

Using Web Photos for Measuring Video Frame Interestingness

Feng Liu

Department of Computer Sciences
University of Wisconsin-Madison
fliu@cs.wisc.edu

Yuzhen Niu

School of Computer Science
Shandong University
yuzhen@cs.wisc.edu

Michael Gleicher

Department of Computer Sciences
University of Wisconsin-Madison
gleicher@cs.wisc.edu

Abstract

In this paper, we present a method that uses web photos for measuring frame interestingness of a travel video. Web photo collections, such as those on *Flickr*, tend to contain interesting images because their images are more carefully taken, composed, and selected. Because these photos have already been chosen as subjectively interesting, they serve as evidence that similar images are also interesting. Our idea is to leverage these web photos to measure the interestingness of video frames. Specifically, we measure the interestingness of each video frame according to its similarity to web photos. The similarity is defined based on the scene content and composition. We characterize the scene content using scale invariant local features, specifically SIFT keypoints. We characterize composition by feature distribution. Accordingly, we measure the similarity between a web photo and a video frame based on the co-occurrence of the SIFT features, and the similarity between their spatial distribution. Interestingness of a video frame is measured by considering how many photos it is similar to, and how similar it is to them. Our experiments on measuring frame interestingness of videos from *YouTube* using photos from *Flickr* show the initial success of our method.

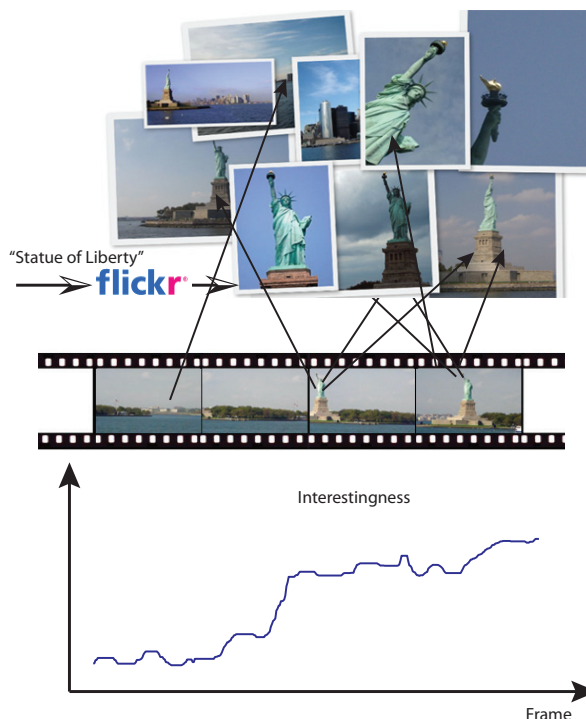


Figure 1: Schematic example. Given a video, our method uses its associated keyword to query photos from web sites, such as *Flickr*, and measures the interestingness of the video frames by voting each frame using the web photo set.

1 Introduction

Many video applications benefit from a metric of *interestingness* of their frames, that is, a measure of the expected subjective response from a viewer. For example, video summarization, abstraction, and collection browsing all benefit from the ability to select the frames of a video that are likely to be most interesting to a viewer.

A computational model of interestingness is elusive because it is subjective, content- and application- dependent. However, we observe that it is possible to find examples of interesting images. Online photo collections tend to contain interesting images because their images are more carefully taken and composed, and have been culled from all those taken (c.f. [Krages, 2005; Peterson, 2003; Kirk *et al.*, 2006]).

Because photos have already been chosen as subjectively interesting, they serve as evidence that similar images are also interesting. In this paper, we propose to leverage these examples in order to create a video frame interestingness metric. Specifically, we present a method that uses collections of web photos to measure the interestingness of frames in videos by matching the frames against collections of web photos. Our initial experiments focus on travel videos of well-known places as it is easy to find appropriate videos and photo collections, and common motifs are easily recognizable.

Using photo collections implicitly captures people's subjective assessment. In contrast, other methods for frame assessment do not consider interestingness as directly. For ex-

ample, visual salience [Itti and Koch, 2001] measures the affinity of the low-level visual system to imagery, and has been used to select frames for video summarization [Ma *et al.*, 2002]. Other low-level image properties, such as image quality and uniqueness, have also been used [Truong and Venkatesh, 2007; Liu *et al.*, 2008a]. However, there is often a gap between these low level metrics and viewers' subjective response as the latter is often influenced by higher level aspects such as content and composition.

Public web photo collections, such as *Flickr*, make large numbers of images available, but these collections are often noisy: while the collected images *tend* to be interesting (because they have been selected), they vary in content and quality. Therefore, to exploit these collections, we must consider entire collections so that statistically, the contents are likely to be indicative of interestingness. The idea of using large, but potentially noisy, collections of web data as a measure of people's assessment has been applied previously for ranking web pages [Liu *et al.*, 2008b] and scene segmentation [Simon and Seitz, 2008].

In this paper, we present an approach where a collection of photographs is used to assess the interestingness of each frame of a video. Our approach first gathers a relevant photo collection by searching for images with the same keywords as the video to be processed. Each frame of the video is compared with all photos in the collection. Frames that are more similar to more photos are judged to be more likely to be interesting. To make the process efficient, we exploit the coherence in the video. Our preliminary experiments suggest that our metric for measuring video frame interestingness is promising.

2 Measuring Video Frame Interestingness

Given a video, our method first uses its associated keywords to search for relevant photos from a photo sharing web site, specifically *Flickr* in our experiments. Then we use these web photos to vote for the interestingness of each video frame by finding its similar photos in the web photo set as described in the following subsections. For clarity, we define notation here. We denote a video as a frame sequence $V = \{V_i | 1 \leq i \leq n\}$, where V_i is its i^{th} frame. We denote a web photo set as $P = \{P_j | 1 \leq j \leq m\}$, where P_j is the j^{th} photo.

2.1 Photo-Frame Registration

Matching a video frame to a photo is an image registration problem. There is a rich literature on image registration. There are two types of methods, direct matching and feature-based matching. Direct methods use pixel to pixel matching, while feature-based methods first detect features from each image, build the feature correspondence between images, and estimate a matching model based on the feature correspondence. A good survey can be found in [Szeliski, 2006]. For our problem, each image and video frame can be taken with varying viewpoint, camera settings and content. Therefore, direct methods are not suitable. We use a feature-based method.

SIFT features are invariant to image scale and rotation, and have been shown to provide robust matching across a substantial range of affine distortion, change in 3D viewpoint,

addition of noise, and change in illumination [Lowe, 2004]. These properties qualify SIFT features as distinguishable features for matching between a video frame and a photo. Therefore, we use a SIFT-feature based method for matching each video frame and photo as follows.

1. Extract SIFT features from video frame V_i and photo P_j .
2. Build the feature correspondences between V_i and P_j . Specifically, for each feature in V_i , we search for its counterpart in P_j . We use the matching criteria recommended in [Lowe, 2004]: if the ratio of its distance to its nearest neighbor and to its second nearest neighbor in P_j is smaller than a given threshold, its nearest neighbor is likely to be its counterpart in P_j . The default threshold used in our experiments is 0.8. Since the number of SIFT features in V_i and P_j is large, we need an efficient search method to find the nearest neighbors. Since each SIFT feature is represented as a 128-dimension vector, we use an approximate nearest neighbor searching method [Arya *et al.*, 1998].
3. Refine the feature correspondences if there are more than 4 pairs of matching features. A theorem of multi-view geometry states that images taken using different camera settings are related with a fundamental matrix [Hartley and Zisserman, 2000] as follows:

$$S_V^T F S_P = 0 \quad (1)$$

where S_V and S_P are the homogenous coordinates of the matching SIFT features in video frame V_i and photo P_j . F is a 3×3 matrix with 7 degrees of freedom. We estimate the fundamental matrix between the video frame and photo using a robust estimation algorithm (RANSAC [Fischler and Bolles, 1981]). Feature matches that do not observe the fundamental matrix constraint are ruled out as outliers.

4. If fewer than 4 feature matches are found, we consider video frame V_i and photo P_j as not related to each other.

2.2 Efficient Photo-Frame Registration

Using the above photo-frame registration method to match a frame to a photo in the photo set is effective, but time-consuming. For a video, the neighboring frames are most likely similar to each other. We improve the basic registration method by exploiting this temporal coherence. We first build SIFT feature tracks in the video, and then match the SIFT feature tracks to each photo in the set. This improved algorithm is detailed below.

Denote the l^{th} SIFT feature track as $T_l = \langle S_{Vl}, \{t | S_{Vl} \in V_t\} \rangle$ and the feature track set as $\Gamma = \{T_l | l = 1, 2, \dots\}$. We build the SIFT feature tracks by matching SIFT features between consecutive frames. Since two consecutive frames do not change much normally, local searching is both more efficient and accurate. We currently use a tile-based method. Specifically, each frame is divided into uniform tiles, and the SIFT points are binned into each tile. When searching for the matching SIFT points, only the points in the corresponding tile will be compared. Based on the matching result between consecutive frames, we can easily build the SIFT fea-

ture tracks by traveling through all the video frames and concatenating matching pairs.

After building the SIFT feature tracks, we match each SIFT track to each photo in the photo set. We again use the approximate nearest neighbor searching method [Arya *et al.*, 1998] to find the matching feature in each photo, and obtain an track-photo incidence matrix $M_{k \times m}$, where k is the total number of SIFT tracks from the video and m is the size of the photo set P .

$$M(l, j) = \begin{cases} i, & T_l \text{ matches feature } S_{P_j, i} \in P_j; \\ -1, & \text{else.} \end{cases} \quad (2)$$

$M(l, j)$ is the index of SIFT feature in photo P_j that matches the l^{th} SIFT track. From Γ and M , we know the feature correspondence between a video frame and a photo. We refine the matching using the fundamental matrix constraint as before.

2.3 Frame Interestingness Estimation

Web photos are usually carefully taken and selected. People often take the photos of a selected scene from a carefully selected viewpoint with carefully selected camera settings (c.f. [Krages, 2005; Peterson, 2003]). People tend to selectively upload the photos, rather than uploading all the photos they have taken. Some bad photos are often erased shortly after they are taken [Kirk *et al.*, 2006]. Thus, these web photos most likely have already been chosen as subjectively interesting. So they can serve as evidence that similar images are also interesting. Based on this observation, we calculate the interestingness of video frames according to their resemblance to the web photos in a variety of aspects based on the SIFT feature-based matching.

We measure the similarity between a video frame V_i and a web photo P_j based on the following criteria.

- **C1:** How much scene content they share. We measure the scene content shared by V_i and P_j according to the number of SIFT features they share since it has been shown that SIFT features are robust in object and scene recognition (c.f. [Lowe, 2004]).
- **C2:** How similar is the composition. That is, does similar content appear in similar image locations. Photo composition is important. For example, the rule of the thirds states that an image should be imagined as divided into nine equal parts by two equally-spaced horizontal lines and two equally-spaced vertical lines, and that important compositional elements should be placed along these lines or their intersections [Peterson, 2003]. If a video frame has similar compositions to the web photos, it is likely to be interesting. We measure the framing scheme similarity between V_i and P_j according to how features are positioned in the image.

The interestingness of each frame is, therefore, the number of features the frame shares with the photo collection, with each feature occurrence weighted by how similar the position of the feature is in the photo and frame:

$$I_i \propto \sum_{T_l \in V_i} \sum_{M_{l,j} \neq -1 \& P_j \in P} \log \frac{1}{\|\hat{S}_{P_j, M_{l,j}} - \hat{S}_{V_i, l}\|_2^2 + \epsilon} \quad (3)$$

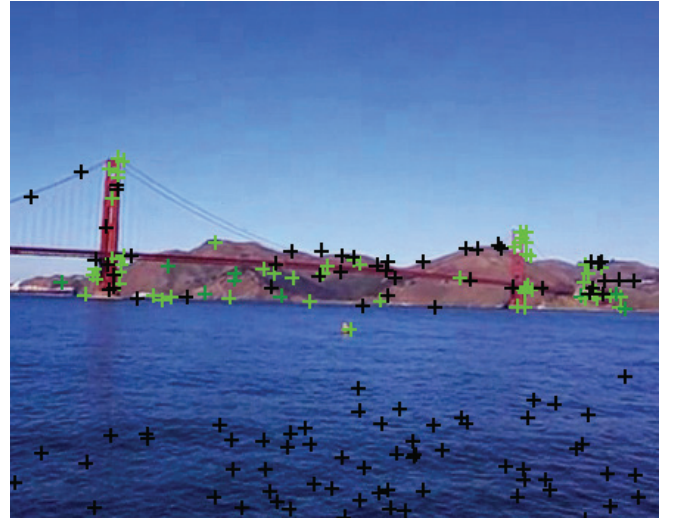


Figure 2: An interesting frame from a video with keyword “Golden Gate Bridge”. SIFT features are marked with “+”. Green (bright) “+” marks indicate “popular” features, i.e. ones that appear often in the corresponding photo set.

where I_i is the interestingness of the video frame V_i , P_j is a photo in the web photo set P , and M is the feature track-photo incidence matrix. $\hat{S}_{V_i, l}$ is the normalized coordinate of the SIFT feature of the feature track T_l in V_i , $\hat{S}_{P_j, M_{l,j}}$ is the normalized coordinate of the $M_{l,j}^{th}$ SIFT feature in photo P_j , and ϵ is a constant to avoid zero-division, with the default value $1.0e-3$. $\log \frac{1}{\|\hat{S}_{P_j, M_{l,j}} - \hat{S}_{V_i, l}\|_2^2 + \epsilon}$ measures the frame composition similarity between the video frame and the photo according to **C2**. The above equation states that I_i , the interestingness of a video frame V_i , is proportional to the total number of SIFT features shared with the web photo set P , weighted by the frame composition similarity. An example of an interesting video frame detected by our algorithm is shown in Figure 2.

3 Experiments

We designed a preliminary experiment to evaluate our method. The goals of the experiment are to

- Verify the feasibility of using web photos to measure the frame interestingness of travel videos.
- Verify if our interestingness metric corresponds with people’s subjective assessment of videos.

Our experiments use travel videos from *YouTube* and photos from *Flickr*. We collected a pool of scene names, and randomly selected 10 from the pool as listed in Table 1. We used these names as keywords to query videos from the *YouTube* travel category, and randomly selected 3 videos for each query from the top 10 retrieved results. Then we downloaded the top 250 photos from *Flickr* with each keyword.

We examined the downloaded *Flickr* photos. The photo content varies significantly, but in each set there exist a few big clusters of common scenes. We roughly clustered the

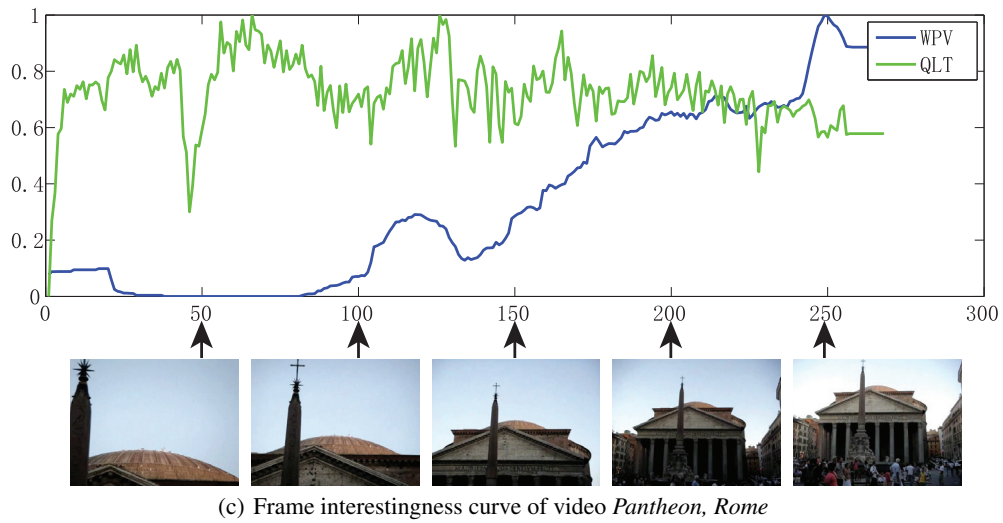
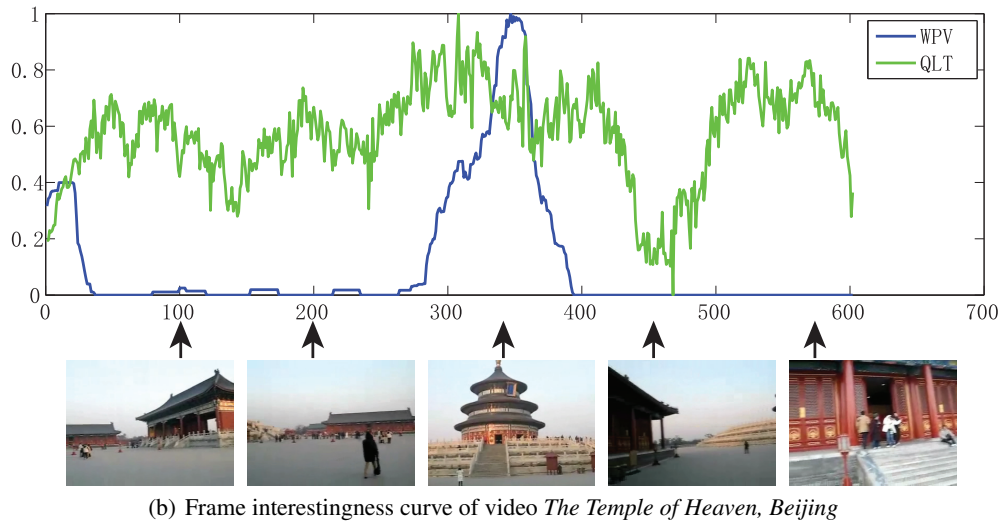
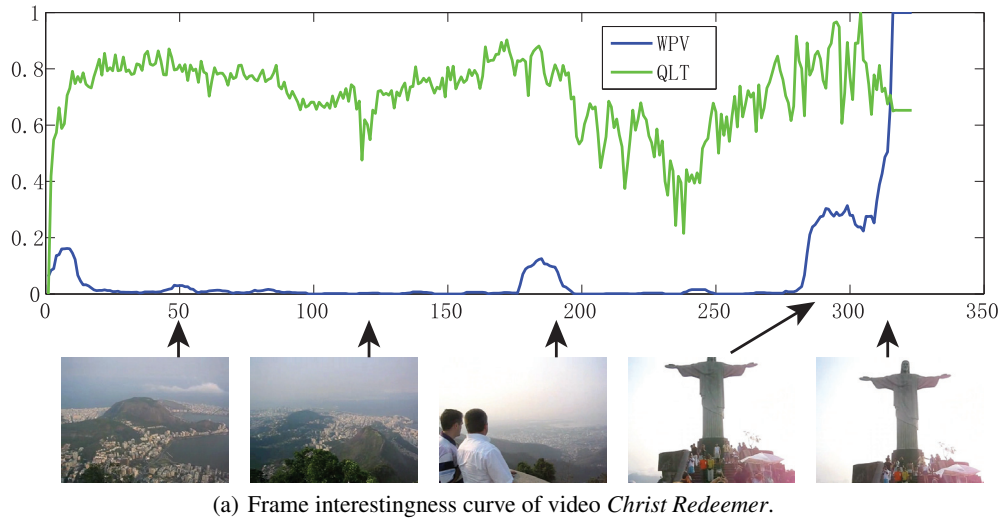


Figure 3: Representative frame interestingness prediction curves. *WPV* is our result and *QLT* is visual quality measurement. Curves are normalized to $[0,1]$.

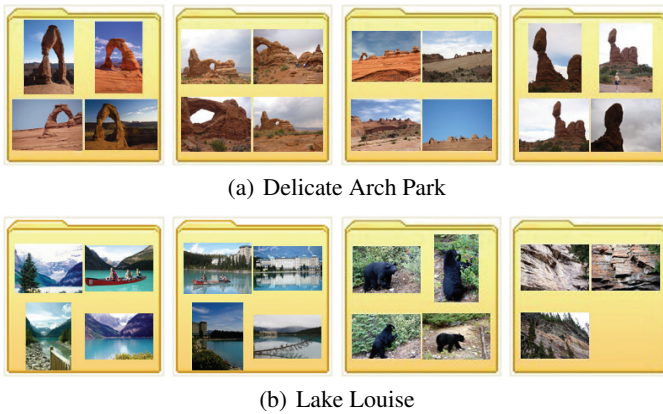


Figure 4: Some representative photo collections.

photos manually into clusters where the photos describe the same scene. The numbers of clusters with more than 10 photos for each photo set are reported in Table 2. Representative photos for some clusters are shown in Figure 4. We find that the big clusters most likely record the interesting scene content, although the number of photos of each of these big clusters is not dominant. The percentage of photos of the largest cluster is reported in Table 2.

We ran our algorithm on each video using the corresponding *Flickr* photo set to vote for its frame interestingness, and obtained the frame interestingness measurement for each video. Measuring the frame interestingness of a video is subjective. To evaluate our results we performed a preliminary user study. The study attempts to verify whether our algorithm’s predication is consistent with users’ evaluation. In the pilot study, 5 graduate students who are familiar with the scenes participated. Each participant watched each video and assessed whether the interestingness curve matched his/her expectation by giving a score ranging from 1 to 5. The higher the score, the better the predication matches the user’s judgement. The average score is 4.5 out of 5. The detail of the result is reported in Table 1.

Some results are shown in Figure 3. We also compared our results to the method (QLT) that uses the visual quality to measure the frame interestingness [Liu *et al.*, 2008a], where the visual quality is measured based on blurriness [Tong *et al.*, 2004] and blockiness [Wang *et al.*, 2002]. These examples suggest that our frame interestingness metric matches people’s subjective assessment. These examples also show that our results are not correlated with the QLT results significantly. One possible reason is that the visual quality of a video frame is not necessarily correlated with its interestingness to people. Figure 3 suggests that our method is more suitable and capable of measuring the frame interestingness of a video.

Although the experiment shows the success of our algorithm, it reveals some failure cases. First, when the quality of the video is bad, for example if it suffers from serious motion blur, our algorithm cannot perform well. The reason is that blurred video frames fail the SIFT feature detection and matching. This often happens when people move the cam-

Keyword	Avg. Score
Christ Redeemer	4.7
Delicate Arch Park	4.7
Statue of Liberty	4.7
Pisa Tower	4.7
Golden Gate Bridge	4.5
Pantheon, Rome	4
Lake Louise	4
The Temple of Heaven, Beijing	5
Taj Mahal, India	4
Eiffel Tower	4.7

Table 1: Experiment results.

Keyword	Num. of Clust.	Largest clust.
Christ Redeemer	4	77(30.8%)
Delicate Arch Park	4	139(55.6%)
Statue of Liberty	2	33(23.6%)
Pisa Tower	1	85(34%)
Golden Gate Bridge	2	161(64.4%)
Pantheon, Rome	5	51(20.4%)
Lake Louise	3	40(16%)
The Temple of Heaven	2	42(16.8%)
Taj Mahal, India	2	70(28%)
Eiffel Tower	2	137(54.8%)

Table 2: Web photo set statistics. For each keyword, we downloaded 250 photos with Common Creative License from *Flickr*, except *Statue of Liberty*, for which we could only download 140 photos. For each photo set, we reported the number of clusters with more than 10 photos in the middle column and the cluster size of the largest cluster in the last column.

era quickly. Since we currently use *YouTube* videos for the experiment, the compression artifacts also contribute some blurring and blocking artifacts. Videos taken at poor lighting conditions can also fail the feature matching. Second, some frames that mainly describe the surrounding scene around important objects are sometimes determined as important as the important objects. The reason is that many web photos record the main object of interest as well as its surrounding context. Our method considers the feature distribution, which helps to relieve this problem. However, when the object of interest does not have enough features while its surrounding has much more features, the frames with the surrounding content are mistaken as important as the frames with the object of interest.

Our observation from this experiment is that, although the content in the photo set varies, there exist large clusters that represent common interesting scene content. This observation supports that the web photo set contains people’s knowledge about what is important in a scene. Our pilot experiment suggests that our method of mining the knowledge in the photo set to measure the frame interestingness of travel videos can capture some of this knowledge. More complete experiments are needed to confirm this.

4 Conclusion

In this paper, we presented a method that uses web photos to vote for the frame interestingness of a travel video. Shared web photos are usually carefully taken and selected. So they can often be regarded as subjectively interesting and serve as evidence that similar images, are also interesting. Based on this idea, we developed an algorithm that matches a video frame to its relevant web photo set and use the web photos to vote for the frame interestingness. Our pilot experiments suggest that web photos capture people's subjective sense of scene interestingness and this knowledge can be exploited to measure the frame interestingness of a travel video.

Our method is complementary to visual quality measurement. It determines video frames containing well positioned interesting content, while the visual quality measurement can be used to select high-quality frames from those interesting ones. Combining them together, especially in specific application scenarios, will lead to better performance than any one of them alone.

While our pilot experiment is encouraging, further experimentation is necessary to both confirm the effectiveness of our method at predicting viewers' subjective expectations, as well as to understand how large and specific an image collection is required for the approach to be effective. Our current user study is still preliminary. Instead of directly measuring the quality of our results, it currently only checks if the user confirms its prediction or not. More studies are needed to measure the accuracy of the prediction. For example, we can show each participant pairs of frames that receive different scores from our method, and ask her/him which is more interesting. Accordingly, we can score our method based on how often participants agree with our method's prediction.

Our interestingness measure currently only looks at the scene content and image composition of a video. It relies on having photos that are similar enough in content to the video. So our initial experiments focus on travel videos of well-known places as it is easy to find appropriate videos and photo collections, and common motifs are easily recognizable. However, even for a travel video, more factors could contribute to its interestingness. For example, the interestingness might come from the people in it or etc. Moreover, extending the approach to more general categories of video content will be challenging.

In future, we plan to apply our method to some real-world applications, such as video summarization, pictorial video representation, and video retrieval. It will be both helpful to solving real-world problems as well as further evaluating our method.

Acknowledgements We thank reviewers for their constructive suggestions. This research was sponsored in part by NSF grant IIS-0416284.

References

- [Arya *et al.*, 1998] Sunil Arya, David M. Mount, Nathan S. Netanyahu, Ruth Silverman, and Angela Y. Wu. An optimal algorithm for approximate nearest neighbor searching fixed dimensions. *J. ACM*, 45(6):891–923, 1998.
- [Fischler and Bolles, 1981] Martin A. Fischler and Robert C. Bolles. Random sample consensus: a paradigm for model fitting with applications to image analysis and automated cartography. *Communications of the ACM*, 24(6):381–395, 1981.
- [Hartley and Zisserman, 2000] R. I. Hartley and A. Zisserman. *Multiple View Geometry in Computer Vision*. Cambridge University Press, ISBN: 0521623049, 2000.
- [Itti and Koch, 2001] L. Itti and C. Koch. Computational modeling of visual attention. *Nature Reviews Neuroscience*, 2(3):194–203, Mar 2001.
- [Kirk *et al.*, 2006] David Kirk, Abigail Sellen, Carsten Rother, and Ken Wood. Understanding photowork. In *CHI '06: Proceedings of the SIGCHI conference on Human Factors in computing systems*, pages 761–770, New York, NY, USA, 2006. ACM.
- [Krages, 2005] Bert P. Krages. *The Art of Composition*. Allworth Communications, Inc., 2005.
- [Liu *et al.*, 2008a] Feng Liu, Yu-hen Hu, and Michael Gleicher. Discovering panoramas in web videos. In *ACM Multimedia*, pages 329–338, 2008.
- [Liu *et al.*, 2008b] Yuting Liu, Bin Gao, Tie-Yan Liu, Ying Zhang, Zhiming Ma, Shuyuan He, and Hang Li. Browser-ank: letting web users vote for page importance. In *SIGIR '08: Proceedings of the 31st annual international ACM SIGIR conference on Research and development in information retrieval*, pages 451–458, 2008.
- [Lowe, 2004] D. G. Lowe. Distinctive image features from scale-invariant keypoints. *International Journal of Computer Vision*, 60(2):91–110, 2004.
- [Ma *et al.*, 2002] Y.F. Ma, L. Lu, H.J. Zhang, and M.J. Li. A user attention model for video summarization. In *Proceedings of ACM Multimedia 2002*, pages 533–542, 2002.
- [Peterson, 2003] Bryan F. Peterson. *Learning to see creatively*. Amphoto Press, 2003.
- [Simon and Seitz, 2008] Ian Simon and Steven M. Seitz. Scene segmentation using the wisdom of crowds. In *ECCV*, pages 541–553, 2008.
- [Szeliski, 2006] Richard Szeliski. Image alignment and stitching: a tutorial. *Found. Trends. Comput. Graph. Vis.*, 2(1):1–104, 2006.
- [Tong *et al.*, 2004] Hanghang Tong, Mingjing Li, Hongjiang Zhang, and Changshui Zhang. Blur detection for digital images using wavelet transform. In *IEEE International Conference on Multimedia & Expo*, 2004.
- [Truong and Venkatesh, 2007] Ba Tu Truong and Svetha Venkatesh. Video abstraction: A systematic review and classification. *ACM Trans. Multimedia Comput. Commun. Appl.*, 3(1):3, 2007.
- [Wang *et al.*, 2002] Z. Wang, H.R. Sheikh, and A.C. Bovik. No-reference perceptual quality assessment of jpeg compressed images. In *IEEE ICIP*, pages Vol I: 477–480, 2002.