Video Zooming

A Visual tracking based approach for video viewing

Rajvi Ajay Shah
MS by Research, IIIT Hyderabad
rajvi.shah@research.iiit.ac.in

Abstract— **In this project, a visual tracking based approach is used to provide zooming in videos. The purpose of zooming in an image is to observe a region of interest. Video being a temporal entity, spatial zooming doesn't serve the purpose. Due to both, camera motion and object motion, the object or region under consideration may move out of the selected region of interest in coming keyframes. To avoid this, RoI is tracked in keyframes and region of zoom is moved accordingly. Here, SIFT keypoints are used for tracking in both scenarios i.e. object motion and camera motion. Instead of using traditional techniques like LK tracker and particle filtering for point tracking, descriptor based keypoint matching is used for the purpose of tracking.**

Keywords- Video Zooming; Video cropping; Object Tracking; SIFT based tracking; Histogram of Hue.

I. INTRODUCTION

Despite the tremendous advancement in the field of computer vision and image processing, the video viewing model as well as user interface has remained more or less the same over decades. Most of the video playback and editing tools present videos in terms of frames and a timeline. This model limits the possibility of content aware interaction, authoring as well as playback. With present state of art in the field of image processing and computer vision, it is possible to redefine this present model to provide a good degree of interactivity in various possible ways.

In this project, it is aimed to explore the ways of providing zooming in a video without losing region of interest with time. Since, video is a temporal entity it is highly possible to lose the region of interest due to both object motion as well as camera motion. Such a scenario is depicted in figure 1. Region marked in blue is showing the result of spatially fixed zoom whereas the region marked in red is the result of tracked zoom.

Due to motion of object, spatially fixed zoom loses the content of interest. Same problem can arise due to camera motion or any relative motion between object and camera. To avoid that a visual tracking based approach is used. In which, the motion of region of interest is computed in every key frame and the zoom window is spatially shifted accordingly.

This can be useful in surveillance and exploration videos where the remote observer wants to concentrate on a specific target object or target area. Other than that it can be used to provide zooming and cropping tools for video editing.

Figure 1. Spatially fixed zoom vs. tracked zoom

II. FEATURE MATCHING BASED APPROACH

Visual tracking is a mature area. A lot of techniques exist for object and camera motion tracking. Instead of using traditional tracking techniques, it was decided to use feature matching based approach commonly used for image matching.

Here, two variants of features are used for the purpose of matching. Both of them use SIFT based key point detection for interest point detection [2].

The use of scale space provides a robust tracking even when region of interest grows or shrinks. The two approaches vary in the way they build descriptors. One approach uses Histogram of oriented gradients through scale space [2] whereas the other approach is color based and uses Histogram of Hue values as described in section 3.

A general framework used in both the approaches is as follows,

1. Select Region of Interest and compute the center of it
2. Compute SIFT Keypoints for Region of interest
3. Build descriptors for computed keypoints
4. While RoI exist
    a. Compute SIFT Keypoints

b. Build descriptors for computed keypoints
c. Match two sets of keypoints using cosine similarity
d. Remove the outliers based on the computed center of object
e. Find mean and occupancy of the matched keypoints
f. Find the distance offset (theta offset can also be found)
g. Move the RoI window by computed offset
h. Use new Region of Interest for matching in next iteration

While this method succeeds for camera motion, it fails to track object motion successfully. The reasons for failure were mainly background keypoints and outliers.

When object of Interest is in motion, in most of the cases its shape deforms, for example human body motion. Due to this estimating the outliers becomes difficult. Also in initial selection, not all the computed keypoints lie on the object body or foreground. If there are more keypoints in background, the mean computation as well as outlier estimation goes wrong. To avoid these problems, background subtraction can be used prior to feature extraction.

Figure 2 shows the result of method discussed above on a video, where camera is in motion. Four keyframes are taken from the video at different timestamp. Region of Interest is marked as a blue rectangle. Accuracy of tracking can be observed.

Figure 3 is a is a snapshot of a video being played in MATLAB movie player with 4x zoom. Image on top left is the keyframe from original video with blue rectangle showing region of interest.

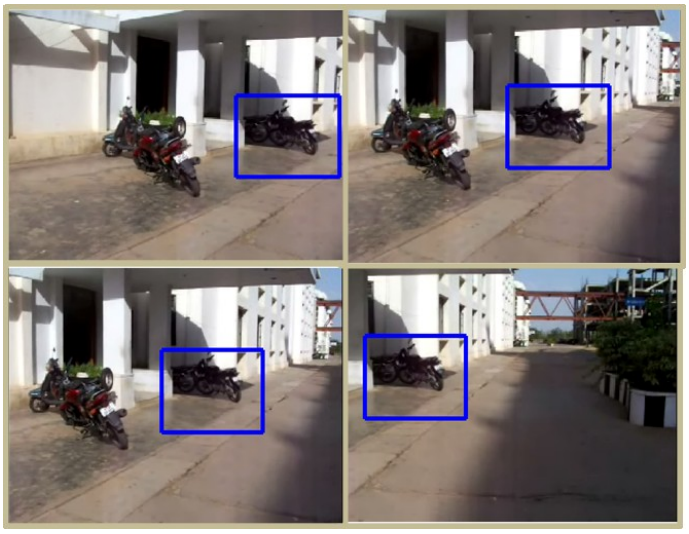

Figure 2. Four keyframes taken from the tracked video

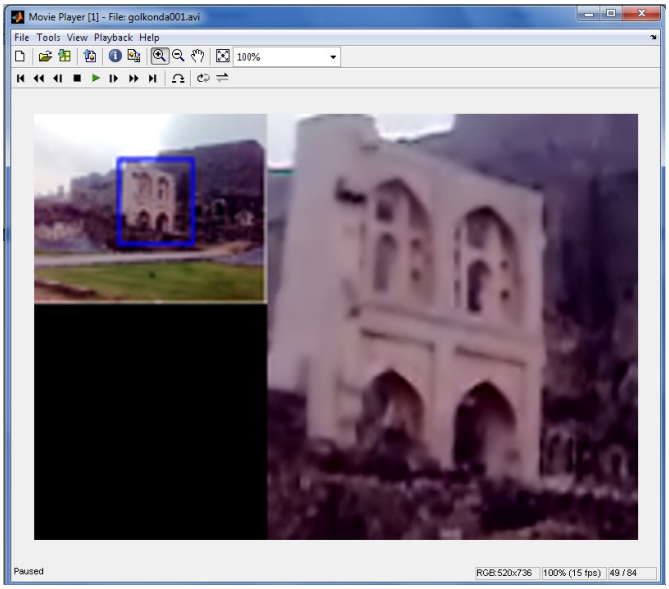

Figure 3. Snapshot of a tracked video played at 4x zoom

III. COLOR BASED DESCRIPTION

Color was a popular feature in earlier days of vision, but due to its sensitivity to illumination variation color based description are becoming obsolete for the purpose of image matching. Though it is a weak feature for image matching where illumination variations are highly expected, it can turn out to be useful for visual tracking in videos.

The fact that in most natural videos, expected illumination variation is very less was the primary motivation for building color based description. Descriptor computation is both simple and efficient and hence an attractive option for real-time applications.

The descriptor can be computed as described below,

1. Represent the key frames in HSV color space
2. Construct scale space of H and V planes
3. Use 'V' plane for interest point detection through scale space
4. For all the detected key points, choose appropriate scale of H plane from the scale of detected keypoint
5. Select a neighborhood of σ2x NxN around location of keypoint
6. Build a Histogram of Hue values of 128 bins the selected neighborhood
7. Normalize the Histogram

This gives us a 128 dimension long feature vector which can be used for matching keypoints.

The following figures show few of the matched key points described by Hue Histograms in H and V planes, as we can see the matching is very efficient.

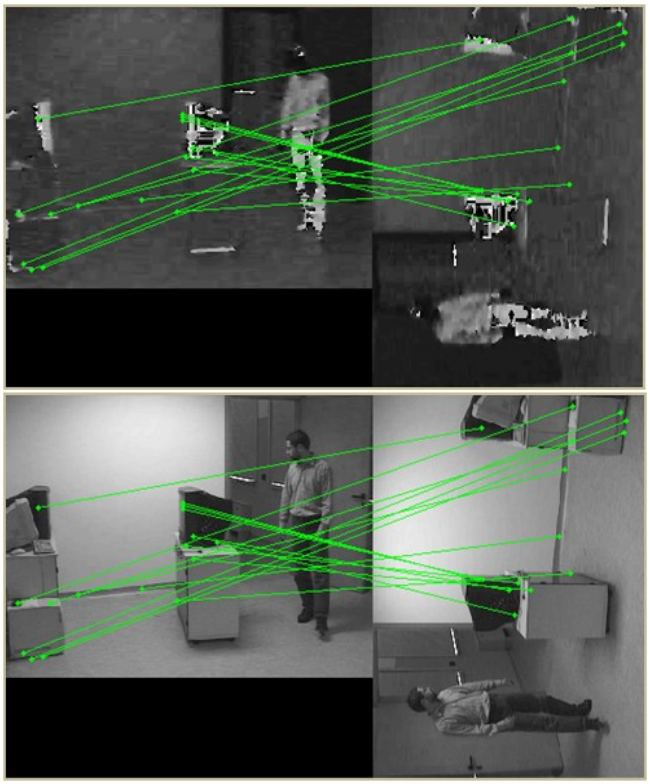

Figure 4. Matched Keypoints in Top and bottom image in H and V planes

IV. Conclusion

It can be concluded that the concept of video zooming based on visual tracking can be realized in practice. Also features used for matching prove useful for the purpose of tracking. With a better model fitting for outlier removal and, the explained approach can be used for general videos having both object and camera motion. Hue histogram based descriptor gives promising results and can be explored upon to yield a faster approach. The explored concept has many potential applications for video authoring as well as video surveillance and exploration.

References

[1] Daniel R. Goldman, A Framework for Video Annotation, Visualization, and Interaction. PhD Thesis, University of Washington, 2007.

[2] David G. Lowe, **"Distinctive image features from scale-invariant keypoints,"** *International Journal of Computer Vision,* 60, 2 (2004), pp. 91-110J.

[3] Yilmaz, A., Javed, O., and Shah, M. 2006. Object tracking: A survey. *ACM Comput. Surv.* 38, 4 (Dec. 2006), 13. DOI= http://doi.acm.org/10.1145/1177352.1177355

[4] Robert Collins, Mean-shift Object Tracking through scale-space. *IEEE Conference on Computer Vision and Pattern Recognition*, 2003.

[5] Jianbo Shi and Carlo Tomasi. Good Features to Track. *IEEE Conference on Computer Vision and Pattern Recognition*, pages 593-600, 1994.

[6] Andrea Vedaldi's implementation of SIFT: http://www.vlfeat.org/