International Journal of Scientific Research in Engineering and Management (IJSREM)
Volume: 04 Issue: 09 | Sept -2020 ISSN: 2582-3930

Identifying Head Movements in 360-degree video: A Convolutional Neural Network Approach

U.R Kulkarni 1 , S.A Patne 2 , S.B Kardile 3 , V.S Mane 4

Department of Computer Engineering, Dr. D.Y Patil Institute of Engineering and Technology, Ambi, Pune

Abstract - Panoramic videos being an interactive media are gaining a lot of popularity. They give a horizontal elongated view where the Field of View (FoV) in the range of $360^o \times 180^o$. FoVs are important to capture the material to be included in the video sequences. The part outside FoV range is not included in the video sequences. This is controlled by human head movements (HM) performed by wearing the head mount display.HM prediction is at extreme level of importance and thus needs attention as it determines the eye fixations too. Thus, head movements and eye fixations are a vital part for perfect capturing of the material in the FoV range. There are works related to this comprising of the deep learning approaches, reinforcement learning and more.The paper establishes a database with the help of face detection and camera. Collecting this database, we will use the convolutional neural networking approach to capture the environment in the FoV.

Key Words: Head movement, Panoramic videos, CNN, Haar cascade.

1. INTRODUCTION

Real time connectivity and communication is a vital part of now-a-days. Panoramic videos are one such media to share the $360^o \times 180^o$ environment view. This type of video capturing needs a perfect hold over the Field of View. The environment outside the Fovs cannot be captured by the subject. This field of view is controlled by controlling the head movements of the subject. Thus, the head movements of the subject determine the Field of View and hence head movements become very much important while deploying attention on the videos. And here arises the most important problem in modeling human focus on panoramic videos. But, seeing in detail these head movement or generally the human attention while focusing on the videos is comprised of the two things: the actual head movements and eye fixations. Eye fixations focus on the high-quality regions to be captured (i.e. the finest or high aimed regions) within the Field of View.

There are some of the approaches proposing the systems for simulating the human attention on panoramic videos and predicting the head movements. Advancing the works from the recent paper [ref no] using the deep reinforcement learning (DRL) approach, the proposed approach uses the convolutional neural network (CNN) approach rather than the DRL approach.It is because of the automatic detection feature of the algorithm. CNN when compared to many of the older algorithms has a capability to automatically detect the important features. The important thing is it does not need any human attention or human supervision. It is considered as the efficient and powerful machine learning algorithm.The engaging experience through the use of panoramic videos can be gained by wearing the head mount displays and moving the head in the space. The database of the imagery can be collected by detecting the face through c

camera where the environment will be captured through the head movements of the subject. The camera will be initialized to detect the face of the subject. Along with the face eyes and nose detection will also be done to detect the exact facial features. Detection phase will continue unless and until the video capturing is completedof the subject.

A control unit will be setup for the re-detection of the subject. An eye tracker and an eye movement detector will perform eye localization and the gaze estimations together once the features are extracted from the subject. The computer vision process tounderstand the detailed manipulation and retrieval of data from the subject is applied once the face detection is completed. For the computer vision process a library in python known as OpenCV is used which an open source library used for machine learning and image processing which is the vital step in database collection of the system proposed. The frameworks Keras and TensorFlow will be used for better extensibility and modularity of the system.

Once the head localization and pose estimation along with detection of the eyes and nose is done then the convolutional neural network (CNN) approach is applied to capture.The frame will be cropped according to the height and width specifications along the X-Y co-ordinates.The convolutional layers extract the features from the input given and the dense layers further which are fully connected uses the data from the layers itself to generate the outputs. As the head of the subject moves the surroundingenvironment will be captured through a camera device and the video will be recorded. To our best knowledge the previous works done till date are a bit on costlier side whereas this mechanism will be truly a life

saver in terms of the cost parameter. We use a simple camera device to capture the video in replacement of the head mounts or the VR's which can lead to an increased cost of the system up to 7 times greater.

2. Related Work

2.1 Predicting head movements

1) Mai Xu. proposed a deep reinforcement learning approach which controlled the Field of View with the help of reward estimations.This is done to predict the head movement scan paths which can be used to check the long-term effect of the behavior of the head movements.

2) More specifically the approach is classified into two types: offline and online head movementprediction. The offline head movement prediction is used to attain human attention of multiple subjects on the panoramic video. Whereas, the online head movement prediction is used to predict the future HM positions of a single subject.

3) Jianyi Wang. has applied the reinforcement learning approach to increase the accumulated reward of the agents' action. He has presented the system by predicting the scan paths of multiple agents which finally yields the HM maps of panoramic videos. The idea behind this is to model human attention based on the past observations which is important in establishing human like computer vision.

2.2 Human eye fixations

1) A CNN approach is applied to detect the eye movements on the images. Saliency detection based on location is the approach used in this system.

2) A deep approach can be seen practically in this paper. Normalized scan paths are the main content to fix the eyesight on the images can be learnt by this study.

3. Motivation

To implement head movement prediction in 360-degree videos by using the convolutional neural network which will help to capture the video with the help of face detection. The motive is to reduce the complexity in data capturing faced during the head movement controlling and also making it more cost friendly and efficient.

4. System Implementation

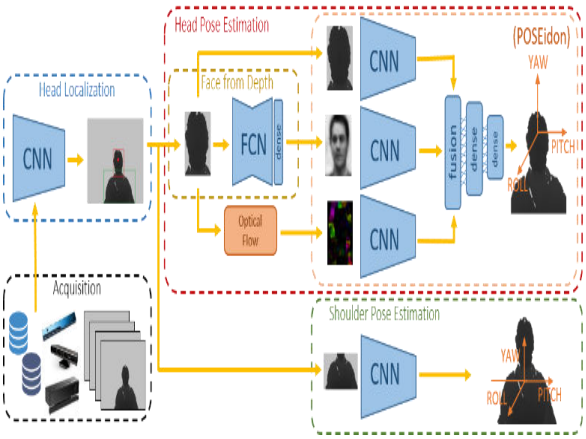
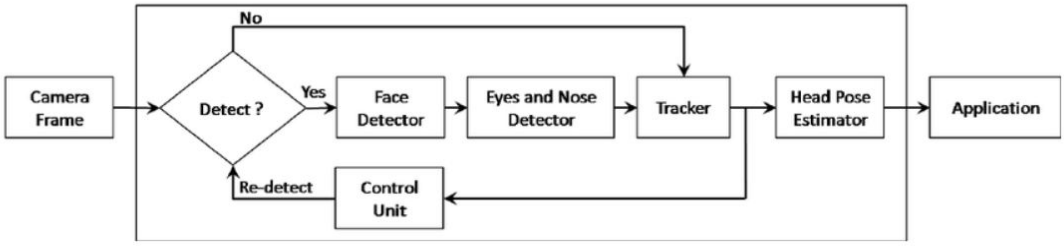


Fig.1 System Architecture

The main objective of the system is to interpret the head movements of the subject to capture the images of the surrounding environment.The above diagram (fig.1) shows the working of head movement prediction system. The first step towards the system implementation is acquisition of the database. This database is collected through detection of head, eye and nose pose estimation. First the camera is initialized and connected. The frame is captured with the camera. Once, the frame is captured the face detection exactly starts at this point of execution. The face is detected with the help of important feature extractor implementation of the system established. The nose and eyes are then detected to get the exact feature combinations of the subject capturing the frame with the head movements. Gaze tracking is carried out at this point of execution to check the eye stimulus for proper and efficient control over the head movements. User will gaze at the screen or the target scene and the gaze estimation will be taken into consideration. Gaze estimation will take keen details of the face which makes the system easy, efficient and cost friendly in terms of database collection of the images. This facial feature detected continuously throughout the execution. After this, the templates are rendered and finally the contents are produced. This rendering is done in some broader aspects as the frame of the face and the database is not only captured but the co-ordinate axes X, Y, Z are manipulated too. The simultaneous calculations of the axes co-ordinates are taking place to cope up with the 3D plane. All the sequenced set of data is passed through the dense layers of the CNN.



Camera calibration and Pose estimation

1) Determine camera parameters from known 3D points or calibration object(s).

2) These parameters are intrinsic and extrinsic. Intrinsic parameters include parameters like focal length, optical center, aspect ratio. The extrinsic parameters are all the pose parameters.

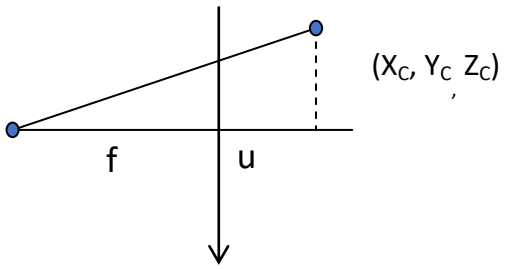
3) How can we do this?

There are view approaches to do this like linear regression, nonlinear optimization but for the panoramic or the 360-degree video capturing we prefer the rotational motion of the panoramas.

4) Image formation equations are

$$\begin{bmatrix} X_c \\ Y_c \\ Z_c \end{bmatrix} = \begin{bmatrix} \mathbf{R} \end{bmatrix}_{3\times3} \begin{bmatrix} X \\ Y \\ Z \end{bmatrix} + \mathrm{t}$$

$$\begin{bmatrix} u \\ v \\ 1 \end{bmatrix} \sim \begin{bmatrix} U \\ V \\ W \end{bmatrix} = \begin{bmatrix} f & 0 & u_c \\ 0 & f & v_c \\ 0 & 0 & 1 \end{bmatrix} \begin{bmatrix} X_c \\ Y_c \\ Z_c \end{bmatrix}$$



Here, X, Y, Z are the co-ordinates of the facial plane in 3D.

We need to draw the 3D co-ordinates with the help of the axes (X, Y, Z)

5) The calibration matrix is,

$$\begin{bmatrix} u \\ v \\ 1 \end{bmatrix} \sim \begin{bmatrix} f & 0 & u_c \\ 0 & f & v_c \\ 0 & 0 & 1 \end{bmatrix} \begin{bmatrix} X_c \\ Y_c \\ Z_c \end{bmatrix} = \mathbf{K}\ \mathbf{X}_c$$

Here K is good enough to consider the radial distortion, non-square pixels and the skew.

6) Pose estimation

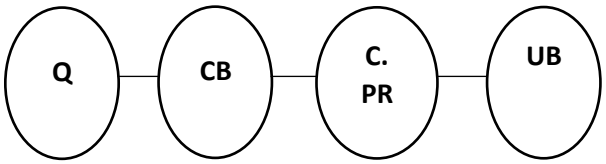
Once the internal camera parameters are known, can compute camera pose as

$U_{ij} = f\ (K, R_j, t_j, x_i)$

$V_{ij} = f\ (K, R_j, t_j, x_i)$

5. Mathematical model

A)



Here, Q = input camera data

CB = predict head pose

C = according to head position find X, Y, Z

PR = predicted result

UB = visualized result

B) Set theory

Let 'S' be as a system which finds the exact matched predefined value **5**and alert

S = In, P, Op

Identify,

input as In = Q (input data)

process as P = CB, C, PR

output as Op = UB

After preprocessing the request, system decides the particular result. If it is identified then system suggests the working.

6. Conclusion

In this paper, we proposed a system for predicting the head movement position using the CNN approach. Firstly, we collect the database with the help of camera. This is an important part of the dataset collection for the consistent head movements of the subjects. The face detection used in this collection of dataset makes it different from other studies over this topic. The database hence collected will be passed through the dense layers of the CNN. The eye and nose detection are performed once the face is detected to make the capturing accurate and exact. The templates captured are then sequenced together and a 360- degree video is established.

7. Acknowledgement

We see this opportunity given by our project guide Prof. Sharmila Chopade and Head of the Department Prof. Mangesh Manake for their precious guidance. The facilities they provided us are valuable too as they helped us for the successful completion of the project. We are also thankful to all the members of the staff of Department of Computer Engineering of Dr. D.Y Patil Institute of Engineering and Technology (DYPIET, Ambi) who helped us in many different ways and gave their efforts to make this project successful.

8. References

1) *M. Jiang, X. Boix, G. Roig, J. Xu, L. V. Gool, and Q. Zhao,* "Learning to predict sequences of human visual fixations,", June 2016.

2*) M. Kummerer, T. S. A. Wallis, L. A. Gatys, and M. Bethge,* "Understanding low- and high-level contributions to fixation prediction,", Oct 2017.

3) *S. S. Kruthiventi, K. Ayush, and R. V. Babu,* "Deepfix: A fully convolutional neural network for predicting human eye fixations,", 2015.

4) *Y.-C. Lin, Y.-J. Chang, H.-N. Hu, H.-T. Cheng, C.-W. Huang, and M. Sun,* "Tell me where to look: Investigating ways for assisting focus in 360 videos,", 2017.

5*) M. Assens, X. G. i Nieto, K. McGuinness, and N. E. O'Connor,* "Saltinet: Scan-path prediction on 360-degree images using saliency volumes,", Oct 2017.

6*) V. R. Gaddam, M. Riegler, R. Eg, C. Griwodz, and P. Halvorsen,* "Tiling in interactive panoramic video: Approaches and evaluation,", 2016.

7) *Y. Liu, S. Zhang, M. Xu, and X. He,* "Predicting salient face in multiple face videos,", 2017.

8)*H.-N. Hu, Y.-C. Lin, M.-Y. Liu, H.-T. Cheng, Y.-J. Chang, and M. Sun,* "Deep 360 pilot: Learning a deep agent for piloting through 360 sportsvideos," in CVPR, 2017.