EGOK360: A 360 EGOCENTRIC KINETIC HUMAN ACTIVITY VIDEO DATASET

Keshav Bhandari[1], *Mario A. DeLaGarza*[1], *Ziliang Zong*[1], *Hugo Latapie*[2], *Yan Yan*[1]

[1]Department of Computer Science, Texas State University, USA
[2]Chief Technology & Architecture Office, Cisco, USA

ABSTRACT

Recently, there has been a growing interest in wearable sensors which provides new research perspectives for 360° video analysis. However, the lack of 360° datasets in literature hinders the research in this field. To bridge this gap, in this paper we propose a novel Egocentric (first-person) 360° Kinetic human activity video dataset (EgoK360). The EgoK360 dataset contains annotations of human activity with different sub-actions, *e.g.*, activity Ping-Pong with four sub-actions which are pickup-ball, hit, bounce-ball and serve. To the best of our knowledge, EgoK360 is the first dataset in the domain of first-person activity recognition with a 360° environmental setup, which will facilitate the egocentric 360° video understanding. We provide experimental results and comprehensive analysis of variants of the two-stream network for 360 egocentric activity recognition. The EgoK360 dataset can be downloaded from `https://egok360.github.io/`.

Index Terms— 360° videos, Kinetic, Egocentric, Activity-recognition, Two-stream Network

1. INTRODUCTION

Wearable devices like Apple smartwatch, GoPro and Google Clip, have been widely used in our daily life nowadays. Meanwhile, the appearance of 360° cameras and the growing services on social media platforms such as Facebook and YouTube are changing the way how we consume multimedia. Having the advantage of 360 field-of-view over perspective videos from traditional cameras, 360° cameras have the superiority in many applications such as self-driving cars, virtual-reality, life-logging, augmented reality, film-making and surveillance [1, 2]. The popularity of 360° videos is also changing computer vision and virtual reality research area recently. Egocentric Activity Recognition (EAR) from videos is one of such fields. However, to the best of our knowledge, there is no public 360 egocentric human activity dataset in literature.

In this paper, we propose a novel 360 egocentric human activity recognition (EgoK360) dataset. The EgoK360 dataset is inspired by action recognition datasets such as UCF-101 [3], HMDB-51 [4] and Kinetics [5]. Our EgoK360 dataset contains three different types of actions: Person-Person, Person-Object, and Singular actions. These categories of actions can be described in the following manner. Person-Person actions involve two or more people interacting with each other such as hugging and speaking with someone; Person-Object actions refer to a person interacting with some objects such as picking up something or moving something from one location to another; Singular person actions involve a single person performing some actions independent of others such as reaching towards something or combing hair. In this paper we perform experiments with two popular action classification deep nerual networks on our introduced EgoK360 dataset, *i.e.*, two-stream network [6] and Inflated 3-Dimensional network (I3D) [5]. In the following sections, we present the related work, datasets, experimental results and conclusions.

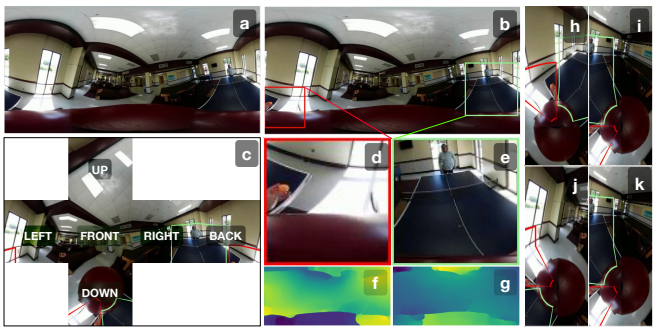

Fig. 1: Sample video frames from EgoK360 Dataset. (a,b) Consecutive frames (I_i, I_{i+1}) for action "Serve" from "Ping-Pong" activity in equi-rectangular projection. (c) Cubemap projection of (b) showing six different cubic faces. (d,e) Cropped section of wearer showing action 'serving' (red box in (b)) and front-view from wearer's perspective (green box in (b)). (f,g) Optical flow ($\vec{u}_i, \vec{v}_i$). (h,i,j,k) Normal field-of-view for front-down, back-down, left-down and right-down.

2. RELATED WORK

Action recognition datasets such as HMDB [4], UCF101 [3] and Kinetcs [7] are widely used in literature. They are captured by perspective cameras (single field-of-view) and have limitations in terms of applications. Singh *et al.* [8] use

a novel dominant motion feature derived from optical flow for egocentric action recognition and also propose a convolutional neural networks (CNN) [9] for end-to-end training. Xia *et al*. [10] present a framework to analyze RGBD videos captured from a robot for activity recognition. Lee *et al*. [11] present an egocentric video summarization approach by identifying important people and object in the video. Two-stream network is the popular architecture in literature for action recognition, such as Two-stream Convnet [6] and Inflated 3D ConvNet (I3D) [5]. I3D architecture is the state-of-the-art in the two-stream genre for action recognition.

In recent years, a few 360° datasets [12, 13, 14, 15, 16, 2] appeared in the applications such as autonomous driving, human-computer interaction, virtual reality, and others. However, they are target to different applications other than Egocentric Activity Recognition in 360° field-of-view (EAR360).

Meanwhile, in the egocentric action recognition field, popular datasets such as EgoHands [17], EGTEA Gaze+ [18] are perspective video datasets with a person interacting with an object or another person. Similarly, large-scale datasets such as Charades-Ego [19] contains both the first-person and third-person videos. Pirsiavash *et al*. [20] present an egocentric dataset for understanding activities and the context in the video. However, all these datasets in literature are only limited to perspective videos.

3. EGOK-360 DATASET

Our EgoK-360 dataset contains activity classes that represent all three categories, *i.e.,* Person-Person, Person-Object, and Singular actions. There are a few differences in the video content compared with other datasets because of the properties of Egocentric 360 videos. Given that the footage encompassing the dataset captured from an egocentric perspective, the Person-Person actions would involve the interaction between the wearer and other people. This differs with Person-Person actions captured in the traditional third-person perspective cameras. Likewise, a Person-Object action such as bouncing a ping-pong ball on a table would only be identified when the particular action was performed by the wearer. Most action classes in EgoK360 are in the Person (singular) category because the egocentric perspective inherently privatizes the action or content recorded.

The 360 field-of-view naturally makes everything egocentric in EgoK360. Egocentric actions entirely depend on the field-of-view where a wearer (first-person) is engaged. Meanwhile, the rest of fields-of-view that he/she is not engaged are irrelevant for action recognition. The significant contribution to action recognition is the wearer's egocentric view and his/her engagement in actions.

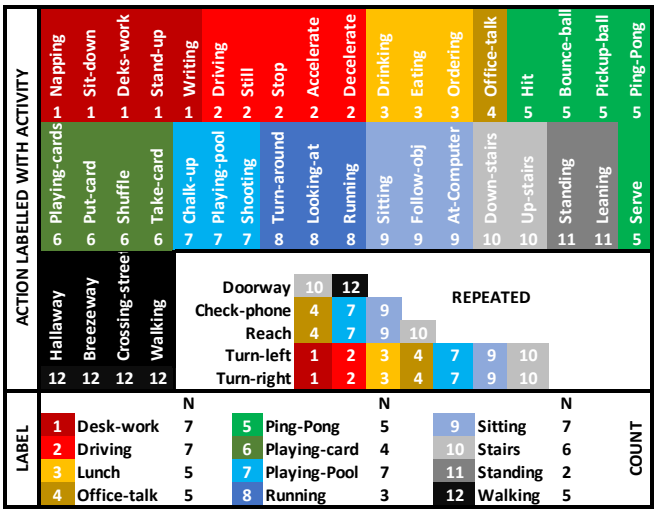

Fig. 2: All activities and action classes in EgoK-360. Actions are colored and numbered with corresponding activity. The same actions may appear in different activities.

3.1. Activity/Action Classes

We show action/activity instances of EgoK360 in Fig. 2. Our dataset contains 12 activities and 45 actions, collectively making 63 activity-action unique cases. An activity is defined as collections of shorter actions. For example, an activity 'driving' is composed of actions such as accelerate, decelerate, idle, stop, driving, turn-left and turn-right. Action classes such as turn-left, turn-right, reach, doorway and check-phone are frequently occurring actions. However, there is a significant difference in these actions depending upon the category of activity. For example, turn-left in driving is completely different than turn-left in activity office-talk. We collected 127 videos with approximately 11 minutes each.

3.2. EgoK-360 Characteristics

The EgoK-360 dataset has its uniqueness of 360 fields-of-view, egocentric and kinetic properties. We discuss the following characteristics of EgoK-360 dataset in terms of its diversity, statistics and properties.

Diversity. Our EgoK-360 dataset contains common different activities in daily life. Around 11% of actions (such as turn-left, turn-right, reach, check-phone and doorway are frequent actions) are overlapping actions. Activity such as desk-work, driving, playing-pool and running have the most number of actions. Activity such as standing has the least number of actions.

Properties. We present sample frame in Fig. 1. The dataset is a collection of videos from a 360 camera projected on the 2D plane using equirectangular projection, as shown in Fig. 1 (a and b). The frames size is 640x320. Frames exhibit huge distortion as shown in Fig. 1 (b-c, h-k) using red and green

bounding box), making it challenging for regular convolution. We calculate optical flows using FlowNet [21].

4. EXPERIMENT

We conduct our experiments using two-stream and I3D networks. We implement two-stream architecture with resnet-101 model pre-trained on UCF101 which outperforms state-of-the-art I3D model. EgoK360 exhibits complexity of spherical representation of 360° video on 2D plane (equirectangular projection) which makes challenge for these models to prioritize a significant field-of-view responsible for the wearer's engagement in certain actions and makes difficult to train. Therefore, 3D-representation of the video does not perform well in the 360 environment.

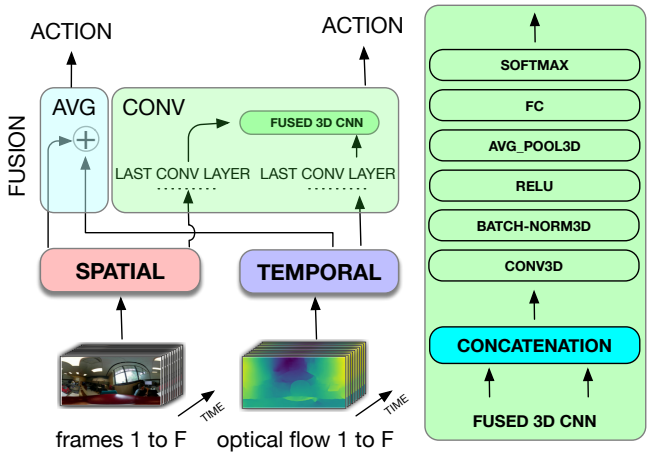

Fig. 3: Spatio-temporal architecture for action/activity classification. We implement resnet-101 and I3D architecture. Average and convolution fusion are adopted. For average fusion we simply average probabilities of two networks and map them into single probability. For convolution fusion, we concatenate (depth-wise) output from last convolution layer and feed to the convolution module.

4.1. Implementation Details

We adopt the network architecture as shown in Fig. 3 in our experiments. Our model inputs are consecutive frames. Videos are down-sampled in the rate of 10 fps. We calculate optical flows beforehand using FlowNet [21]. We adopt the two-stream and I3D architectures with average and convolution fusion. For two-stream architecture, video is represented as 2D inputs with $[N \times F_c \times H \times W]$ dimensions. For I3D architecture, video is represented as 3D input with $[N \times C \times F \times H \times W]$ dimensions, where $F_c = F \times C$. Here F is the number of frames, N is the batch-size, C, H and W are channel, height and width of the frames.

4.2. Two-stream Architecture

Residual learning framework [22] provides convenient optimization and rapid high accuracy as network becomes deeper. With this in mind, we change the two-stream architecture by replacing the spatial and temporal network with resnet-101 model pre-trained on UCF101. We use size of 10 to stack frames in sequence for both spatial and temporal networks. This brings the channel size changing from 3 to 20 in the temporal network and to 30 in the spatial network. The resnet-101 requires input as 3 channel images. To fix this, we use the method in the cross-modal learning [23]. We do not observe better results compared to UCF101 implemented with the same architecture, which achieves at least 80% accuracy.

4.3. I3D Architecture

We implement I3D architecture as proposed in [5]. I3D architecture relies on 3D receptive fields for video representation. Spatial and temporal network receive an input of 3 and 2 for the channel size respectively, along with depth of 10. The original idea in [5] using the entire video as one training sample. However, we do not achieve performance increase on EgoK-360 dataset. We run experiments with the depth of 10 which is the optimal for our case.

4.4. Fusion

In this paper we use both average and convolution fusion techniques. In average fusion, we take an average of probabilities from the last layers. For convolution fusion, we implement convolution-module inspired by [22]. We use the output from the last convolution layer and concatenate the features which later fed into the fusion convolution module. We freeze our spatio-temporal module and train the fusion layer. We can also train the spatio-temporal network along with the fusion layer.

4.5. Results

We present our experimental results in Table 1 and visualization in Fig. 4. We observe that activity classification accuracy is higher than action classification accuracy as shown in Table-1. The range of activity and action classification accuracy in I3D architecture is higher than in two-stream architecture.

The average fusion is remarkably better than the convolution fusion in our case. This quantitative results can be explained using Fig-4 visualization. We observe interesting results on two different architectures. Fusion techniques make a huge difference in action/activity classification in resnet whereas spatial and temporal streams have significant differences in I3D architecture. From Fig. 4 we can infer that convolution fusion performs better whenever two streams have significant gap in accuracy.

Architecture	Network Mode							
	flow		rgb		avg_fused		conv_fused	
	Activity	Action	Activity	Action	Activity	Action	Activity	Action
Resnet	73.94	61.09	**77.05**	58.22	76.53	62.44	74.71	56.87
I3D	57.24	43.4	74.13	55.31	68.74	50.88	**74.47**	56.63

Table 1: Experimental results of EgoK360 datasets on two-Stream (modified version with trained Resnet-101 on UCF101) and I3D Architecture. (*Top accuracy in bold*)

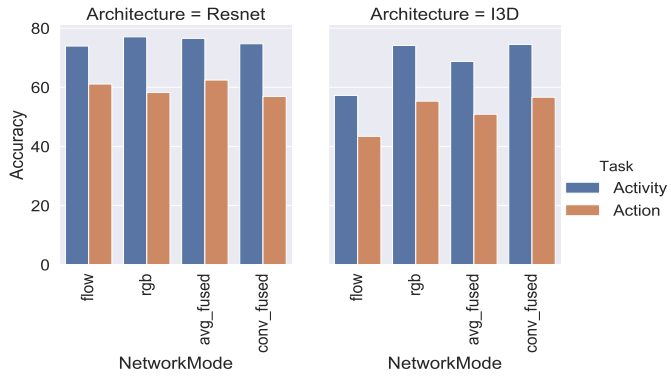

Fig. 4: Visualization results of Table 1. The figure shows accuracy for each architecture and classification mode (activity *vs* action). Overall 'action' classification is better in resnet-101. We observe convolution fusion in resnet architecture and I3D makes significance difference in flow (temporal) and rgb (spatial) stream. Similarly, temporal-stream performs better in resnet compared with I3D architecture. Conv_fusion has same effect on both cases where as avg_fusion comparatively improve resnet architecture. In general resnet architecture shows consistent metrics relative to I3D arhcitecture.

We also investigate how well this two-stream architecture generalizes with our dataset. We use the technique presented in [24] to visualize the activation map. We show the activation map with a randomly selected action in Fig. 5. We derive these activation maps from the I3D and two-stream spatial network.

These activation maps represent salient features learned by the model. The model infers most edges as trivial regions, as the dataset has massive distortion near edges. We can visually inspect and analyze this behavior. For example, in the Fig. 5 activity Bounce_ball (playing Ping-Pong activity), the salient features are away from the actual region where a person wearing a camera is bouncing a ball. This region lies on the left-bottom-corner and has massive distortion. It is nearly impossible to judge the meaning of these activation maps accurately. However, if we carefully inspect the salient features learned by both architectures, we can conclude that the model is inferring the action classification task from other features rather than salient features as expected. The reason behind this poor response of the model is due to the naïve convolution, which is not rotation invariant. Features on EgoK360 have different spatial properties depending upon the position in equirectangular plane. This can be improved with techniques such as [25, 26].

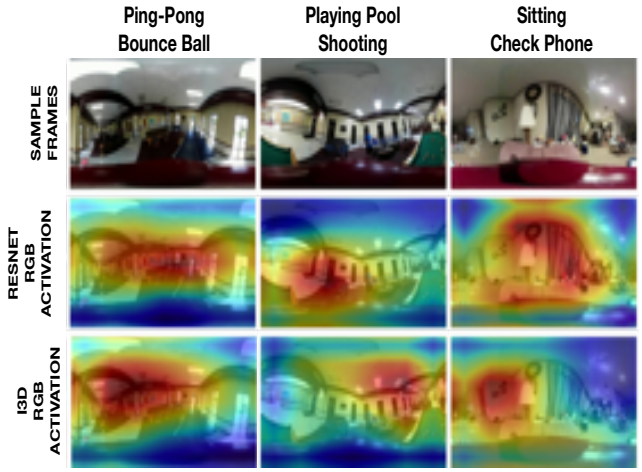

Fig. 5: Activation map showing salient features learned by the spatial network. The top row shows RGB frames, the second row represents activation map from two-stream networks, and the bottom row shows the activation map from I3D model.

5. CONCLUSION

This paper introduces EgoK360 dataset with annotations of 63 unique activity and action classes. This dataset is challenging because of distortion, wide field-of-view and activities/actions properties. We implement two popular two-stream architectures in the experiments. We modify the two-stream convents architecture by replacing each stream with resnet-101. It outperforms state-of-the-art I3D architecture. EgoK-360 is the first to address egocentric activity recognition in 360 environment. We believe EgoK360 dataset will be beneficial to the EAR360 research.

Acknowledgements: This research was partially supported by NSF CSR-1908658, NeTS-1909185, and gift donation from Cisco Inc. This article solely reflects the opinions and conclusions of its authors and not the funding agents.

6. REFERENCES

[1] S. Hecker, D. Dai, and L. Van Gool, "End-to-end learning of driving models with surround-view cameras and route planners," in *ECCV*, September 2018.

[2] C. Häne, L. Heng, G. Lee, F. Fraundorfer, P. Furgale, T. Sattler, and M. Pollefeys, "3d visual perception for self-driving cars using a multi-camera system: Calibration, mapping, localization, and obstacle detection," *IVC*, vol. 68, pp. 14–27, 2017.

[3] K. Soomro, A. Roshan Zamir, and M. Shah, "UCF101: A Dataset of 101 Human Actions Classes From Videos in The Wild," *arXiv e-prints*, p. arXiv:1212.0402, Dec 2012.

[4] H. Kuehne, H. Jhuang, E. Garrote, T. Poggio, and T. Serre, "Hmdb: a large video database for human motion recognition," in *ICCV*. IEEE, 2011, pp. 2556–2563.

[5] J. Carreira and A. Zisserman, "Quo Vadis, Action Recognition? A New Model and the Kinetics Dataset," *arXiv e-prints*, p. arXiv:1705.07750, May 2017.

[6] K. Simonyan and A. Zisserman, "Two-Stream Convolutional Networks for Action Recognition in Videos," *arXiv e-prints*, p. arXiv:1406.2199, Jun 2014.

[7] W. Kay, J. Carreira, K. Simonyan, B. Zhang, C. Hillier, S. Vijayanarasimhan, F. Viola, T. Green, T. Back, P. Natsev, M. Suleyman, and A. Zisserman, "The Kinetics Human Action Video Dataset," *arXiv e-prints*, p. arXiv:1705.06950, May 2017.

[8] S. Singh, C. Arora, and C. V. Jawahar, "Generic action recognition from egocentric videos," in *NCVPRIPG*, Dec 2015, pp. 1–4.

[9] S. Singh, C. Arora, and CV Jawahar, "First person action recognition using deep learned descriptors," in *CVPR*, 2016, pp. 2620–2628.

[10] Lu s, I. Gori, J. K Aggarwal, and M. S Ryoo, "Robot-centric activity recognition from first-person rgb-d videos," in *WACV*. IEEE, 2015, pp. 357–364.

[11] Y. J. Lee, J. Ghosh, and K. Grauman, "Discovering important people and objects for egocentric video summarization," in *CVPR*, June 2012, pp. 1346–1353.

[12] Y. Rai, J. Gutiérrez, and P. Le Callet, "A dataset of head and eye movements for 360 degree images," in *Proceedings of the 8th ACM on Multimedia Systems Conference*. ACM, 2017, pp. 205–210.

[13] W. Lo, C. Fan, J. Lee, C. Huang, K. Chen, and C. Hsu, "360 video viewing dataset in head-mounted virtual reality," in *Proceedings of the 8th ACM on Multimedia Systems Conference*. ACM, 2017, pp. 211–216.

[14] H. Hu, Y. Lin, M. Liu, H. Cheng, Y. Chang, and M Sun, "Deep 360 pilot: Learning a deep agent for piloting through 360 sports videos," in *CVPR*. IEEE, 2017, pp. 1396–1405.

[15] C. Wu, Z. Tan, Z. Wang, and S. 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.

[16] H. Cheng, C. Chao, J. Dong, H. Wen, T. Liu, and M. Sun, "Cube Padding for Weakly-Supervised Saliency Prediction in 360° Videos," *arXiv e-prints*, p. arXiv:1806.01320, Jun 2018.

[17] S. Bambach, S. Lee, D. J. Crandall, and C. Yu, "Lending a hand: Detecting hands and recognizing activities in complex egocentric interactions," in *ICCV*, December 2015.

[18] Y. Li, M. Liu, and J. M Rehg, "In the eye of beholder: Joint learning of gaze and actions in first person video," in *ECCV*, 2018, pp. 619–635.

[19] G. A Sigurdsson, A. Gupta, C. Schmid, A. Farhadi, and K. Alahari, "Charades-ego: A large-scale dataset of paired third and first person videos," *arXiv preprint arXiv:1804.09626*, 2018.

[20] H. Pirsiavash and D. Ramanan, "Detecting activities of daily living in first-person camera views," in *CVPR*. IEEE, 2012, pp. 2847–2854.

[21] E. Ilg, N. Mayer, T. Saikia, M. Keuper, A. Dosovitskiy, and T. Brox, "Flownet 2.0: Evolution of optical flow estimation with deep networks," in *CVPR*, Jul 2017.

[22] K. He, X. Zhang, S. Ren, and J. Sun, "Deep Residual Learning for Image Recognition," *arXiv e-prints*, p. arXiv:1512.03385, Dec 2015.

[23] M. Luo, X. Chang, Z. Li, L. Nie, A. G. Hauptmann, and Q. Zheng, "Simple to Complex Cross-modal Learning to Rank," *arXiv e-prints*, p. arXiv:1702.01229, Feb 2017.

[24] B. Zhou, A. Khosla, A. Lapedriza, A. Oliva, and A. Torralba, "Learning Deep Features for Discriminative Localization," *arXiv e-prints*, p. arXiv:1512.04150, Dec 2015.

[25] Y. Su and K. Grauman, "Kernel Transformer Networks for Compact Spherical Convolution," *arXiv e-prints*, p. arXiv:1812.03115, 2018.

[26] Y. Su and K. Grauman, "Learning Spherical Convolution for Fast Features from 360 ° Imagery," *arXiv e-prints*, p. arXiv:1708.00919, Aug 2017.