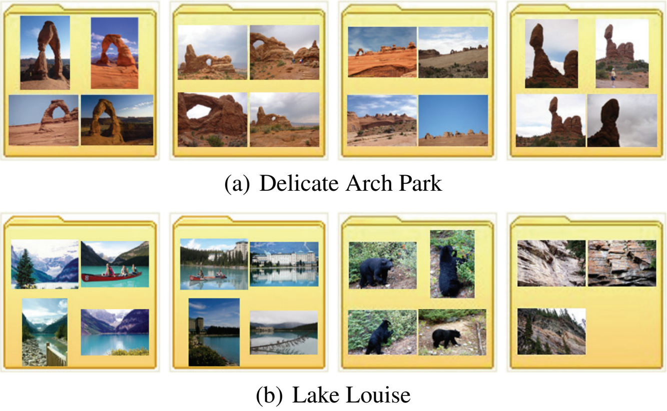

(a) Delicate Arch Park

(b) Lake Louise

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 blockness [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.

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.

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 camera 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.