Discovering Panoramas in Web Videos

Lingfeng Huang, Fang Wang

11/2016

Abstract

Panoramas have been widely used in may applications in multimedia, but the main constraint for panoramas is that they must be taken by people who physically present at the place. In this project, we will implement our version of Discovering Panoramas in Web Videos by Liu et al. to solve the problem by selecting optimal segments within a given web videos, then perform synthesizing to obtain panoramas. This whole procedure is basically a optimization problem where we optimize the three criteria which are wide field of view, mosaicability, and high image quality.

1 Introduction

The emerge of the idea "Panorama" has to be dated back to early 20 A.D. and was a means of generating an 'panoptic' view of a vista [**wikipedia**]. Nowadays, with the help of advancement of technology, people are able to create desired panorama by simply rotating their cell phones and clicking the shot button. The process of synthesizing panorama is relatively straight forward. First step is to take successive photos from the same optical center and next step is finding the alignment between each image and warping accordingly, and final step is interpolating the warped image and applied certain blending to remove the visible seams. However, the problem with creating panoramas using above approach is that people are required to physically appear at the place where they take the images, which means that if people want to take a panorama of Time Square in New York, they have to fly over New York to do so.

Compared to sequence of images, although some segments within videos have relatively low image quality and also moving object, they are still shot from the same optical center and cover a wide field-of-view. Lui et al. suggests an approach that synthesizing panoramas by identifying proper segments within videos as panorama source [**Lui**]. They convert the problem into a optimization problem, and set up three constraints in order to evaluate the video segments. Lui ei al. indicates that in order to be a appropriate panorama source, a video segment should cover a wide field-of-view based on the definition of panorama imagery, be ”mosaicable” and the frames should have high image quality[**Lui**].

2 Visual Quality Measure

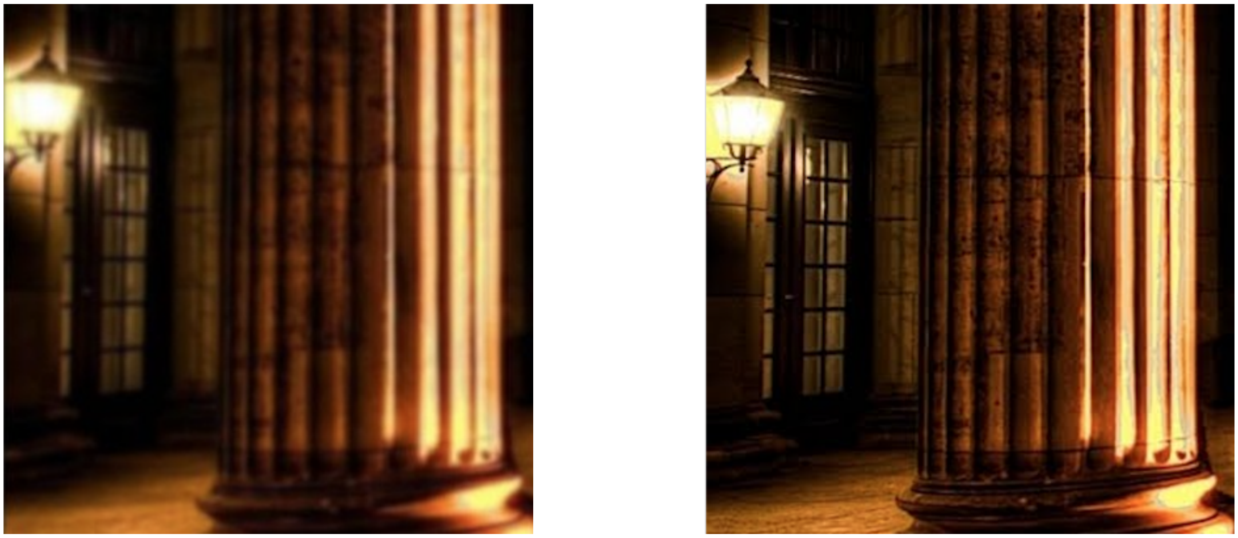
We measure the visual quality of a single frame based on two terms, one is incorrectness of the motion model $E_{vm}(S_i)$ and the source image visual quality $E_{vv}(S_i)$. Then by setting up the visual quality distortion $E_v(S_i)$, we can obtain the visual quality measure.

$$E_v(S_i) = \alpha_m E_{vm}(S_i) + \alpha_v E_{vv}(S_i) \tag{1}$$

By default, we set both weights α_v and α_m to be 1.0.

2.1 Source Image Visual Quality

The source image visual Quality $E_{vv}(S_i)$ is defined as how blurry and blocky the image is. We use the idea of Tong et al.'s method of measuring blurring artifacts by using Haar Wavelet Transform [**Tong**].

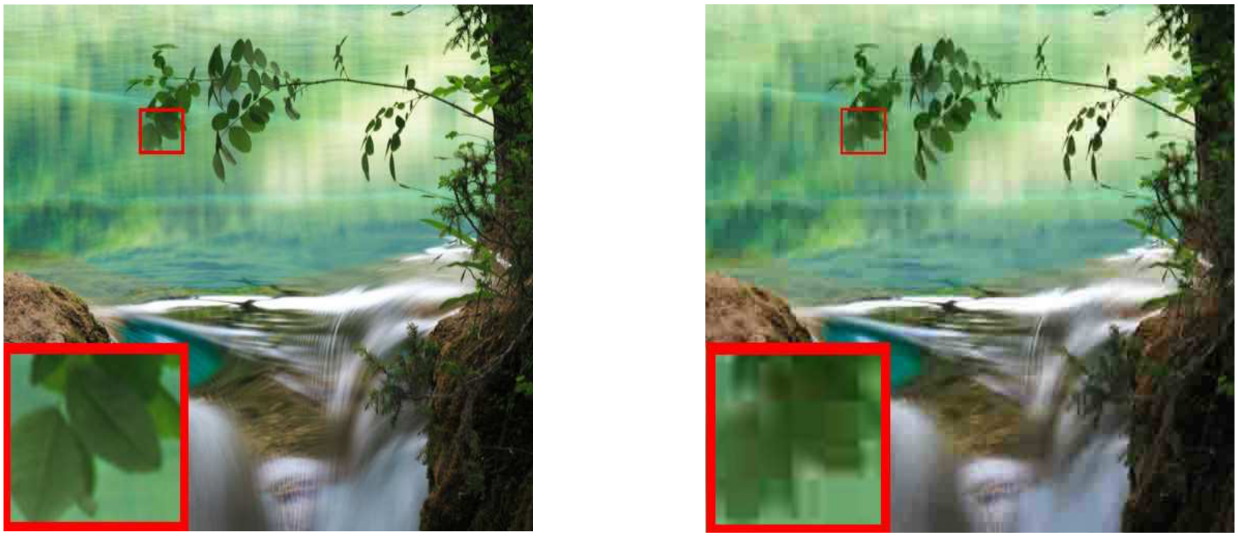


(a) Blurriness: 0.8086

(b) Blurriness: 0.3648

Figure 1: Blurriness measure compare

The blockiness is measured by using the method of Wang et al. which estimates the average difference across block boundaries modulated by image activities [**Wang**].



(a) Blockiness: 0.204

(b) Blockiness: 0.479

Figure 2: Blockiness measure compare

After obtain the blockiness and blurriness from all the frames within a given video segment, we calculate the visual distortion for this segment as follows:

$$E_{vv}(S_i) = \sum_{I_k \in S_i} \gamma q_{bk}(I_k) + (1-\gamma) q_{br}(I_k) \tag{2}$$

where $q_{bk}(I_k)$ is the measurement of blockiness of given frame, and $q_{br}(I_k)$ is the measurement of blurriness. Weight γ is set to 0.45.

2.2 Incorrectness of Motion Model

In order to achieve the "mosaicablity", we use a homography to model the motion between frames. By matching SIFT feature points, we are able to locate significant points between frames and thus obtain the homography. In practice, getting a high quality panorama from video requires the inter-frame motion is closed to its homography and few casual videos can achieve that. Therefore, we measure the error using the real motion vector from SIFT feature points and the predicted value by homography between two successive frames.

$$E_{vm}(S_i) = \sum_{I_k \in S_i} \frac{1}{n_k} \sum_{p_{j,k} \in S_i} \| mv(p_{j,k}), mv_h(p_{j,k}) \| \tag{3}$$

We first for each adjacent frame I_k and I_{k+1} find its matching SIFT feature pairs, and calculate homography using RANSAC based on these feature pairs. The notation $mv(p_{j,k})$ is the motion vector of j^{th} SIFT feature point of frame I_k and $mv_h(p_{j,k})$ is the predicted motion vector by homography at j^{th} feature point. Then the error of j^{th} feature point is taking the L1 norm of these two terms. Then we average the errors of all feature pair in each frame and obtain the incorrectness of motion model by summing up all the average.

3 Extent of Scene

The extent of scene $\varepsilon(S_i)$ is defined as the scene covered by the video segment. Although it could be good if we maximizing the covering area to obtain a larger field of view, larger field of view often means more distortion. Hence, we want to choose the reference frame where the distortion is minimum, in other words, we want the minimum area covered by segment S_i

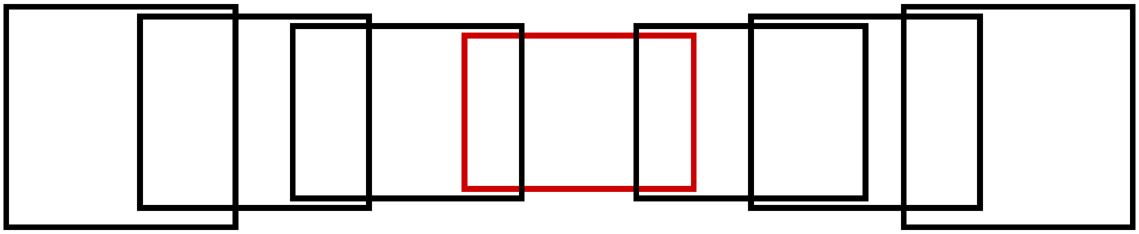


Figure 3: All the other frames in the video are alignment to the reference frame, and we want to find the optimal reference frame so that the area covered by aligning all the frames to be minimized

In order to find the optimal reference frame, we want to search for all possible combination of panorama alignment as follows:

$$r = \arg\min_{f_r \in S_i} \bigcup_{f \in S_i} I(f, f_r) \tag{4}$$

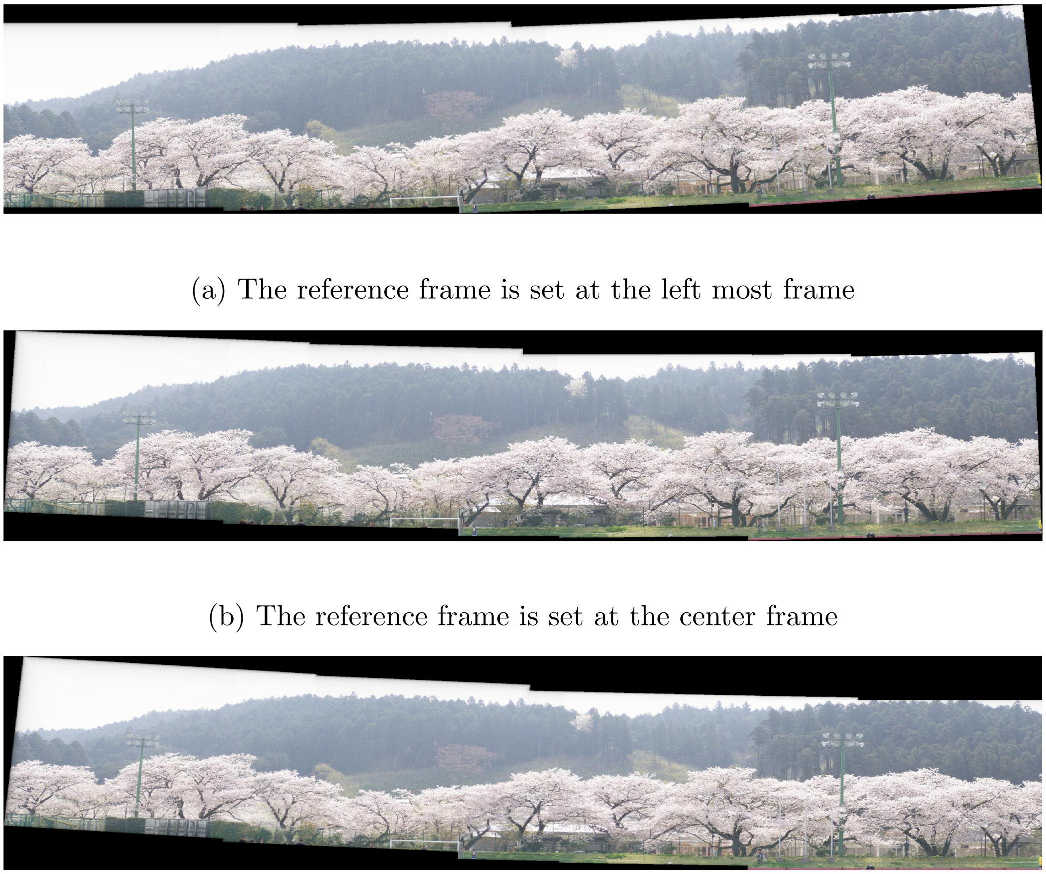
Then the extent of scene is defined as:

$$\varepsilon(S_i) = \bigcup_{f \in S_i} I(f, r) \tag{5}$$

In Lui et al. paper, they used generic polygon clipping for finding the minimum area covered. However, we used another approach. Although the accuracy declines, the complexity reduces in decent amount, we choose to find the distance between the center of reference frame and that of all other frames as follows:

$$\frac{1}{n} \sum_{f \in S_i} \| C_r, f_r \| \tag{6}$$

Where n is the number of frames within the segment, C_r is the center pixel of the reference frame and f_r is the center pixel in a given frame. Figure 4 demonstrates when reference frame is the most left frame (the first frame in the image sequence), the middle frame and the right



(a) The reference frame is set at the left most frame

(b) The reference frame is set at the center frame

(c) The reference frame is set at the right most frame

Figure 4: Difference reference frame results in different look of panorama

most frame. However, due to the relatively small distortion since this image sequence only has 5 images, the visual difference between three panoramas is subtle.

4 Experiments

We randomly select some videos from Youtube and manually chop the video into several segments. Lui et al. use histogram based shot boundary detection to divide the video, and in order to simplify the process, we choose to feed our algorithm with preprocessed video segments.

5 Limitations

First of all, this whole process is not fully automated. At this point, we are not able to pass a whole video and get several possible results. Second, the complexity of this algorithm seems relatively high since we need to recalculate the homographies between adjacent frame for each iteration when we calculate the extent of scene. In addition, the run time for computing blurriness for each frame is also relatively high.