LADDERNET: KNOWLEDGE TRANSFER BASED VIEWPOINT PREDICTION IN 360° VIDEO

Pengyu Zhao[†] *Yuanxing Zhang*[†] *Kaigui Bian*[†] *Hu Tuo*[¶] *Lingyang Song*[†]

[†]Peking University, Beijing, China
[¶]iQIYI Co. Ltd., Beijing, China
[†]{pengyuzhao, longo, bkg, lingyang.song}@pku.edu.cn, [¶]tuohu@qiyi.com

ABSTRACT

In the past few years, virtual reality (VR) has become an enabling technique, not only for enriching our visual experience but also for providing new channels for businesses. Untethered mobile devices are the main players for watching 360-degree content, thereby the precision of predicting the future viewpoints is one key challenge to improve the quality of the playbacks. In this paper, we investigate the image features of the 360-degree videos and the contextual information of the viewpoint trajectories. Specifically, we design ladder convolution to adapt for the distorted image, and propose LadderNet to transfer the knowledge from the pre-trained model and retrieve the features from the distorted image. We then combine the image features and the contextual viewpoints as the inputs for long short-term memory (LSTM) to predict the future viewpoints. Our approach is compared with several state-of-the-art viewpoint prediction algorithms over two 360-degree video datasets. Results show that our approach can improve the Intersection over Union (IoU) by at least 5% and meeting the requirements of the playback of 360-degree video on mobile devices.

Index Terms— Untethered virtual reality, image distortion, viewpoint prediction

1. INTRODUCTION

Virtual reality (VR) combines cutting-edge technologies in multimedia, sensors, Internet technologies, and artificial intelligence. VR content provides 360-degree panoramic view-sound field and immersive experience, and no matter where people are, they can feel themselves in the landscape. The rise of virtual reality may completely change real estate, games, tourism, industrial manufacturing, and other industries. Many companies attempt to combine VR with the Internet, big data, and cloud computing, and thus VR truly becomes a solution to the core technology from a toy.

Researchers have proposed many streaming strategies based on estimating the location of viewpoints [1] for upgrading the service of playing VR videos on the commercial immersive playback devices. Most of the strategies solve the problem by constructing optimization problems over some quality of experience (QoE) metrics [2, 3]. Other common strategies on this issue resort to group-based learning [4] and irregular tile shapes [5]. Usually, the streaming affects the economic aspect of the VR video-on-demand services, as the quality of streaming determines the engagement of users and the expenditure for content delivery.

The viewpoint-relevant streaming strategies always assume a high precision on predicting the exact location of future viewpoints of users. However, the viewpoints are usually difficult to predict, as the prediction may encounter the following three challenges: 1) **Cold start**. Mining the pattern among the historical trajectories requires complicated and annotated clustering models with accumulated trajectories data, whereas these works cannot handle the real-world scenario where newly released 360-degree videos are flooded every day. 2) **Distorted image**. Unlike the monocular videos, the contents in the 360-degree videos under the equirectangular projection scheme are always distorted, which can be hardly recognized unless being recovered by the perspective projection. 3) **Low latency requirement**. The untethered mobile devices usually own weak computing capabilities and receive media data from the Internet with limited bandwidth, thereby the computation latency should be minimized to guarantee the QoE of the playback.

Two categories of works, divided by whether they attempt to understand the distorted image directly, are proposed to address these challenges. In some specific scenarios where the possible points of interests are all gathered around the equator of the sphere with only slight distortion, conventional object detection algorithms [6] can help understand the point of interest. These methods show good performance on predicting the viewpoint in the subsequent frame, but may not predict accurately for the viewpoints to the several subsequent frames. Besides, the saliency map and motion map [7] are investigated to avoid detection of the distorted objects, while they are sensitive to noise in the real-world videos [8]. Regarding feature retrieval over the distorted images, state-of-the-art algorithms operate on the perspective projection rather than the distorted projection. Successful attempts resort to knowledge transfer, including SPHCONV [9] which imitates VGG on the cropped field of view and SphereNet [10] which wraps spherical filters. Saliency map of panoramic frames [11] and attention mechanism [12] are also patched to learn the features of the 360-degree videos, while a trial on reinforcement

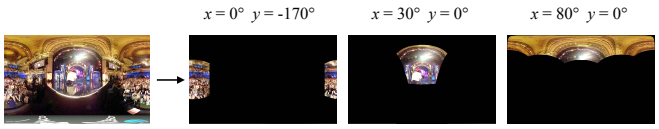

Fig. 1. Perspective fields of view to equirectangular projections given three viewpoints. The image will distort to different sizes and shapes depending on the viewpoints.

learning [13] under strong assumptions further improvement with complicated computation. These models rely on high computing capabilities, which may not be appropriate for the on-demand playback of 360-degree videos on mobile devices.

In this paper, we propose *ladder convolution* to solve the distortion of the equirectangular image. Based on the ladder convolution, we introduce *LadderNet* to transfer the knowledge of residual nets (ResNet) [14] over the perspective projection, and retrieve the decent image features from the distorted field of view. The image features are then concatenated with the embeddings of viewpoints and fed to long short-term memory (LSTM) for viewpoint prediction. Evaluation over two open-source datasets reveals the performance of our approach. Compared to several state-of-the-art viewpoint prediction strategies, our approach improves the Intersection over Union (IoU) by at least 5% without overwhelming computing resources or requiring long running time.

2. PROBLEM FORMULATION

Suppose we have a 360-degree video set V with size $|V|$ and each video $v \in V$ is attached with a trajectory set for tracking each user's viewpoint when viewing this video, denoted as τ_v. A trajectory consists of a sequence of polar angle (latitude) and azimuth angle (longitude) in a spherical coordinate system which indicates the viewpoint at a specific timestamp. Considering the processing capability of the mobile device, we extract and predict the viewpoints in 200 milliseconds granularity rather than every frame of the video. We retrieve the frames corresponding to the viewpoints to supply content information for the prediction. Besides, we calculate the region that would be viewed according to the viewpoint, defined as a *perspective field of view*. Based on the perspective field of view, we could recover the undistorted content to be displayed on the mobile device, defined as a *perspective projection*. Let $r = \langle (x_1, y_1, F_1), (x_2, y_2, F_2), \ldots, (x_{|r|}, y_{|r|}, F_{|r|}) \rangle$ indicate a trajectory, where x_*, y_*, and F_* stand for the latitude, longitude, and the equirectangular projection frame at a timestamp respectively. For simplicity, we fix the angle of field of view as $110^\circ \times 110^\circ$. Our task is to predict the viewpoints in the subsequent timestamps given the previous viewpoint trajectories and the retrieved frames over the entire video.

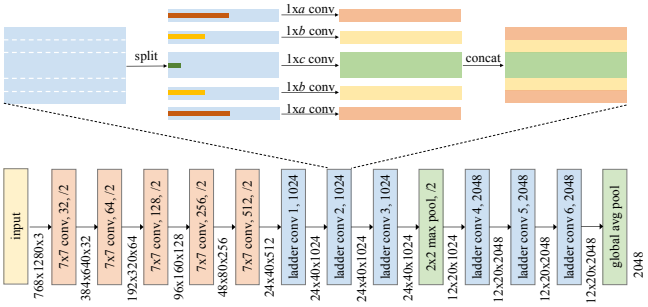

Fig. 2. The unfolded structure of LadderNet, which contains convolutional layers, Ladder-Conv layers, and pooling layers. The Ladder-Conv layer splits the sphere into strips and deploys various kernel on each strip, and thus it can adapt to the distorted image and reduce the number of parameters.

3. LADDER CONVOLUTION BASED KNOWLEDGE TRANSFER

In this section, we introduce the design of LadderNet, which is proposed to retrieve image features from the perspective field of view by transferring the knowledge on the perspective projection from ResNet.

3.1. Ladder Convolution

In a 360-degree video, the distortion is location-dependent - a rectangular perspective projection does not correspond to a rectangular region in the equirectangular projection, and its shape and size depend on the latitude of the viewpoint. Fig. 1 illustrates three perspective fields of view given the viewpoints as examples. It is obvious that perspective fields of view are different and not orthogonal to the image, while the contents inside the perspective fields of view are distorted. However, the legacy convolution is not appropriate for this issue, since the receptive fields of the neurons are the same in the convolution layer. Therefore, we propose ladder convolution to learn the knowledge of the region around the viewpoints from the distorted video frame by differentiating the receptive fields of the neurons in different locations.

Ladder-Conv Layer. As the shape and size of the perspective field of view are determined by the latitude of the viewpoint (row of the image), we are expected to use the separate kernels for each row of the equirectangular image. However, as the shapes of the perspective fields of view between adjacent rows tend to be similar, it is reasonable to make some of them share the kernels. Then, the number of the parameters of the network can be significantly reduced. Specifically, we propose *Ladder-Conv* layer which splits the sphere into strips by latitude, with separate 1-d kernels attached to each strip. As the equirectangular projection oversamples the spherical image in the polar regions, the length of the kernel for the polar strip should be longer than the kernel for the equator strip. To further reduce the complexity of the layer without loss of generality, we make the strips symmetric distributed over the

Table 1. Design of the Ladder-Conv layers by layer ID.

1		2		3		4		5		6	
Rows	Conv.	Rows	Conv.	Rows	Conv.	Rows	Conv.	Rows	Conv.	Rows	Conv.
1-5	1×5	1-4	1×5	1-4	1×5	1-2	1×5	1-2	1×5	1-2	1×3
6-20	1×3	5-8	1×3	5-7	1×3	3-3	1×3	3-3	1×3	3-11	1×1
21-24	1×5	9-17	1×1	8-18	1×1	4-10	1×1	4-10	1×1	12-12	1×3
-	-	18-21	1×3	19-21	1×3	11-11	1×3	11-11	1×3	-	-
-	-	22-24	1×5	22-24	1×5	12-12	1×5	12-12	1×5	-	-

sphere, and thereby we could use same 1-d kernels for the symmetric strips.

LadderNet. The structure of LadderNet is illustrated in Fig. 2. The size of the input image to the LadderNet is 768×1280 pixels, which is generated by downsampling of the ultra high-dimension (UHD) frame. We use five 7×7 kernels with stride 2 to increase the receptive field size while maintaining the information through the channel expansion. Six Ladder-Conv layers are utilized to solve the different perspective field of view in various latitude. They also reduce the number of parameters to meet the computing requirement of the playback of 360-degree video. Since the perspective field of view might be split into two if it passes through the 180° longitude, we use the global average pooling to address this phenomenon and compress the feature maps into a single vector to represent the image feature of the frame.

3.2. Knowledge Transfer Method

Uniform Sphere Sampling. To improve the training efficiency of the Ladder-Conv layers, we investigate an efficient uniform sampling of the spherical image, rather than conducting convolutions for the viewpoints at each pixel of the sphere. The equirectangular projection oversamples the spherical image in the polar regions, and thus the polar regions require sparser sampling than the equator regions. Specifically, we sample viewpoints along each row of the equirectangular projection that encircles the sphere in a way that the distance between adjacent points in each circle is inversely proportional to the latitude. Each pixel on the sphere is guaranteed to be covered by at least one sampled perspective fields of view, and thus the loss of information could be minimized. We calculate the perspective field of view on the equirectangular projection frame, i.e., retaining the color of each pixel in the perspective field of view and leaving the other pixels as black. The LadderNet receives the perspective field of view and finally outputs the extracted features corresponding to each viewpoint.

Knowledge Transfer from ResNet-50. We expect the LadderNet to automatically extract the features of the perspective projection from the distorted perspective field of view. We transfer the image features of the perspective projection with length 2048 learned by ResNet-50 to LadderNet, which captures the color, undistorted object, and distribution information. The LadderNet is thereby designed to generate the same output from the distorted image as the ResNet-50. During training, we minimize the L2 loss between the image features corresponding to the sampled viewpoints for the parameters of the LadderNet.

Hyper parameters. The LadderNet can be trained offline. We use the Adam optimization [15] with learning rate as 0.003. The specific design of Ladder-Conv layer is demonstrated in Tab. 1.

4. PREDICTING VIEWPOINTS BY LSTM

We implement LSTM to learn the long short-term dependency of the contextual information in the viewpoint trajectories. Given a trajectory r, the LSTM network is expected to make the prediction at timestamp t to predict the viewpoints in the future, based on the trajectory by the end of timestamp t.

4.1. Importing Image Features

The image features indicate what information the users are interested in. As mentioned in Sec. 3, the pre-trained LadderNet retrieves the image features from the perspective field of view. However, we expect the LadderNet could not only extract the "local" features of the perspective fields of view, but also learn the "global" image features of the entire equirectangular frame. It enlightens us to retrieve the *global feature* f_t by feeding the entire frame to the LadderNet, and uses this global feature for decoding the future viewpoints, i.e., $f_t = \text{LadderNet}(F_t)$.

4.2. LSTM Structure in LadderNet

To expand the representation capability, we transform the viewpoints from two real values to a high dimensional embedding through a fully-connected layer. The parameters of the embedding learning layer can be trained along with the LSTM cells. We expect the LSTM can figure out the points of interests of the users based on the previous part of the viewpoint trajectory to the video, and embed the knowledge in the hidden state of the LSTM cell. Then, given the image feature of the next frame, the LSTM is able to retrieve the evidence indicating for the location that the user may be interested in. We concatenate the image feature with the embedding of viewpoint location and feed them into the LSTM cell to generate the location-relevant vector,

$$e_t, h_t = \text{LSTM}(f_{t+1}, \text{EMBEDDING}(x_t, y_t); h_{t-1}), \quad (1)$$

where h_t is the hidden state of the t-th LSTM cell. Finally, we set a fully-connected layer to decode the prediction of the viewpoints at timestamp $t+1$, i.e.,

$$(x_{t+1}^{(t)}, y_{t+1}^{(t)}) = \text{DECODER}(e_t). \quad (2)$$

4.3. Training

We could recurrently feed the predicted viewpoint back to the LSTM and obtain the prediction $(x_{t'}^{(t)}, y_{t'}^{(t)})$ for any timestamp t' given the historical trajectory by the end of timestamp t. As we are expected to predict accurately for several subsequent timestamps, the number of which is denoted as γ, in

the future, the model should minimize the loss

$$L(V;\theta) = \sum_{v\in V}\sum_{r\in\tau_v}\sum_{t=1}^{|r|-\gamma}\sum_{t'=t+1}^{t+\gamma} ||(x_{t'}^{(t)}, y_{t'}^{(t)}) - (x_{t'}, y_{t'})||^2 \tag{3}$$

where θ represents the parameter of the models to be learned from the training data.

5. EVALUATION

5.1. Settings

Datasets. We collect two datasets of UHD 360-degree videos with viewpoint trajectories. The first dataset [16] includes 16 videos with at least 3 minutes in length, which provide enough contextual information. The second dataset [17] consists of 18 videos with only 20 seconds length. Most objects in the videos are severely distorted because of the equirectangular projection that the object detection algorithms cannot accurately identify them under pilot experiments like faster R-CNN [18] and YOLO [19]. The evaluations are cross-validated by taking one video as the unseen test video and the other videos as the training videos. 90% of the trajectories corresponding to the training videos are used for training, and the remaining trajectories are treated as part of the test set.

Algorithms for Comparison. We compare the proposed LadderNet on the viewpoint prediction task with several common viewpoint prediction strategies in the streaming system, including: *Reactive* uses the latest viewpoint; *Linear regression* (LR) predicts by the previous 4 viewpoints; *LSTM* predicts by the viewpoints from the beginning of the trajectory. We also compare LadderNet with several state-of-the-art viewpoint prediction algorithms which may be affordable on the mobile devices regarding the runtime memory consumption and execution time: *Deep 360 Pilot* (D360P) [6] selects the main objects in the entire distorted image by faster R-CNN [18] and predicts viewpoint; *SphereNet* [10] learns the image features by wrapping filters on the sphere, which are then fed to the LSTM to make the prediction; *ResNet* [14] learns the image features from the entire distorted image, which are then fed to the LSTM to make the prediction.

Performance Metrics. To quantify our results on predicting viewpoints for the playback of 360-degree videos on mobile devices, we report both Intersection over Union (IoU) and Running Time (RT). **IoU** measures how much the perspective field of view centered at the predicted viewpoint overlaps with that of the ground truth at each frame. **RT** evaluates the average computing time for predicting viewpoint at any timestamp.

5.2. Performance of LadderNet

IoU of Predicting Viewpoints. The IoU metric indicates to what extent the prediction is near the ground truth, which is calculated by the intersection over the union of the perspective field of view under the angle of $110^\circ \times 110^\circ$. We examine the IoU for the prediction of the viewpoints in the subsequent 4 seconds, namely 20 timestamps. Performance of the seven compared algorithms is plotted in Fig. 3. We find that the algorithms working on the perspective projection (i.e., SphereNet, LadderNet) improve the IoU compared to the algorithms working directly on the distorted image (i.e., D360P, ResNet). With the help of the image features and contextual information, LadderNet outperforms the state-of-the-art algorithms by at least 5% on average.

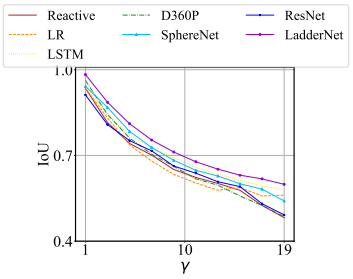

Fig. 3. The performance on IoU of 7 algorithms on both datasets with various γ.

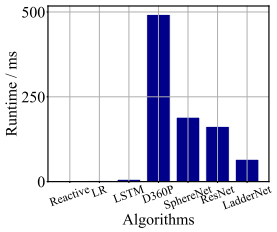

Fig. 4. The performance on RT of 7 algorithms on both datasets.

Runtime Latency for Predicting Viewpoints. We depict the average runtime for generating viewpoint for one timestamp from the compared algorithms in Fig. 4. The content-independent algorithms provide predictions in up to tens of milliseconds, while the content-aware algorithms require hundreds of milliseconds. Due to the reduction of the number of parameters, LadderNet significantly decreases the running time. Compared to the other content-aware algorithms, LadderNet leaves much space for executing heuristic QoE-driven streaming strategies on mobile devices.

6. CONCLUSION

In this paper, we propose LadderNet to retrieve image features from the distorted equirectangular projection by transferring the knowledge from ResNet-50 on the perspective projection. LadderNet includes ladder convolution layers, which split the sphere into symmetric strips and conducts convolution on each strip to adapt for the distorted content and reduce the number of parameters. We combine the image features and the contextual information of the viewpoint trajectories to predict the viewpoints at a future timestamp. Evaluations reveal that LadderNet outperforms several state-of-the-art viewpoint prediction algorithms and it is appropriate for being deployed on mobile devices.

Acknowledgment

This work is partially supported by the National Key Research and Development Program No. 2017YFB0803302, the National Natural Science Foundation of China No. 61572051 and CERNET Innovation Project NGII20160124.

7. REFERENCES

[1] Jill M Boyce, Yan Ye, Jianle Chen, and Adarsh K Ramasubramonian, "Overview of shvc: Scalable extensions of the high efficiency video coding standard," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 26, no. 1, pp. 20–34, 2016.

[2] Feng Qian, Bo Han, Qingyang Xiao, and Vijay Gopalakrishnan, "Flare: Practical viewport-adaptive 360-degree video streaming for mobile devices," in *ACM MobiCom*, 2018.

[3] Jian He, Mubashir Adnan Qureshi, Lili Qiu, Jin Li, Feng Li, and Lei Han, "Rubiks: Practical 360-degree streaming for smartphones," in *Mobisys*. ACM, 2018.

[4] Lan Xie, Xinggong Zhang, and Zongming Guo, "Cls: A cross-user learning based system for improving qoe in 360-degree video adaptive streaming," in *2018 ACM Multimedia Conference*. ACM, 2018, pp. 564–572.

[5] Chao Zhou, Mengbai Xiao, and Yao Liu, "Clustile: Toward minimizing bandwidth in 360-degree video streaming," in *IEEE INFOCOM 2018*. IEEE, 2018, pp. 962–970.

[6] Hou-Ning Hu, Yen-Chen Lin, Ming-Yu Liu, Hsien-Tzu Cheng, Yung-Ju Chang, and Min Sun, "Deep 360 pilot: Learning a deep agent for piloting through 360 sports videos," in *Proc. CVPR*, 2017, pp. 1396–1405.

[7] Ching-Ling Fan, Jean Lee, Wen-Chih Lo, Chun-Ying Huang, Kuan-Ta Chen, and Cheng-Hsin Hsu, "Fixation prediction for 360 video streaming in head-mounted virtual reality," in *Proceedings of the 27th Workshop on Network and Operating Systems Support for Digital Audio and Video*. ACM, 2017, pp. 67–72.

[8] Yujie Li, Atsunori Kanemura, Hideki Asoh, Taiki Miyanishi, and Motoaki Kawanabe, "A sparse coding framework for gaze prediction in egocentric video," in *2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2018, pp. 1313–1317.

[9] Yu-Chuan Su and Kristen Grauman, "Learning spherical convolution for fast features from 360 imagery," in *Advances in Neural Information Processing Systems*, 2017, pp. 529–539.

[10] Benjamin Coors, Alexandru Paul Condurache, and Andreas Geiger, "Spherenet: Learning spherical representations for detection and classification in omnidirectional images," in *Proceedings of the European Conference on Computer Vision (ECCV)*, 2018, pp. 518–533.

[11] Ziheng Zhang, Yanyu Xu, Jingyi Yu, and Shenghua Gao, "Saliency detection in 360 videos," in *Proceedings of the European Conference on Computer Vision (ECCV)*, 2018, pp. 488–503.

[12] Xin Ji, Wei Wang, Meihui Zhang, and Yang Yang, "Cross-domain image retrieval with attention modeling," in *Proceedings of the 2017 ACM on Multimedia Conference*. ACM, 2017, pp. 1654–1662.

[13] Mai Xu, Yuhang Song, Jianyi Wang, MingLang Qiao, Liangyu Huo, and Zulin Wang, "Predicting head movement in panoramic video: A deep reinforcement learning approach," *IEEE transactions on pattern analysis and machine intelligence*, 2018.

[14] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun, "Deep residual learning for image recognition," in *Proc. CVPR*, 2016, pp. 770–778.

[15] Diederik P Kingma and Jimmy Ba, "Adam: A method for stochastic optimization," *arXiv preprint arXiv:1412.6980*, 2014.

[16] Chenglei Wu, Zhihao Tan, Zhi Wang, and Shiqiang Yang, "A dataset for exploring user behaviors in vr spherical video streaming," in *Proceedings of the 8th ACM on Multimedia Systems Conference*. ACM, 2017, pp. 193–198.

[17] Erwan J David, Jesús Gutiérrez, Antoine Coutrot, Matthieu Perreira Da Silva, and Patrick Le Callet, "A dataset of head and eye movements for 360 videos," in *Proceedings of the 9th ACM Multimedia Systems Conference*. ACM, 2018, pp. 432–437.

[18] Shaoqing Ren, Kaiming He, Ross Girshick, and Jian Sun, "Faster r-cnn: Towards real-time object detection with region proposal networks," in *Advances in neural information processing systems*, 2015, pp. 91–99.

[19] Joseph Redmon, Santosh Divvala, Ross Girshick, and Ali Farhadi, "You only look once: Unified, real-time object detection," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016, pp. 779–788.