Books, Mar. 1969. 4

[8] M. Gaffron. Left and right in pictures. *Art Quarterly*, (13):312–331, 1950. 6

[9] Y. Ke, X. Tang, and F. Jing. The design of high-level features for photo quality assessment. In *in IEEE Computer Society Conference on Computer Vision and Pattern Recognition,2006,vol.1,pp.419426*, pages 419–426, 2006. 2

[10] J. Levy. Lateral dominance and aesthetic preference. *Neuropsychologia*, 14(4):431 – 445, 1976. 6

[11] Y. Luo and X. Tang. Photo and video quality evaluation: Focusing on the subject. In *Proceedings of the 10th European Conference on Computer Vision: Part III*, pages 386–399, Berlin, Heidelberg, 2008. Springer-Verlag. 2, 7

[12] A. Maass and A. Russo. Directional bias in the mental representation of spatial events: nature or culture? *Psychol Sci*, 14(4):296–301, 2003. 7

[13] J. P. McLaughlin. Aesthetic preference and lateral preferences. *Neuropsychologia*, 24(4):587 – 590, 1986. 6

[14] J. P. McLaughlin, P. Dean, and P. Stanley. Aesthetic preference in dextrals and sinistrals. *Neuropsychologia*, 21(2):147 – 153, 1983. 6

[15] A. Moorthy, P. Obrador, and N. Oliver. Towards computational models of the visual aesthetic appeal of consumer videos. In K. Daniilidis, P. Maragos, and N. Paragios, editors, *Computer Vision ECCV 2010*, volume 6315 of *Lecture Notes in Computer Science*, pages 1–14. Springer Berlin / Heidelberg, 2010. 1, 2, 3, 5, 6, 7

[16] I. Nachson, E. Argaman, and A. Luria. Effects of Directional Habits and Handedness on Aesthetic Preference for Left and Right Profiles. *Journal of Cross-Cultural Psychology*, 30(1):106–114, 1999. 7

[17] N. Sundaram, T. Brox, and K. Keutzer. Dense point trajectories by gpu-accelerated large displacement optical flow. In K. Daniilidis, P. Maragos, and N. Paragios, editors, *Computer Vision ECCV 2010*, volume 6311 of *Lecture Notes in Computer Science*, pages 438–451. Springer Berlin / Heidelberg, 2010. 2, 3, 5

[18] H. Tong, M. Li, H. jiang Zhang, J. He, and C. Zhang. Classification of digital photos taken by photographers or home users. In *In Proceedings of Pacific Rim Conference on Multimedia*, pages 198–205. Springer, 2004. 2

[19] M. Werlberger, W. Trobin, T. Pock, A. Wedel, D. Cremers, and H. Bischof. Anisotropic Huber-L1 optical flow. In *Proc. of the British Machine Vision Conference (BMVC)*, September 2009. 2, 3

[20] H. Wölfflin. *Über das Rechts und Links in Bilde. Gedanken zur Kunstgeschichte*. Schwabe, Basel, Switzerland, 1957. 6

[21] C. Zach, T. Pock, and H. Bischof. A duality based approach for realtime TV-L1 optical flow. volume 4713 of *LNCS*, pages 214–223. Springer, 2007. 2, 3