Smart 360 Pilot: Automatic 360° degree video cinematography agent

Jaskaran Singh, Kartik Mathur, Aniket Kumar, Aryan Bansal

Thapar Institute of Engineering & Technology, Patiala - 147003, INDIA

jsingh11_be19@thapar.edu, kmathur_be19@thapar.edu, akumar2_be19@thapar.edu, abansal1_be19@thapar.edu

***Abstract*— With the continuous advancement of multimedia technology, 360° videos are becoming widely popular. These videos allow the viewer to immersively view content in a manner that appears three-dimensional. But there are many limitations to viewing videos in this manner such as the requirement of specialized equipment like VR Headsets. Another requirement is that the viewer must constantly adjust their field of view, either by clicking mouse buttons or by adjusting the position of their head. In case the viewer does not have the special equipment, the view appears warped and is generally a dissatisfactory experience. To tackle these issues, the proposed method should take a 360° video and using computer vision, automatically predict the key area of interest in the video. Next, it should use a de-warping algorithm to map and flatten the given section into an N-FOV (normal field of view) video output. The video generated by this method should appear as a normal two-dimensional stream whilst preserving as much information as possible. As opposed to other methods that use deep learning (hence, being computation intensive). The method proposed will use the notion that areas of interest are usually concentrated to a small section in 360° videos. This means the proposed method is by virtue, efficient and has reasonable computational resource requirements. The performance of the method is evaluated by the time it takes to generate video outputs as well as the qualitative measure of these outputs. The experimental results have sufficiently established the efficiency and quality of the proposed method.**

Keywords— 360° video piloting, De-warping panoramic video, computer vision, video cascading, N-FOV video generation

I. INTRODUCTION

The surge of advancements in the field of multimedia technology have made it possible for media formats such as 360° panoramic videos to become widespread and popular. While these types of videos offer an immersive and futuristic three-dimensional viewing experience, they also come with a host of limitations and requirements. Physical devices such as VR headsets or computing devices with specialized sensors like gyroscopes are required for a good consumption experience with 360° videos. Another limitation is that due to the way these videos are rendered in three-dimensional space in virtual reality, they require the viewer to constantly adjust their head so that they can follow the area of interest in the video, which usually is a small and localize area in the whole 360° space. To tackle all these limitations and provide a generalized two dimensional video viewing experience for these videos, in this document we have proposed a method that automatically extracts the location of the area of interest in the video and uses a de-warping algorithm to map it onto a flattened two-dimensional space while trying to retain as much information as possible while also aiming to not let the resulting video become uncanny.

The methods already proposed in other publications such as by Hou-Ning Hu *et al.*[1] are sufficiently capable of producing the required results, but due to their use of deep learning they have computation intensive implementation, often requiring dedicated gpu systems to function. Most implementations also require a linux operating environment as well as large training datasets to produce their deep learning based models.

The proposed method requires minimal computation power, is efficient and produces results in reasonable time. This is possible because the proposed method uses video cascading features available in Python's OpenCV library. The video stream is split into its component frames (which are images). The detection filter, comprising of video cascading detects the key area of interest in the first frame and then again every 8th frame in the stream of frames. This is based on the fact that digital video formats usually run at 24 frames per second. This means each second, the key area is recognized 3 times, which is practically sufficient for smooth transitions in case of change in key area. Once the key area is recognized, the de-warping algorithm recognizes the co-responding viewport and maps the viewport onto a flat two-dimensional space (N-FOV). Lastly, the resulting stream of resulting frames is compiled into a video format.

The rest of the paper is organized as follows: Section II consists of descriptions of related work, its key achievements and limitations. Section III explains the proposed method in detail. The experimental results of the proposed method are given in Section IV. Finally, conclusion of the work is given in Section V.

II. RELATED WORK

A brief analysis of the work related to the field of 360° video cinematography is given in this section. The work related to piloting of Sports centered 360° videos is primarily described in Hong-Ning Hu *et al* [1] and the work related to conversion of Panoramic videos to N-FOV videos is primarily described in Yu-Chuan Su *et al.* [5]. These two publications are the primary references upon which the methods proposed in this literature are based. Other supplementary references have been derived from other publications such as Duc V. Nguye *et al.* [2] which describes the tile based approach used in the proposed method.

Hong-Ning Hu *et al* [1] have proposed a method with a deep-learning based agent that pilots through 360◦ sports videos automatically. For each frame, the agent observes the panoramic image whilst having knowledge of previously selected viewing angles. The task of the agent is to shift the current viewing angle to the next predicted one. Specifically, it leverages an object detector to propose candidate objects of interest. Then, a neural network (CNN) is used to select the primary object. Given the primary object and previously selected viewing angles, the method regresses a shift in viewing angle to move to the next one. It uses the policy

gradient method to train the pipeline, by minimizing: (1) regression loss measuring the gap between the selected and ground viewing angles, (2)smoothness loss, encouraging smooth transitions between viewing angles, and (3) maximizing an expected reward of focusing on a foreground object. These techniques allow the method to provide a deep-learning based, activity aware 360° piloting agent.

In this method, a major drawback exists in the form of the requirement of a dedicated NVIDIA GPU with CUDA cores to run the program, which is further supplemented by the fact that the implementation can only run in a Linux operating environment.

Yu-Chuan Su *et al.* [5] have proposed a method where given a 360° video, the goal is to direct a virtual camera to capture a natural normal field-of-view (N-FOV) video. By selecting "where to look" within the panorama video at each frame. In this method, Yu-Chuan Su *et al.* [5] first compile a dataset of 360° videos from the internet, along with human-edited NFOV camera trajectories to facilitate evaluation. Next, they propose a data-driven approach to solve the piloting problem. Their method uses N-FOV videos to discriminatively identify space measures of "glimpses" of interest at each time instant, and then uses dynamic programming to select optimal human-esque camera trajectories. Along these trajectories, they compile a flattened N-FOV video.

The primary drawback of this model is that it is data intensive. To train a sufficiently acceptable ML model, the training data requirements are extensive and may be a hurdle to obtain.

III. PROPOSED METHOD

The primary focus of this paper is to design and develop a piloting agent for automatic cinematography of 360° panoramic videos into a N-FOV two-dimensional video. This will allow a viewer to view a panoramic video as if it were any other regular video, where the agent has automatically selected the key area of interest. This process can briefly be divided into 4 major steps. A panoramic video is first divided into a set of its component frames. Then on each 8th frame, a detection filter is apply to predict the key area of interest, this result is in the form of a specific region in the equirectangular frame. This region is passed onto the de-warping algorithm. This algorithm uses the calculated angles of z-axis and y-axis to convert the key area of frame into a normal perspective view. Now, that the frames are flattened, all the resulting frames can be merged back to form an N-FOV regular video.

Step 1: Any video is essentially a stream of image frames. So we will first split the input video into its component frames and store these frames in sequential order. Most digital videos run at 24 frames per second, so each second of video will generate 24 frames.

Mathematically, a set 'S' which is a video, where *"Fi"* is any frame in the video can be represented as:

$$S = \left\{ \sum_{i=1}^{n} F_i \right\} \tag{1}$$

Step 2: Now, we go through each frame one by one, starting with the first frame. On every frame number that is a

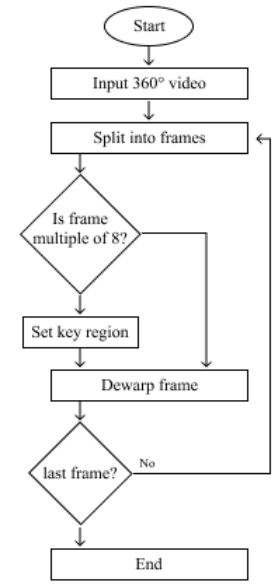

Fig 1: Flow chart of the proposed model

multiple of 8, we will apply our detection filter. This filter will use video cascading to determine key areas of interest. For example, in case it detects a face, it will determine the location of the face and output it for the algorithm in the next step to use for de-warping. This video cascading algorithm uses a cascading file such as Haar Cascade to determine these key locations.

$$x, y = Filter(F_n) \tag{2}$$

Step 3: Using the key location co-ordinates generated every 8th frame, we will now run a de-warping algorithm on each frame. The co-ordinates of key area from the previous step are taken. The 360° field of view is divided into " n" number of sections, for eg: 3, so that there are 3 sections each of 120°. The panoramic equirectangular image's width is also divided into same number of sections. Say the image width is 300 pixels. There will be 3 sections of 100 pixels in the image whose pixels range from pixel 1 to pixel 300. Now, lets say the x co-ordinate of key area is at 50th pixel.

Using this information, we can easily deduce that the key area lies in section 1 of the panoramic video. So the angle of field of view will be -120° (left is negative, right is positive) The same can be done for the y co-ordinate of the key area, and another angle can be obtained.Let the angles of field of view of the viewport be ϕ and θ respectively. Our algorithm now knows the area from where it has to crop, de-warp and project the co-responding pixels onto a flat two-dimensional space. Each de-warped frame is stored.

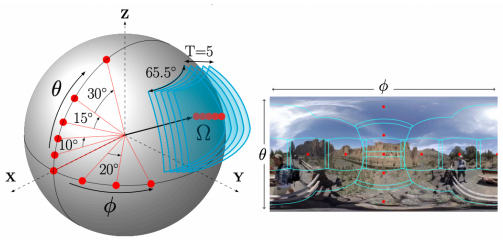

Fig 2: Placement of ϕ and θ angles

Fig 3: A frame from the input 360° video

Fig 4: Key area selection using the processes

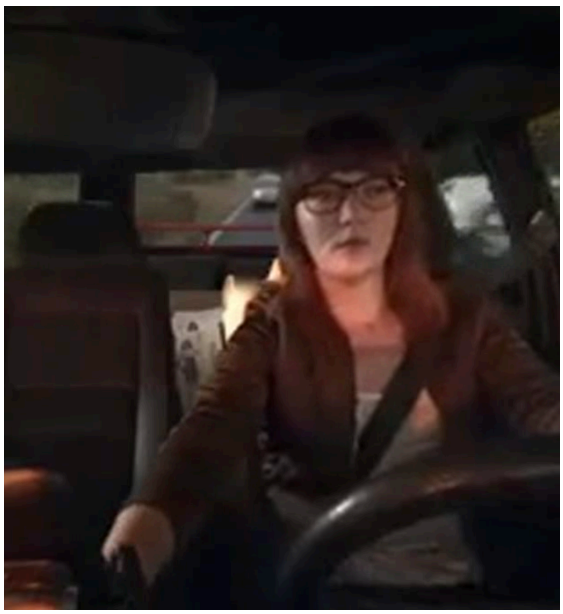

Fig 5: The de-warped and cropped output

Step 4: Now that each de-warped frame has been stored. These frames are combined once again at the initial frame rate to produce the required video. The process is now complete.

IV. EXPERIMENTAL RESULT

The performance of the proposed method is evaluated qualitatively as well as quantitatively and compared with the existing methods. The experiments have been performed on a machine running Windows 10 operating system, 8 GB of RAM & an Intel Core-i5 processor. The software and libraries used are Python 3.9.8 and OpenCV 3. The method was tested by using 11 360° panoramic equirectangular videos comprising of various real life sports situations like surfing, boating, kayaking, cycling etc. as well as different indoor and outdoor situations. Based on the experimentation highest accuracy was achieved in outdoor situations where key areas of movement are limited and there is sufficient lighting. Indoor situations with good lighting conditions also showed promising accuracy.

The snapshots of various steps of the proposed algorithm are given in Fig[3-5] where a person can be seen in a driving situation, the cropped output snapshot shows the final output produced after application of the proposed method onto the input video.

A. Quantitative Performance

As apparent from the snapshots attached the method was :

(1) Successful in separating a component frame out of the video stream.

(2) Successful in appropriately identifying the key area of interest in the panoramic equirectangular sphere.

(3) Successful in cropping the required region onto a flat N-FOV two-dimensional space.

We must also note an apparent degrade in output image quality, but this can be attributed to the fact that panoramic videos have to fit approximately five times data of a regular video into the same 1920x1080 pixel video frame, So, when the cropped image is mapped onto a larger image frame, it will always show degradation of quality to some degree.

To summarize, the quantitative performance of the proposed method is sufficiently acceptable.

B. Qualitative Performance

Qualitative results have been discussed with the experiment above, in this section the quantitative analysis of the proposed method is presented. In order to perform the evaluation of the automatic cinematography results, a numerical accuracies are computed. These accuracies focus on the ability of the proposed method to identify key areas of interest in a given panoramic video.

The parameters which are used for the quantitative calculations are defined as follows:
True Positive (TP): the number of times correctly predicted key area; False Positive (FP): the number of times incorrectly predicted key area; False Negative (FN): the number of times no key area was predicted. On the basis of these 3 parameters, three measures are computed as follows :

$$Correctness = \left(\frac{TP}{TP + FP}\right) X 100 \quad (3)$$

$$Completeness = \left(\frac{TP}{TP + FN}\right) X 100 \quad (4)$$

$$Quality = \left(\frac{TP}{TP + FP + FN} \right) X100 \quad (5)$$

On the basis of these measures of accuracy, the results from the proposed method have been compiled using 11 experiments on various panoramic videos and tabulated in Table 1.

Table 1: Accuracy of proposed method

Input Video						
	TP	***FP***	***FN***	***Correctness***	***Completeness***	***Quality***
Indoor Video	3	1	0	75%	100%	75%
Outdoor Videos	5	1	1	83.33%	83.33%	71.4%

V. CONCLUSION

This paper hence provides a sufficiently acceptable and efficient model for the creation of an automatic piloting agent for cinematography of 360° panoramic videos. The method can detect key areas of interest on a per-frame basis in an equirectangular frame and convert it into the corresponding panoramic frame. The method is neither computation intensive nor does it require specialized computing equipment. The results are produced within reasonable run time and are of acceptable accuracy across a wide range of videos. However, the accuracy of the proposed method is limited by the quality of the input video and the proposed method may not produce a smooth transitioning output in case of extremely dynamic video cases. The proposed method is hence more suitable for less dynamic and stable panoramic videos whilst being efficient, practical and smart.

REFERENCES

1. Hou-Ning Hu and Yen-Chen Lin and Ming-Yu Liu and Hsien-Tzu Cheng and Yung-Ju Chang and Min Sun *"Deep 360 Pilot: Learning a Deep Agent for Piloting through 360° Sports Video"* Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR) - 2017
2. Duc V. Nguyen, Huyen T Tran , Truong Cong Thang, "*An Evaluation of Tile Selection Methods for Viewport-Adaptive Streaming of 360-Degree Video*" *,* The University of Aizu - 2020
3. Y. Bao, H. Wu, T. Zhang, A. A. Ramli, and X. Liu. 2016. *"Shooting a moving target: Motion-prediction-based transmission for 360-degree videos".* In IEEE Big Data. Washington, DC, 1161--1170.
4. P. R. Alface, J. Macq, and N. Verzijp. 2012. *"Interactive omnidirectional video delivery: A bandwidth-effective approach".* Bell Labs Technical Journal 16, 4 (March 2012), 135--147
5. Yu-Chuan Su, Dinesh Jayaraman, and Kristen Grauman, *"Pano2Vid: Automatic Cinematography for Watching 360◦ Videos*". The University of Texas at Austin - 2017
6. Mai Xu*, Senior Member, IEEE, Yuhang Song*, Jianyi Wang, Minglang Qiao, Liangyu Huo and Zulin Wang, "*Predicting Head Movement in Panoramic Video: A Deep Reinforcement Learning Approach*" Journal of LaTex Class Files - 2019
7. M. Xu, C. Li, Y. Liu, X. Deng, and J. Lu, "*A subjective visual quality assessment method of panoramic videos*," in ICME, July 2017, pp. 517– 522.
8. Y. S. de la Fuente, R. Skupin, and T. Schierl, "*Video processing for panoramic streaming using hevc and its scalable extensions,*" Multimedia Tools and Applications, pp. 1–29, 2016.