Gaze-Driven Video Re-Editing

Bachelor's Project (BTP-I and BTP-II)

Advisors: Prof. Kavita Vemuri and Prof. Vineet Gandhi

Contents

Chapter 1

Introduction

Due to the numerous types of media devices, a video of certain aspect ratio will need to be viewed on various devices with different aspect ratios.

Displaying a video on a different aspect ratio may cause content to be modified or reduce user appeal. The task is to preserve the narrative or impact of the video by re-editing based on the implicit information revealed in tracking the user's gaze as they watch the video.

Chapter 2

BTP I

Initial work involved doing a literature survey of existing video retargeting methods, and then study andimplementing the reference research paper "Gaze-Driven Video Re-Editing" by Jain et al, to evaluate and possibly improve the performance of the system.

2.1 Understanding the paper

The paper proposed searching for a cropping window path through the video cube while maximizing the number of gaze points. The primary editing operations that were considered in the paper were pans, which are gradual shifts of horizontal position as the user's attention shifts, and cuts, which are quick shifts from one part of a scene to another.

2.1.1 Data Collection

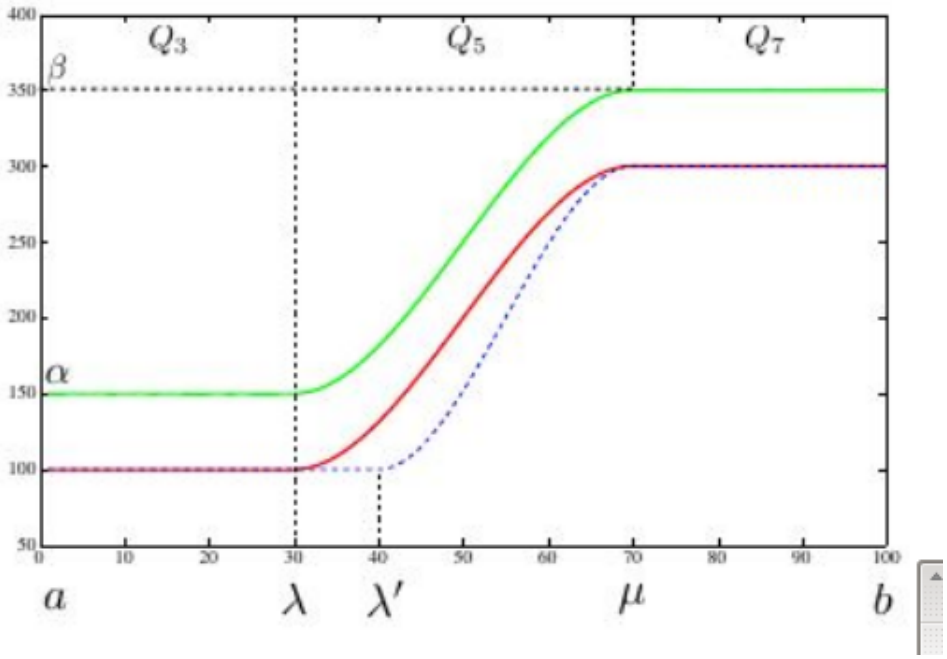
In total, 10 participants' data was collected, but only the gaze data of the top 6 participants whose data was recorded most accurately was used. Data cleaning and preprocessing was done: missing data was replaced with zeros, and it was ensured there were at least 4 usable gaze points at any instant.

2.1.2 Implementation

The cropping window for the ith frame was parametrized by the position of its center(xi,yi), and a 1:1 aspect ratio i.e. [720x720] was considered.

2.1.3 Representation: Model to fit the data

A piecewise B-Spline curve was used to represent the eye movements which consist of saccades, fixations, and smooth pursuit of the target.



In the above image, the flat segment-1 corresponds to stationary window, the smoothly varying segment is for pans, and the flat segment-2 is for the stationary window.

Model parameters

Here, (a,b) denote the first and last frame of the shot, and by changing the control points (α,β) while keeping the knots (λ, μ) same vertically shifts the curve.

Constraints

The distance between λ and μ indicates how fast the camera is allowed to pan.

$$\lambda - \mu = 40 \tag{2.1}$$

(α,β) must lie within the boundary of the original screen.

$$360 < \alpha, \beta < 920, a < \lambda, \mu < b \tag{2.2}$$

2.1.4 Procedure for every RANSAC trial

1. Pick 4 random frames for each curve A and B.
2. Initialize the parameters (α, β, λ, μ).
   (a) For curve A = [360,920,120,200]

(b) For curve B = [920,360,120,200]

3. Assign the constraints.
4. Compute the B-spline curves with the initial parameters and trial frames.
5. Using MATLAB minimization solver 'fmincon', obtain the optimal parameters by minimizing the sum of squared errors.
6. Re-compute the B-spline with the optimal parameters for all the frames.
7. Check for cuts between the two curves and obtain the final path.
8. Count the inliers within one-third of the cropping window.

To compute the final path

1. Compute the median of all the gaze points for every frame.
2. Compute the sum of distances of the median from curve A and B for the first 30 frames. Initialize the final path by selecting the curve with least error.
3. Re-compute the inliers for the final path.

Cuts

The conditions for a cut to take place at any instant are:

1. If the shift in the median across consecutive frames crosses a threshold.
2. If the two curves are more than 250 pixels apart.
3. Discard the candidate cut if another candidate cut occurs within next 30 frames and skip to the next cut.
4. Discard the final cut if the shift in the median is still closer to the current path.

2.2 Test Videos

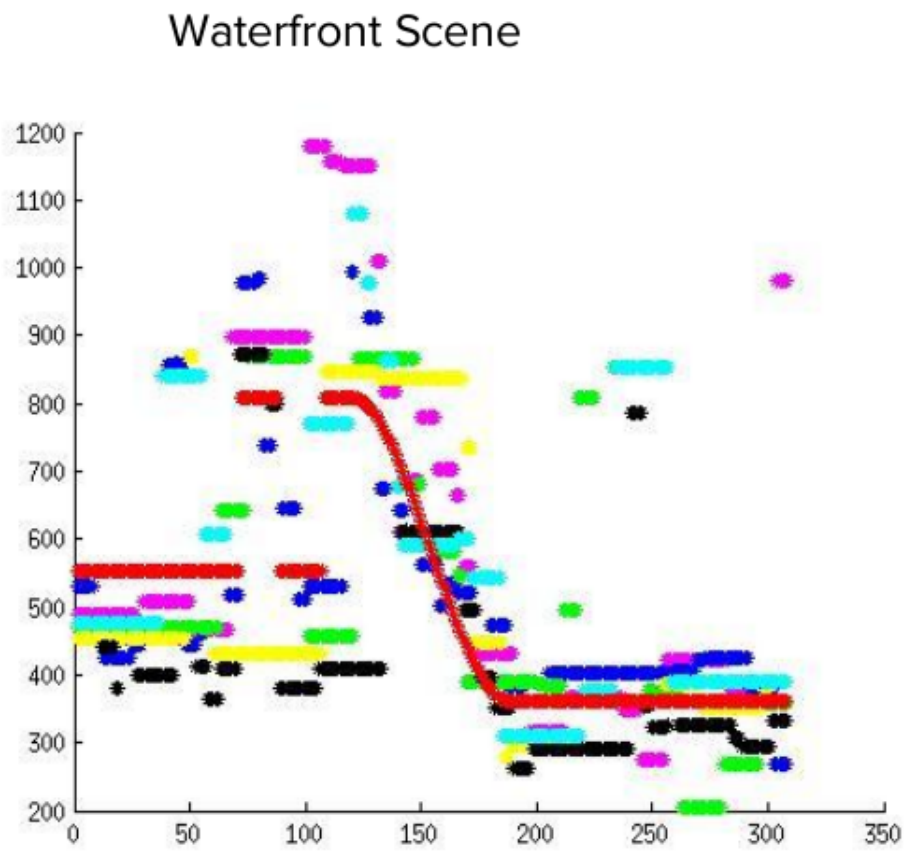
1. Herbie Rides Again (Waterfront scene) - 13 seconds
2. Herbie Rides Again (Chauffeur scene) - 13 seconds
3. Analyze This (Hotel conversation) - 15 seconds
4. Analyze This (Shooting scene) - 17 seconds

5. Batman vs Superman trailer conversation - 14 seconds
6. Fast Five (action scene) - 10 seconds
7. Transformers 3 (highway chase) - 14 seconds
8. Transformers 3 (slow motion scene) - 14 seconds
9. Brazil vs Germany 3rd goal (sports video) - 10 seconds

2.3 Results

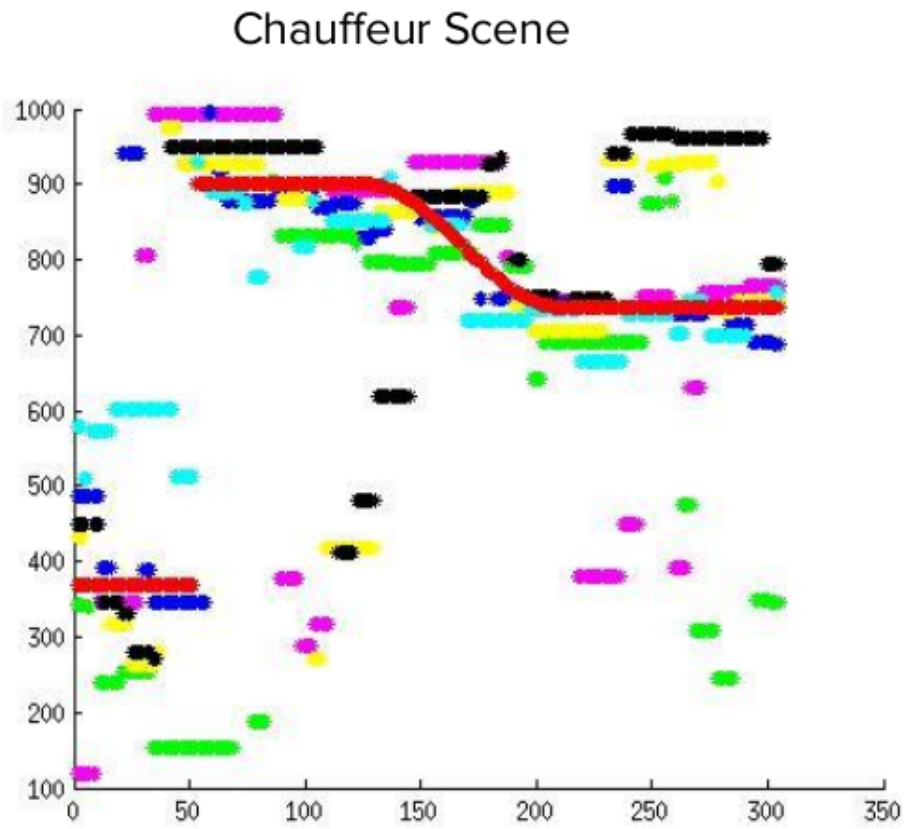
After implementing the approach described in the research paper, it was tested on a number of hand picked video clips. The following are the results that were obtained on some of the video clips that used for testing. It was not expected that the results match exactly, since the eye data used was different.

2.3.1 Waterfront Scene



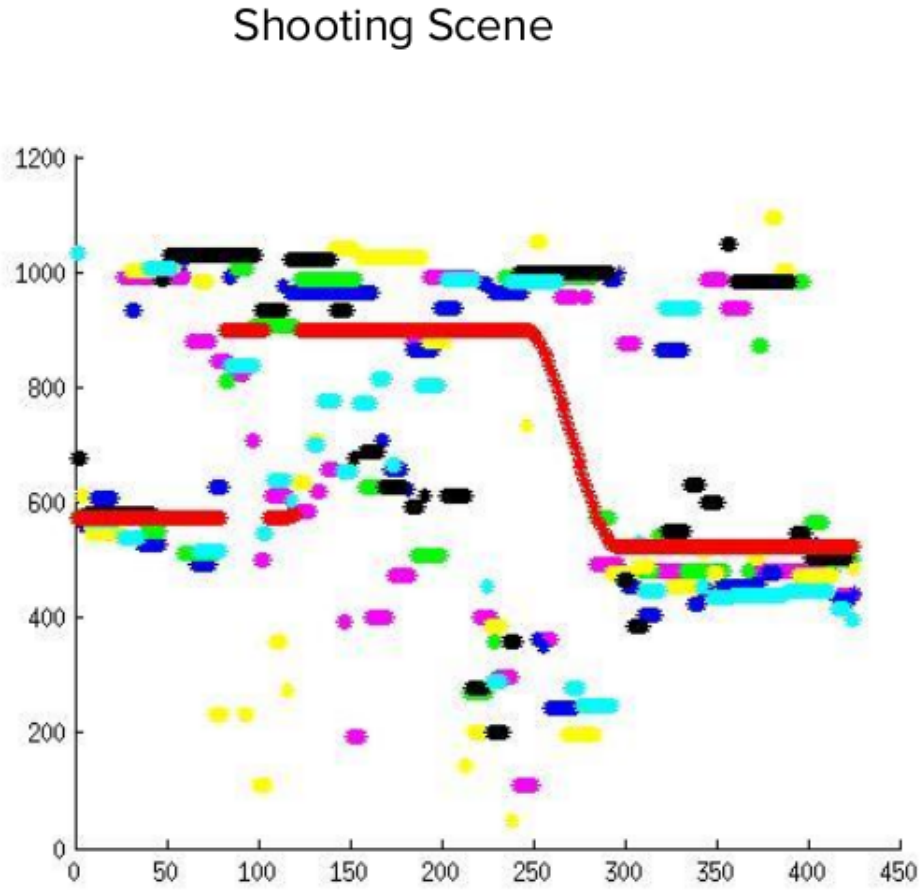
There were three cuts as compared to one depicted in the paper. The additional cut captures the movement of the man at the table (users looked at him as he looked like he was about to speak).

2.3.2 Chauffeur Scene



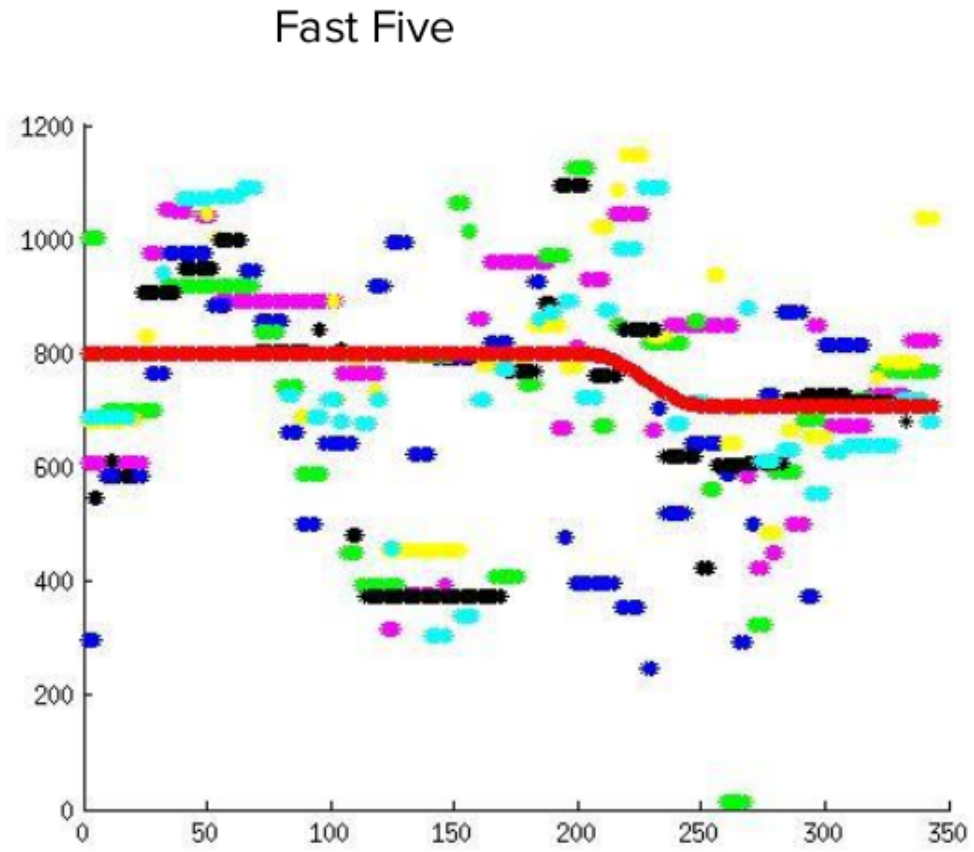
For this video clip, the cut obtained was perfectly timed, the pan was smooth, and there was no relative motion between our pan and the movement of the man, hence the video looked well edited.

2.3.3 Shooting scene



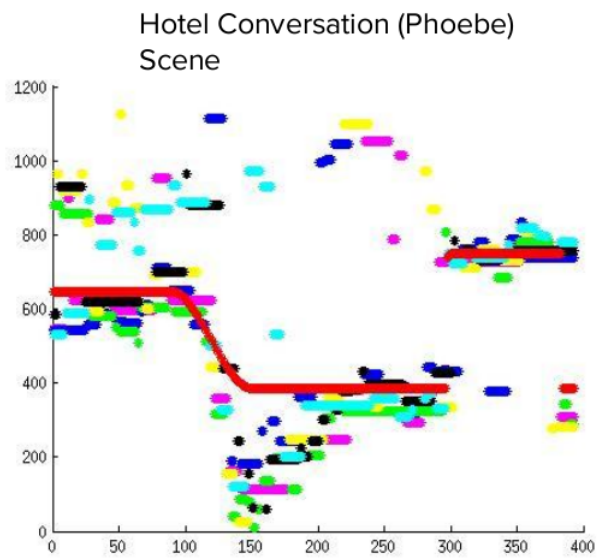
In this video clip, the gun, which was a key object, was cropped out. Also, the cut between the shooter and person shot seemed impossible to capture (too sudden).

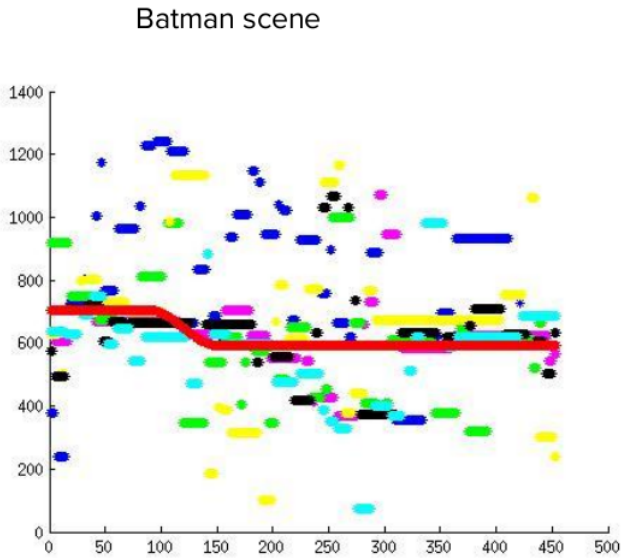
2.3.4 Fast Five Scene

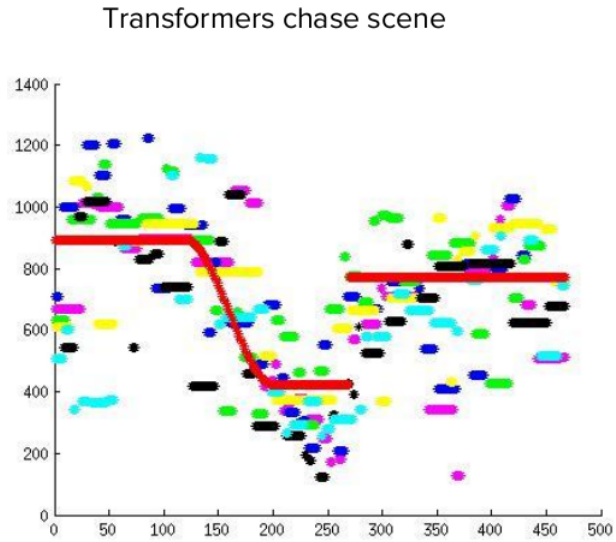


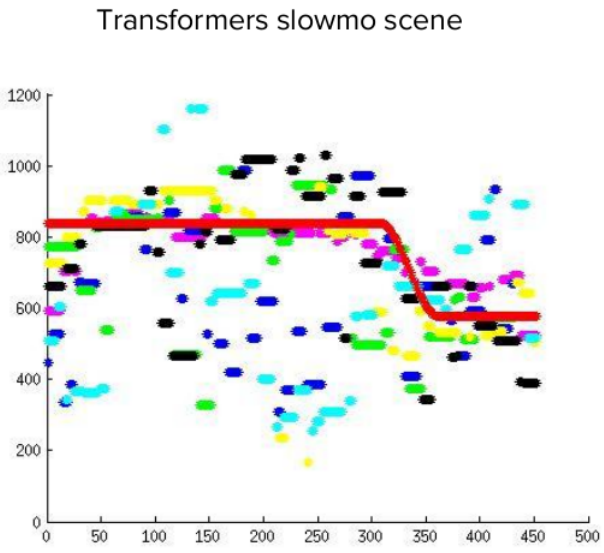
In this video clip, O'Brian's face was partially cropped as the clip already had a sudden cut from the cars to the driver, the man firing in the corner was also cropped out. The main problem noticed with this clip was that there were many outliers due to the fast movements/action.

2.3.5 Some other video clips









Chapter 3

BTP II

After evaluating the performance of the B-Spline model, it seemed that there could be at most only one pan in a single shot. In order to remove such an assumption, the focus shifted to improve the cropping window algorithm and model camera movement such as pans and cuts, as an L1 norm convex optimization problem.
The cropping window path is initialized to the median of gaze points computed for each time frame, excluding outer fence outlier points. Points where cuts can occur are identified, and each segment is optimized individually.

$$1/2 * \Sigma(x(t) - median(t)) + \lambda 1|x(t) - x(t+1)| + \lambda 2|x(t) - 2 * x(t+1) + x(t+2)|$$

$$... + \lambda 3|x(t) - 3 * x(t+1) + 3 * x(t+2) - x(t+3)| \tag{3.1}$$

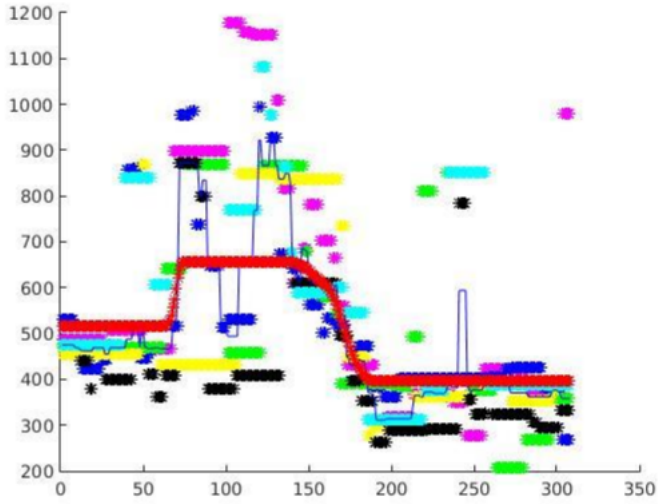
3.1 Cuts

A cut is made if it satisfies all the following conditions:

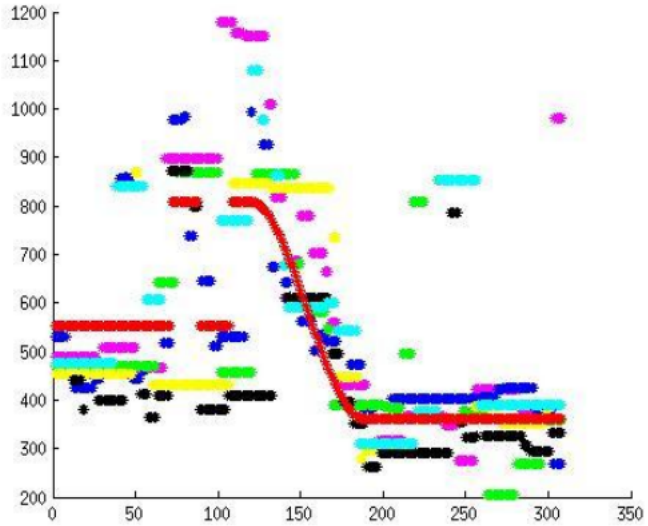
- The distance between medians of consecutive frames exceeds one third the width of the cropping window
- The acceleration between consecutive frames crosses a threshold
- The length of the segment after a cut is detect is at least as long as 2*fps before another cut occurs

3.2 L1 norm vs RANSAC

3.2.1 Waterfront Scene

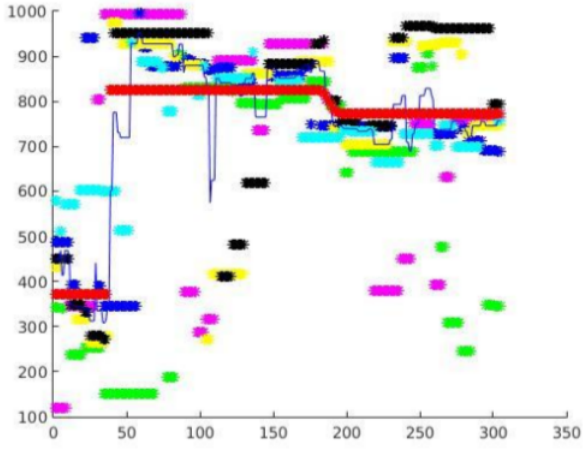


L1 Norm

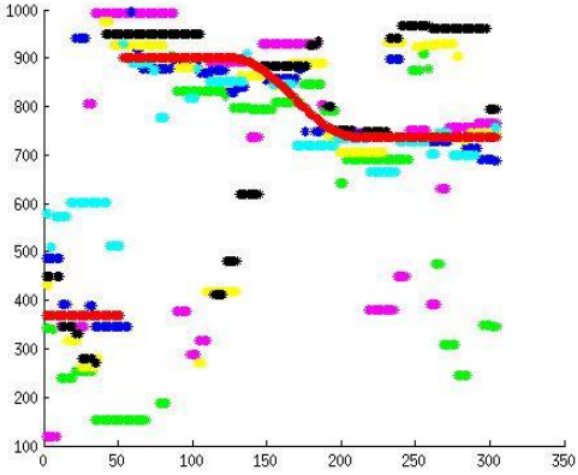


RANSAC

3.2.2 Chauffeur Scene

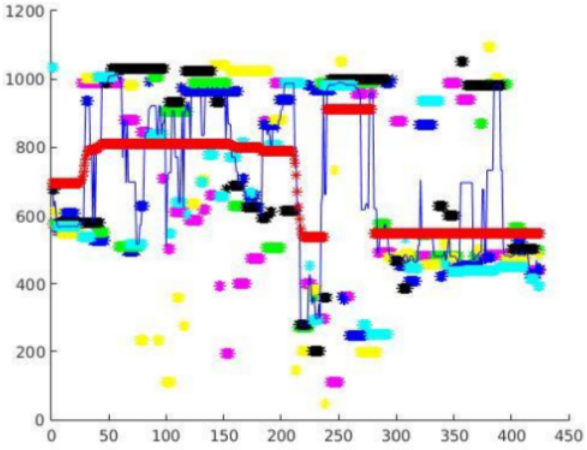


L1 Norm

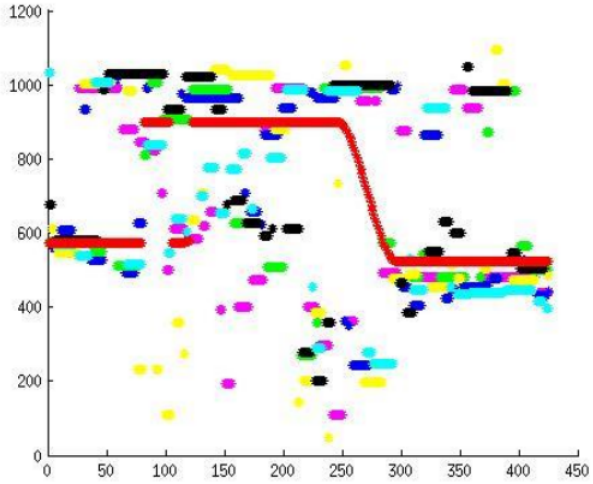


RANSAC

3.2.3 Shooting Scene

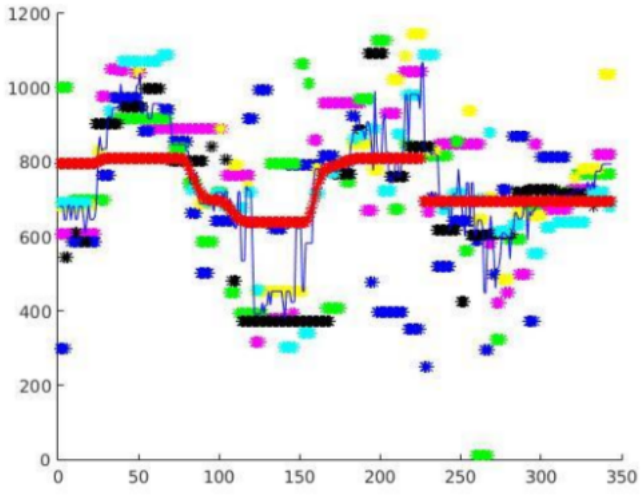


L1 Norm

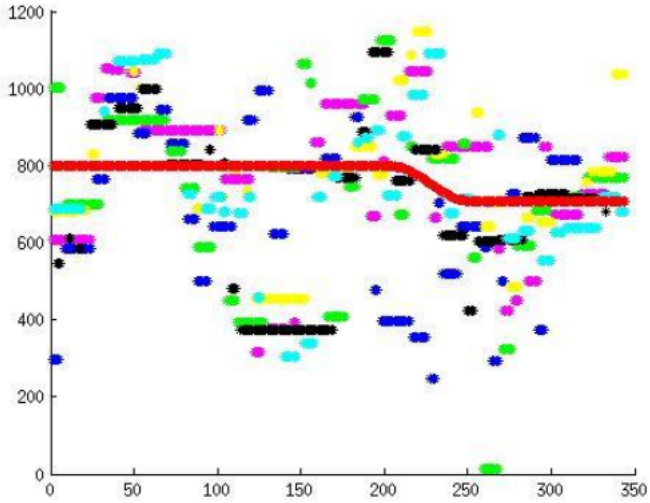


RANSAC

3.2.4 Fast Five Scene

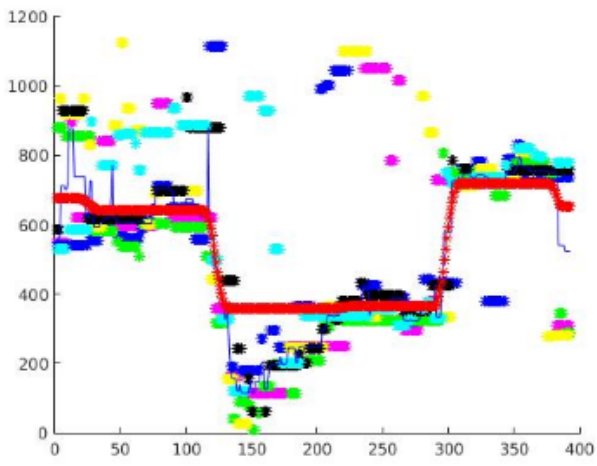


L1 Norm

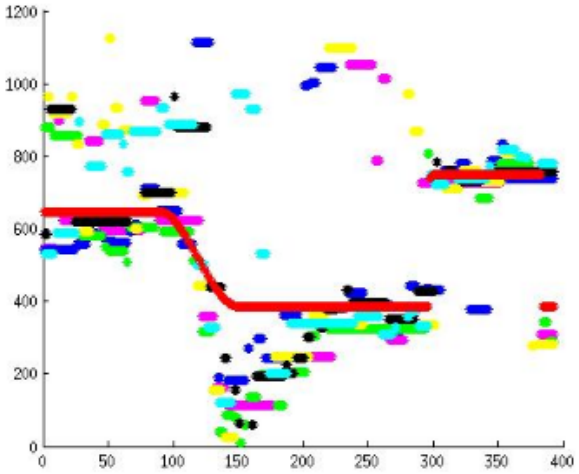


RANSAC

3.2.5 Hotel Conversation Scene



L1 Norm



RANSAC

3.3 Testing on longer videos

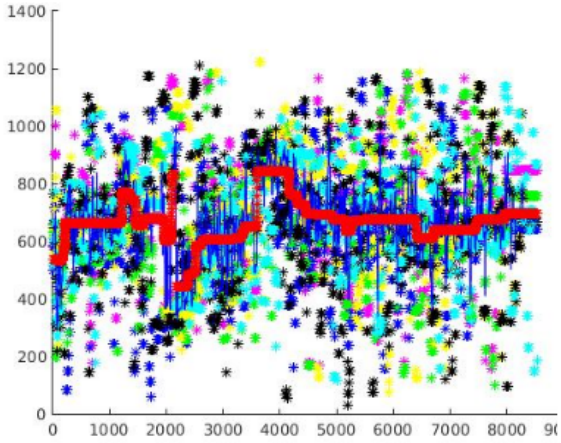
In order to make an more comprehensive analysis, we took videos which were much longer, ranging from 1 - 4 minutes. The intention was to compare how the B-Spline model fares on a sequence of shots, which may contain multiple pans. Each video was manually segmented into scenes, and B-spline was applied.

1. Matrix clip1 - 4:25s
2. Matrix clip2 - 4:02s
3. Dos3 (French theater scene) - 1:02s
4. Harry Potter scene - 4:47s

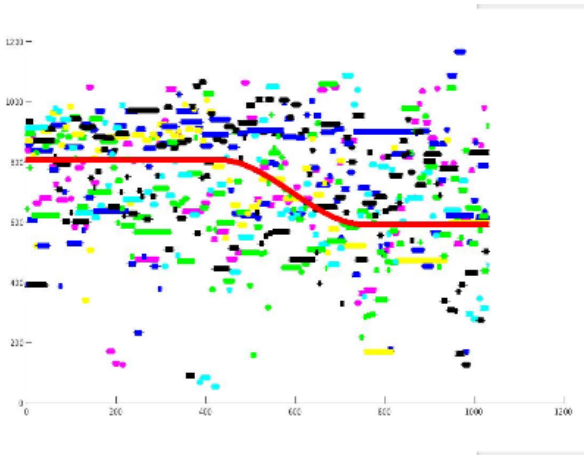
3.4 Results

3.4.1 Matrix clip1

Aspect ratio [720 x 720]



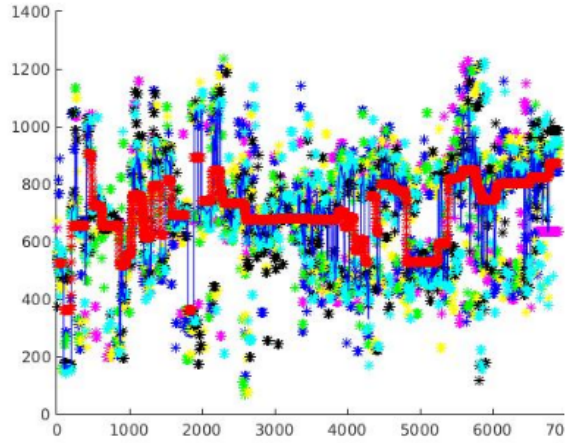
L1 Norm



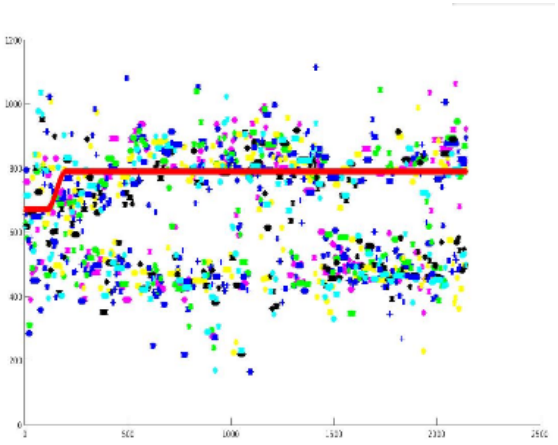
RANSAC Segment

3.4.2 Matrix clip2

Aspect ratio [530 x 530]



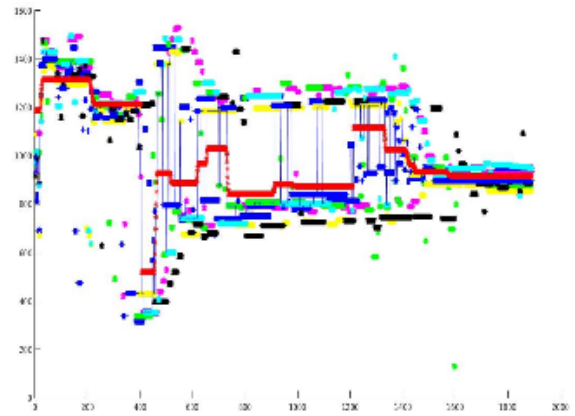
L1 Norm



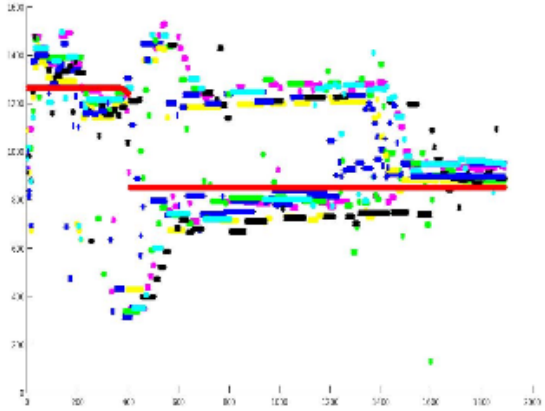
RANSAC Segment

3.4.3 Dos3 (French theater scene)

Aspect ratio [720 x 720]



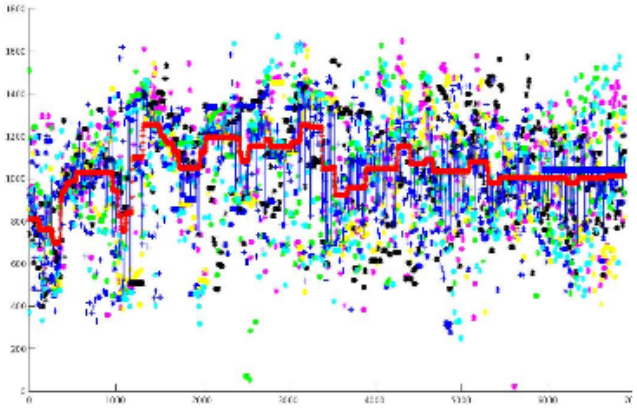
L1 Norm



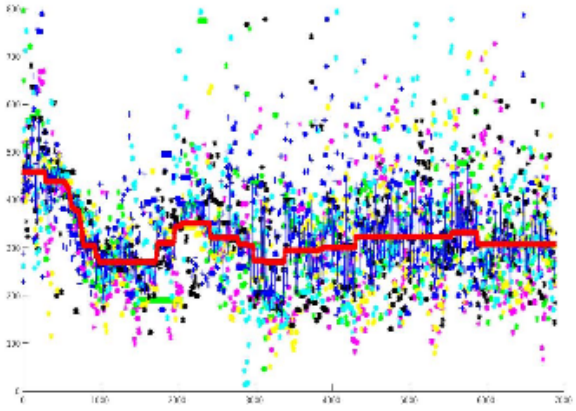
RANSAC

3.4.4 Harry Potter scene

L1 norm optimization was done along only X, aspect ratio [800 x 800]
L1 norm optimization was done along X and Y, for aspect ratio [1000 x 540]



X optimization



Y optimization

Chapter 4

Conclusions

The advantage of gaze driven editing over retargeting methods such as seam carving is that there are no distortions or aliases produced. However, the cropping window algorithm does not allow all aspect ratios to be considered, as it may partially crop out faces and other important regions.

4.1 RANSAC Algorithm

The following are the conclusions we drew about the algorithm proposed in the research paper.

1. This method necessarily models a single pan shot.
2. Works decently for static scenes, conversations.
3. For fast moving or action scenes, due to multiple stimuli, the field of view needs to be bigger.
4. Sometimes key objects are cropped out, resulting in an incomplete picture.
5. Audio may be used as a cue for editing.
6. The automated tool can't seem to capture and generalize the eye gaze patterns of every user.

4.2 L1 Norm Optimization

1. Can model multiple pans
2. If data is split between two parts in the screen, the model may not capture any one side. Instead the path settles at a point between the two ROIs and may miss both regions.

3. Optimization along both axes is possible
4. The distance between the optimal end point of one segment and optimal start point of the next segment may not be sufficiently spaced apart, though the cut conditions for medians were satisfied. Hence may cause jarred segments.

4.3 Source Code

The source code along with the eye tracking data (csv files) can be found at the github repository.

Future Work

- Focus on developing a mechanism which performs smoother, more reliable camera cuts
- Automating the creation of video sequences from director's viewpoint
- Real time video editing