International Journal on Recent and Innovation Trends in Computing and Communication
ISSN: 2321-8169 Volume: 11 Issue: 9s
DOI: https://doi.org/10.17762/ijritcc.v11i9s.7420
***Article Received**: 08 May 2023 **Revised:** 29 June 2023 **Accepted:** 28 July 2023*

Understanding First-Person and Third-Person Videos in Computer Vision

Sheetal Girase[1], Mangesh Bedekar[2]
[1]School of Computer Engineering
Dr. Vishwanath Karad MIT World Peace University,
Pune 411038, Maharashtra, India
sheetal.girase@mitwpu.edu.in
[2]School of Computer Engineering
Dr. Vishwanath Karad MIT World Peace University,
Pune 411038, Maharashtra, India
Mangesh.bedekar@mitwpu.edu.in

Abstract— Due to advancements in technology and social media, a large amount of visual information is created. There is a lot of interesting research going on in Computer Vision that takes into consideration either visual information generated by first-person (egocentric) or third-person(exocentric) cameras. Video data generated by YouTubers, Surveillance cameras, and Drones which is referred to as third-person or exocentric video data. Whereas first-person or egocentric is the one which is generated by GoPro cameras and Google Glass. Exocentric view capture wide and global views whereas egocentric view capture activities an actor is involved in w.r.t. objects. These two perspectives seem to be independent yet related. In Computer Vision, these two perspectives have been studied by various domains like Activity Recognition, Object Detection, Action Recognition, and Summarization independently. Their relationship and comparison are less discussed in the literature. This paper tries to bridge this gap by presenting a systematic study of first-person and third-person videos. Further, we implemented an algorithm to classify videos as first-person/third-person with the validation accuracy of 88.4% and an F1-score of 86.10% using the Charades dataset..

Keywords- First-person videos, Third-person videos, Video Classification, Charades Dataset.

I. INTRODUCTION

Today's world is not only about the internet but also about smartphones. Smartphones have replaced traditional cameras. Social networks like Instagram and Snapchat ha**ve** become popular since photos and videos can be shared faster. No wonder YouTube has its own share in it. The use of surveillance cameras and drones is also becoming much more popular. These are contributing to creating much more visual data and capturing a third-person perspective. In Third-Person Vision (TPV) either the camera is mounted on a wall or it is been held by somebody who is capturing that environment. For example, Surveillance cameras mounted on stations, subways, inside houses, etc. along with the shooting of a scene by the cameraman as shown in Fig. 1(a). In nutshell, In a third-person video, The understanding of the information related to the action being performed and the action performer appears on the forefront of the video[1].

Figure 1 (a) Third-person surveillance camera setup

Figure 1 (b) First-person body camera and Google glass

The introduction of wearable cameras in the 1990s by Steve Mann has revolutionized the IT industry and created a deep impact on our daily lives[2]. After the introduction of the first Wearable wireless webcam in 1994 by Mann at MIT, Boston many researchers started working on wearable cameras to record lifelogs. Since 2002, the popularity of

International Journal on Recent and Innovation Trends in Computing and Communication
ISSN: 2321-8169 Volume: 11 Issue: 9s
DOI: https://doi.org/10.17762/ijritcc.v11i9s.7420
***Article Received**: 08 May 2023 **Revised:** 29 June 2023 **Accepted:** 28 July 2023*

wearable devices such as GoPro cameras and Google glasses as shown in Fig. 1(b) has changed the nature of the task at hand. In First-person Vision (FPV) the person behind the camera is a point of interest in which images/videos are captured from the perspective of one's self[[1]. Today's wearable cameras are compact and robust capturing images and videos with good resolution with different frame rates. In computer vision, for the tasks like interaction identification, action recognition, person identification, pose estimation, etc., the human actor is the main focus. Thus recognizing, detecting and localizing human is a vital component. As shown in Fig. 2 the wearable cameras are having applications in surveillance domain, for capturing extreme sports activity, Rescue operation monitoring.

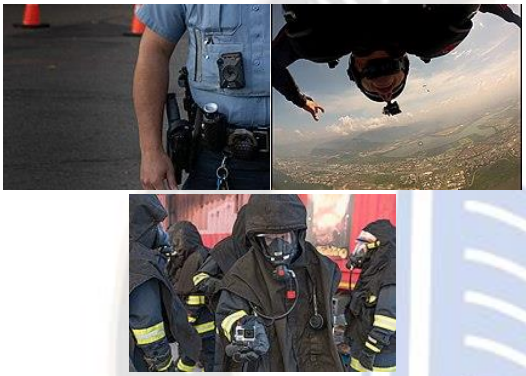

Figure 2. Applications of body cameras (i.e. first-person cameras) in patrolling, extreme sports like Skydiving, and Firefighters in rescue monitoring are increasing

As per [3], First Person Vision (FPV) is arguably the most commonly used term. In fact nowadays Egocentric has also gain popularity and been used interchangeably with the FPV. Similar is the case with Third Person Vision (TPV) and Exocentric. Also, while reading literature we came across terminology first-person and third-person which also refers to egocentric (FPV) and exocentric (TPV) respectively and will be used in the remainder of this paper.

First-person Vision (FPV)[4] also referred to as the Egocentric perspective which bears distinct characteristics from the more traditional Third-Person Vision (TPV) also referred to as the Exocentric perspective. Traditional third-person cameras usually give a wide and global view of the high-level appearances within a video. While first-person cameras can capture objects and people at a much finer level of granularity [5]. In the past, both these perspectives are studied independently. However, the relationship between FPV and TPV has yet to be fully explored. Chenyou Fan et.al.[6][7]in his work tries to find correspondence between multiple first- and third-person cameras for joint scene understanding. Since first-person data suffers from low scale and diversity Yanghao [8]in his work proposes a method that embeds key egocentric signals into the third-person video pre-training pipeline, in such a way that models could benefit from both the scale and the diversity of third-person video datasets creating strong video representations. Whereas, [9]proposed a framework to link first-person and third-person video data. All these works were efforts towards enhancing the action recognition and activity recognition tasks. Also, to address the challenge of fewer first-person datasets as compared to the third-person dataset.

While studying we did not come across any literature which compares and shares characteristics of first-person and third-person videos together. This paper tries to address the characteristics of these two perspectives along with their comparison. Also, we have treated this as a supervised learning problem that classifies videos using pre-trained models.

The key contributions of this paper includes:

- Presenting systematic and comparative study of egocentric and exocentric perspectives in the domain of computer vision. These perspectives are studied separately for various vision tasks like Action Recognition, Activity Recognition, Video Summarization, etc.
- Using Charades Dataset, we have classified the videos as first-person or third-person in a supervised manner using ResNet50 pre-trained model. To the best of our knowledge, we are the first ones to classify the video based on these perspectives.
- To classify videos in a more real-time mode, trained our model using 100 distinct daily life videos in a controlled environment of the Kitchen/Dining and Living room with objects like people, stairs, fridge, cupboard, bread, etc.

II. RELATED WORK

As per Starkville Daily News, by 2022, an average person is predicted to spend 100 minutes per day watching online videos with estimated 45 billion cameras [10]. Nowadays traditional cameras (third-person) are everywhere whether it is a shopping mall, hospital, museum, or home. Also, due to advancements in technology, wearable cameras are affordable and have become popular for recording lifelogging applications like surfing, hiking, law enforcement, and geriatric care (for old people)[11].

While reading papers we did not come across any literature that performs classification of videos based on the way they are captured i.e. first-person and third-person. Drawing inspiration from several computer vision tasks performed on first-person and third-person videos, this video classification experiment will give useful insights from the captured video data. Since these perspectives are independent of each other

International Journal on Recent and Innovation Trends in Computing and Communication
ISSN: 2321-8169 Volume: 11 Issue: 9s
DOI: https://doi.org/10.17762/ijritcc.v11i9s.7420
***Article Received**: 08 May 2023 **Revised:** 29 June 2023 **Accepted:** 28 July 2023*

they have been studied separately. That is, prior work for third-person videos for performing various computer vision tasks like action recognition[12], object recognition, activity recognition[12][13][14][15], and video summarization[11][16][17][2].

Alejandro[17] presented a systematic study on first-person video analysis between 1997 and 2014, highlighting, the most commonly used features, methods, challenges, and opportunities within the field. From a computer vision point, videos from first-person devices pose a lot of challenges because the camera is either head-mounted or chest mounted on the actor who is constantly moving, so the motion is highly non-linear and unpredictable[18]. It showcases objects, and people with whom an actor interacts and is centered in the camera view. In third-person videos the cameras are static so the video captures linear motion and the video is not shaky/blurry. On the other hand, the Objects, people, and their interactions may or may not be in the point of view.

In the first-person videos, the first-person cameras are mounted on the wearer's head so by observing optical flow it is possible to track the wearer's head motion [19] [20][21]. Since for different persons the variations of head motion and actions can be different which makes the problem more challenging to recognize actions [21]. The majority of the previous works have focused on third-person videos based on space-time interest points (STIP)[22], including local and global features based on spatio-temporal changes[12][22], key point tracking based trajectory features[23], motion changes based on depth information[24], and action features based on human pose changes[24].

TABLE I. COMPARATIVE STUDY OF FIRST-PERSON AND THIRD-PERSON VIDEOS

Characteristics	**First-person videos**	**Third-person videos**
Intention[25]	First-person videos are unconstrained in nature. Most of the videos are lifelogging so there is no specific intention in the recording. As the user moves his head the intention keeps changing. There is no control over what to record	In third-person videos, the cameraman decides the part of the scene that needs to be focused on.
Attention[26]	Spontaneous	Focused by Cameraman
Capturization	The first-person video captures the wearer's interaction with the objects, animals, and other people. It captures his ongoing activities and goals.	The third-person video captures everything in the field of vision i.e. it gives a global view of the high-level appearance and events in a scene[6].
Content[27][28]	In lifelogging videos, there is no control over the content as the wearer is continuously shooting his daily life experience. So there are many chances that the content will be repetitive and irrelevant. E.g. day to day activities of a college going student	The third-person camera's intentions are very focused and specific to the extent to record life experiences that are worth remembering E.g. birthday celebration
Quality	Many of the first-person videos are blurry due to the camera placed over the head or chest.	In this case, as the cameras are fixed most of the time the recordings are stabilized
General Applications	Sports and Adventure, Social life experiences	News, movies, music videos
Context	In case of the First-person videos generally, the context is unknown and diverse	In most of the Third,-person videos the context of the video is well defined beforehand and is the same throughout the videos

A general video can be represented by low-level features like color, shape, intensity, texture, SIFT, HOG (gradient-based), optical flow, and high-level features like geometric and model-based. These features are taken into consideration in computer vision and deep learning to perform tasks like Action Detection/Recognition, Object Detection, video summarization, Object Interaction, social interaction, etc. Research has focused on extracting features like gaze, ego-motion cues, hand, ego-action, and interaction for first-person videos due to head and body movements. It encodes unique characteristics driven by the camera wearer's attention and interaction with the surrounding.

III. METHODOLOGY

For video classification the pre-trained network ResNet-50 architecture is exploited with some design changes. The workflow described in Fig. 3 is used for the video classification task. The transfer learning mode and fine-tuning mode are used to train the model on the Charades dataset. In transfer learning, the model is trained in two modes. In the first mode, the pre-trained weights of the Resnet-50 model have been used without changing the weights of the layers in the model. During fine-tuning mode, the weights of the layers are updated during training. For the transfer learning, the model is created without the top node i.e. by removing the classifier layer from the pre-trained Resnet-50 model. The classifier is designed to classify the images as first-person or third-person. For the head of the Resnet-50 model, a combination of

International Journal on Recent and Innovation Trends in Computing and Communication
ISSN: 2321-8169 Volume: 11 Issue: 9s
DOI: https://doi.org/10.17762/ijritcc.v11i9s.7420
***Article Received**: 08 May 2023 **Revised:** 29 June 2023 **Accepted:** 28 July 2023*

AveragePooling2D, Dense, Dropout layers and a softmax classifier have been used as shown in the Fig. 4.

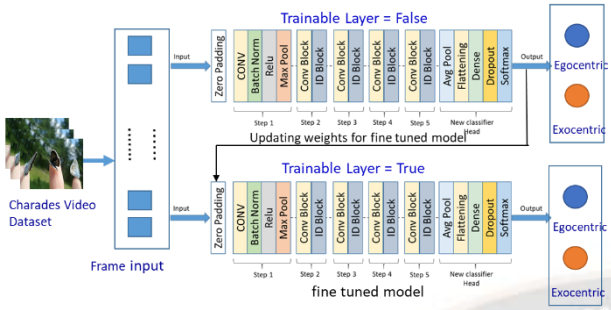

Figure 3. Video Classification Workflow

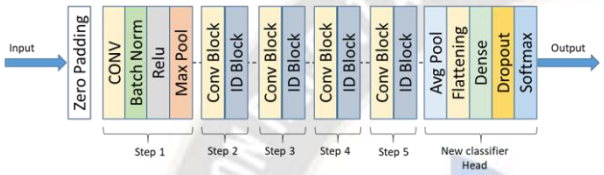

Figure 4. ResNet50 Modified Architecture

IV. RESULT AND DISCUSSION

A. *Dataset*

Charades[29] dataset is used for the experimentation which is composed of 9848 videos of daily indoor activities collected through Amazon Mechanical Turk. It contains almost 66,500 temporal annotations for 157 action classes, 41,104 labels for 46 object classes, and 27,847 textual descriptions of the videos. The data set has labeled each video as an egocentric (first-person) or an exocentric (third-person) video. From this dataset, 100 videos have been selected for the training that includes videos from the kitchen, and living room with a person performing certain actions like drinking water, opening a cupboard, etc. The validation and testing set has 10 videos selected independently of the Training Set.

B. *Experiments*

For the experimentation, every input video is converted into set of frames and as a representation 1 frame per second will be selected as an input to the model. E.g. if the video is recorded at 30 frames per second, then with 1 frame per sec, 1 frame out of every 15 frames will be used. This helps in reducing redundant data. Frames are resized to 224 x 224 so that it can be used by the pre-trained image classification CNN models like ResNet-50 and VGGNet.

The modified pre-trained Resnet-50 is trained on Google Colab with 12 GB of GPU. The hyperparameters tuned includes learning rate of 0.001, an epoch of 20, and batch size of 64 and the momentum of 0.9. The learning rate decay is used to reduce the learning rate as the number of epochs are increasing. This ensures that the model converges well during the training. Every test is repeated 20 times and the average accuracy and F1-score are calculated.

The weights calculated at the end of the warm up mode are used for initializing the weights for the fined tune mode. Now, in this fine-tuned mode, the training data is allowed to update the weights for all the layers with a reduced learning rate to converge well. The training is repeated on the same dataset with half the number of epochs and half the learning rate. Fig 4 and Fig 5 shows Training loss and accuracy curve for transfer learning and Fine tuning mode. Fig. 5 i.e. fine-tuned mode shows improvement of accuracy by about 2%.

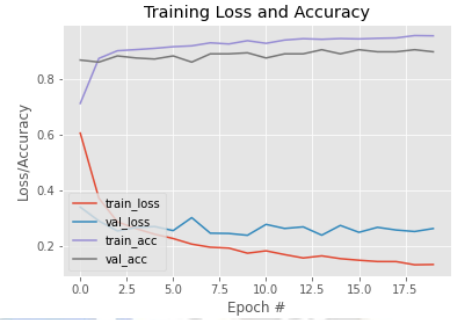

Figure 4. Transfer learning mode

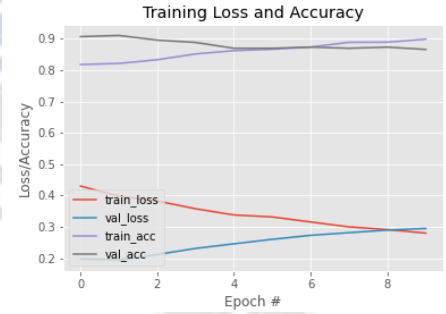

Figure 5. Fine tuning mode

C. *Evaluation Metrics*

The evaluation metrics used include accuracy, precision, recall, and F1-score. They are computed based on the predicted and actual relevance scores of video frames as described below. Since the data is unbiased accuracy is considered as the main metric for evaluating model's performance.

1) *Accuracy*

Accuracy is defined as the percentage of correct predictions for the input from test data. In this case it is calculated by observing True Positive (TP), True Negative (TN), False Positive (FP) and False Negative (FN) samples as below:

$$Accuracy = (TP + TN)/(TP + TN + FP + FN) \quad (1)$$

2) *Loss*

Here Binary cross entropy is used to calculate the error as below:

International Journal on Recent and Innovation Trends in Computing and Communication
ISSN: 2321-8169 Volume: 11 Issue: 9s
DOI: https://doi.org/10.17762/ijritcc.v11i9s.7420
***Article Received**: 08 May 2023 **Revised:** 29 June 2023 **Accepted:** 28 July 2023*

$$Binary\ Cross\ Entropy\ Loss = -\frac{1}{N}\sum_{i=1}^{N}(\log(pi)) \quad (2)$$

Where N is the Number of data samples, p is the predicted probability of the sample

3) Precision

Precision is the ratio of correctly predicted positive examples divided by the total number of positive examples that were predicted[30].

$$Precision = TP/(TP + FP) \quad (3)$$

4) Recall

A recall is a metric that quantifies the number of correct positive predictions made out of all positive predictions that could have been made[30]. It is calculated as below:

$$Recall = TP/(TP + FN) \quad (4)$$

5) F1-Score

F1-score is the harmonic mean of precision and recall. It is calculated as below:

$$F1 - score = 2TP/(2TP + FP + FN) \quad (5)$$

D. Results

This section describes the quantitative and qualitative results of the detailed experiments conducted on our models. We used ResNet50 as well as VGG model by conducting 20 test runs each in the warm up as well as fine-tuned mode and the average accuracy has been calculated. As shown in Fig. 6, we plotted a graph of average accuracy for ResNet 50 and VGG model. The VGG model is performing slightly better as compared to the ResNet 50. The observed accuracy for both the models with a fine-tuned mode is above 90%.

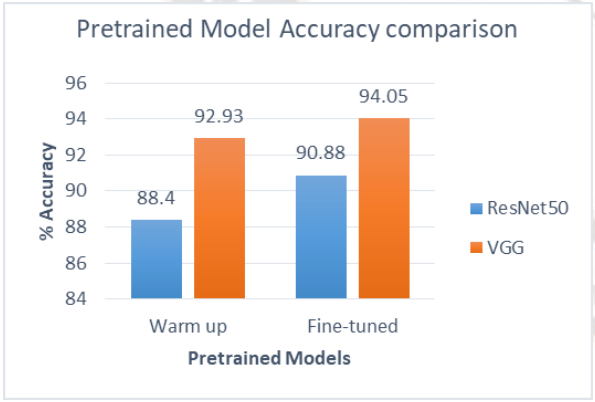

Figure 6. Model Accuracy comparison for video classification

For the prediction, the threshold value of 0.6 has been considered. While performing qualitative analysis we observed a few exceptional cases like some videos are shaky or blurred, some videos have been recorded in a poor lighting conditions also we saw some egocentric videos having steady human figure which normally is not a characteristic of the egocentric videos.

We took out 5 videos from each of these above mentioned exceptional cases. And have performed qualitative analysis. For the same we calculated prediction ratio as below:

$$Prection\ Ratio = no\ of\ frames\ of\ the\ class / total\ no\ of\ frames\ of\ that\ class \quad (6)$$

Quantitative and qualitative analysis results are as follow:

1) Low Lighting Condition videos

Since the videos are taken by humans in an uncontrolled manner some videos do face very low lighting conditions i.e. dark videos. Five such videos of different lengths have been given as an input as shown in the Table II. The table shows the class of the video as Ego/Exo along with the total frames of that video and predicted class for each frame as shown in Fig 7. From the same we have calculated the #of samples belonging to each class and calculated prediction ratio for each video. From Table II it is clear that third-person videos are classified correctly by the model irrespective of the lighting conditions while the first-person videos are also classified appropriately.

TABLE II. PREDICTION FOR LOW LIGHTING CONDITIONS

Video	Original Classification	Total frames	Ego	Exo	Prediction Ratio
6SL1REGO	EGO	16	15	1	94%
7PNXYEGO	EGO	29	23	6	79%
0HSFHEGO	EGO	32	32	0	100%
7GC82EGO	EGO	33	28	5	85%
7LJI0EGO	EGO	80	71	9	89%
6SL1R	Exo	15	0	15	100%
7PNXY	Exo	24	0	24	100%
0HSFH	Exo	33	0	33	100%
7GC82	Exo	23	0	23	100%
7LJI0	Exo	46	0	46	100%

For the first video i.e.65L1REGO we have produced the qualitative results as shown in the Fig. 7 likewise we have calculated it for each of the above mentioned videos and calculated prediction ratio. Higher the value better the classification.

International Journal on Recent and Innovation Trends in Computing and Communication
ISSN: 2321-8169 Volume: 11 Issue: 9s
DOI: https://doi.org/10.17762/ijritcc.v11i9s.7420
***Article Received**: 08 May 2023 **Revised:** 29 June 2023 **Accepted:** 28 July 2023*

Figure 7. Sample frames of the video 6SL1REGO with poor lighting conditions

2) Shaky/blurry videos

Shaky and blurry videos are mostly the first-person videos which are recorded in an uncontrolled environment like river rafting, hiking, daily lifelogging where a camera wearer is busy performing various activities. As shown in Table III, Shaky third-person videos are appropriately classified by the model with almost 100% prediction ratio and for first-person videos results are varying as per the input video. We have demonstrated here qualitative results for the 6JF1AEGO video where prediction ratio achieved is 100%.

TABLE III. PREDICTION FOR FLICKERING SHAKY VIDEOS

Video	Original Classification	Total frames	Ego	Exo	Prediction Ratio
6JF1AEGO	EGO	32	32	0	100%
1DUKWEGO	EGO	11	3	8	27%
1HDWSEGO	EGO	41	41	0	100%
1GK9YEGO	EGO	26	17	9	65%
3AFH1EGO	EGO	11	10	1	91%
6JF1A	Exo	35	0	35	100%
1DUKW	Exo	13	0	13	100%
1HDWS	Exo	36	6	36	100%
1GK9Y	Exo	26	1	25	96%
3AFH1	Exo	4	0	4	100%

Figure 8. Sample frames of the video 6JF1AEGO for shaky/blurry video

International Journal on Recent and Innovation Trends in Computing and Communication
ISSN: 2321-8169 Volume: 11 Issue: 9s
DOI: https://doi.org/10.17762/ijritcc.v11i9s.7420
***Article Received**: 08 May 2023 **Revised:** 29 June 2023 **Accepted:** 28 July 2023*

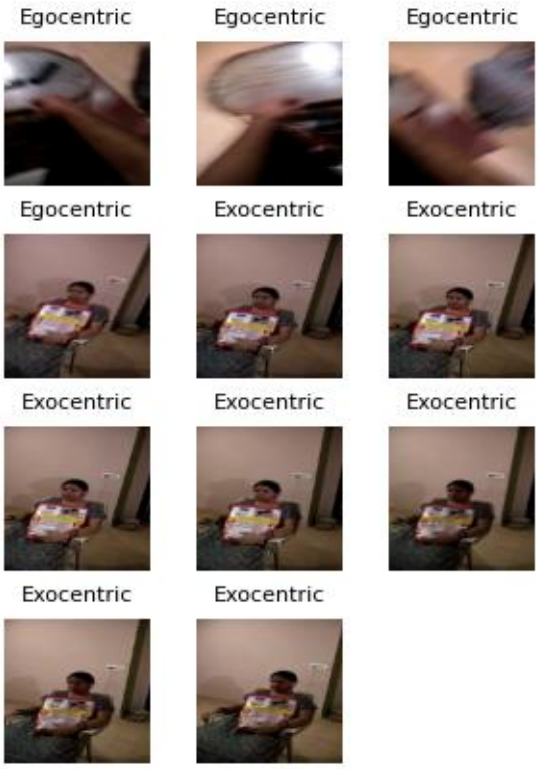

Figure 9. Sample frames of the video 1DUKWEGO for shaky video

As shown in the Fig. 9 in Table III, except for the video 1DUKWEGO all other predictions match with the ground truth, where the accuracy and the model can accurately predict the class. For the video 1DUKWEGO, the prediction class is wrongly classified. One reason could be the presence of a person sitting in the video.

3) Presence of a person in the video

While chit-chatting with a friend if an actor is wearing the head/chest-mounted camera it is challenging to predict if the video is first/third-person. As shown in the table IV, all third-person videos with person in the video have been classified accurately while first-person videos have not been classified correctly as per the ground truth.

TABLE IV. PREDICTION FOR HUMAN EGO VIDEOS

Video	Original Classification	Total frames	Ego	Exo	Prediction Ratio
6FN5GEGO	Ego	33	0	33	0%
7TMD7EGO	Ego	25	6	19	24%
8DSQ2EGO	Ego	37	7	30	19%
9IYNHEGO	Ego	33	0	33	0%
7YRTBEGO	Ego	41	4	38	10%
6FN5G	Exo	33	0	33	100%
7TMD7	Exo	22	0	22	100%
8DSQ2	Exo	37	0	37	100%
9IY9H	Exo	36	0	36	100%
7YRTB	Exo	39	0	39	100%

Figure 10. Sample frames of the video 6FN5GEGO for the presence of a person in the video

The presence of a person sitting in the first-person video poses a challenge for classifying the video since the images are superimposed, for the person to be present in the first-person videos. The model wrongly classifies all first-person videos and correctly classifies all third-person videos.

V. CONCLUSION

In a complex and highly dynamic environment, third-person videos capture a global view of high-level appearance while first-person videos capture ground-level aspects at a finer level of granularity[6] this may help in finding the correspondence for understanding various tasks like object tracking, scene understanding, activity recognition, etc. capturing distinct viewpoint.

We proposed a novel approach to classify videos based on the perspective of first-person/third-person. For the same the charades dataset is used to train and test the model performance. Our experiments exhibit good accuracy of 88.4% and an F1-score of 86.10%. We discussed a few

International Journal on Recent and Innovation Trends in Computing and Communication
ISSN: 2321-8169 Volume: 11 Issue: 9s
DOI: https://doi.org/10.17762/ijritcc.v11i9s.7420
***Article Received**: 08 May 2023 **Revised:** 29 June 2023 **Accepted:** 28 July 2023*

exceptional cases like shaky videos, videos shot in a dark environment, and the presence of a person in the first person video. In the earlier two situations model classifies correctly. However, in the presence of a person and a stand still video where a camera wearer is talking to the person the model is not able classify video appropriately. In the future, we intend to improve our model to address this situation.

REFERENCES

[1] S. A. Behrostaghi, "Relating First-person and Third-person Vision," 2018.

[2] H. A. Ghafoor, A. Javed, A. Irtaza, H. Dawood, H. Dawood, and A. Banjar, "Egocentric Video Summarization Based on People Interaction Using Deep Learning," Math. Probl. Eng., vol. 2018, 2018, doi: 10.1155/2018/7586417.

[3] A. Betancourt, P. Morerio, C. S. Regazzoni, and M. Rauterberg, "An Overview of First Person Vision and Egocentric Video Analysis for Personal Mobile Wearable Devices," Circuits Syst. Video Technol., vol. (Under Rev, no. (Under Review), pp. 744–760, 2014.

[4] M. Devyver, a Tsukada, and T. Kanade, "A wearable device for first person vision," 3rd Int. Symp. Qual. Life …, pp. 1–6, 2011, [Online].Available: http://scholar.google.com/scholar?hl=en&btnG=Search&q=intitle:A+Wearable+Device+for+First+Person+Vision#7.

[5] G. Liu, H. Tang, H. Latapie, and Y. Yan, "Exocentric to Egocentric Image Generation Via Parallel Generative Adversarial Network," ICASSP, IEEE Int. Conf. Acoust. Speech Signal Process. - Proc., vol. 2020-May, pp. 1843–1847, 2020, doi: 10.1109/ICASSP40776.2020.9053957.

[6] C. Fan et al., "Identifying First-person Camera Wearers in Third-person Videos," no. 1.

[7] M. Xu, C. Fan, Y. Wang, M. S. Ryoo, and D. J. Crandall, "Joint Person Segmentation and Identification in Synchronized First- and Third-Person Videos," Lect. Notes Comput. Sci. (including Subser. Lect. Notes Artif. Intell. Lect. Notes Bioinformatics), vol. 11205 LNCS, pp. 656–672, 2018, doi: 10.1007/978-3-030-01246-5_39.

[8] Y. Li, T. Nagarajan, B. Xiong, and K. Grauman, "Ego-Exo: Transferring Visual Representations from Third-person to First-person Videos," Proc. IEEE Comput. Soc. Conf. Comput. Vis. Pattern Recognit., pp. 6939–6949, 2021, doi: 10.1109/CVPR46437.2021.00687.

[9] G. A. Sigurdsson, A. Gupta, C. Schmid, A. Farhadi, and K. Alahari, "Actor and Observer: Joint Modeling of First and Third-Person Videos," Proc. IEEE Comput. Soc. Conf. Comput. Vis. Pattern Recognit., pp. 7396–7404, 2018, doi: 10.1109/CVPR.2018.00772.

[10] "135 Video Marketing Statistics You Can't Ignore in 2022." https://invideo.io/blog/video-marketing-statistics/ (accessed Dec. 20, 2022).

[11] A. Rathore, C. Arora, P. Nagar, and C. V. Jawahar, "Generating 1 minute summaries of day long egocentric videos," MM 2019 - Proc. 27th ACM Int. Conf. Multimed., pp. 2305–2313, 2019, doi: 10.1145/3343031.3350880.

[12] D. Das Dawn and S. H. Shaikh, "A comprehensive survey of human action recognition with spatio-temporal interest point (STIP) detector," Vis. Comput., vol. 32, no. 3, pp. 289–306, 2016, doi: 10.1007/s00371-015-1066-2.

[13] Q. Li, R. Gravina, Y. Li, S. H. Alsamhi, F. Sun, and G. Fortino, "Multi-user activity recognition: Challenges and opportunities," Inf. Fusion, vol. 63, pp. 121–135, 2020, doi: 10.1016/j.inffus.2020.06.004.

[14] A. B. Sargano, P. Angelov, and Z. Habib, "A comprehensive review on handcrafted and learning-based action representation approaches for human activity recognition," Appl. Sci., vol. 7, no. 1, 2017, doi: 10.3390/app7010110.

[15] D. Surie, T. Pederson, F. Lagriffoul, L. E. Janlert, and D. Sjölie, "Activity recognition using an egocentric perspective of everyday objects," Lect. Notes Comput. Sci. (including Subser. Lect. Notes Artif. Intell. Lect. Notes Bioinformatics), vol. 4611 LNCS, pp. 246–257, 2007, doi: 10.1007/978-3-540-73549-6_25.

[16] H. I. Ho, W. C. Chiu, and Y. C. F. Wang, "Summarizing first-person videos from third persons' points of views," Lect. Notes Comput. Sci. (including Subser. Lect. Notes Artif. Intell. Lect. Notes Bioinformatics), vol. 11219 LNCS, pp. 72–89, 2018, doi: 10.1007/978-3-030-01267-0_5.

[17] A. Betancourt, P. Morerio, C. S. Regazzoni, and M. Rauterberg, "The evolution of first person vision methods: A survey," IEEE Trans. Circuits Syst. Video Technol., vol. 25, no. 5, pp. 744–760, 2015, doi: 10.1109/TCSVT.2015.2409731.

[18] S. Bambach, "A Survey on Recent Advances of Computer Vision Algorithms for Egocentric Video," 2015, [Online]. Available: http://arxiv.org/abs/1501.02825.

[19] M. S. Ryoo, B. Rothrock, and L. Matthies, "Pooled motion features for first-person videos," Proc. IEEE Comput. Soc. Conf. Comput. Vis. Pattern Recognit., vol. 07-12-June, no. Figure 1, pp. 896–904, 2015, doi: 10.1109/CVPR.2015.7298691.

[20] D. Thapar, C. Arora, and A. Nigam, "Is Sharing of Egocentric Video Giving Away Your Biometric Signature?," Lect. Notes Comput. Sci. (including Subser. Lect. Notes Artif. Intell. Lect. Notes Bioinformatics), vol. 12362 LNCS, pp. 399–416, 2020, doi: 10.1007/978-3-030-58520-4_24.

[21] S. Narayan, M. S. Kankanhalli, and K. R. Ramakrishnan, "Action and interaction recognition in first-person videos," IEEE Comput. Soc. Conf. Comput. Vis. Pattern Recognit. Work., pp. 526–532, 2014, doi: 10.1109/CVPRW.2014.82.

[22] I. Laptev and T. Lindeberg, "Space-time interest points," Proc. IEEE Int. Conf. Comput. Vis., vol. 1, pp. 432–439, 2003, doi: 10.1109/iccv.2003.1238378.

[23] H. Wang and C. Schmid, "Action recognition with improved trajectories," Proc. IEEE Int. Conf. Comput. Vis., pp. 3551–3558, 2013, doi: 10.1109/ICCV.2013.441.

[24] M. Li, H. Leung, and H. P. H. Shum, "Human action recognition via skeletal and depth based feature fusion," Proc. - Motion Games 2016 9th Int. Conf. Motion Games, MIG 2016, pp. 123–132, 2016, doi: 10.1145/2994258.2994268.

[25] A. Garcia, C. Tan, J. Lim, and A. Tan, "Summarization of Egocentric Videos : A Comprehensive Survey," no. section

International Journal on Recent and Innovation Trends in Computing and Communication
ISSN: 2321-8169 Volume: 11 Issue: 9s
DOI: https://doi.org/10.17762/ijritcc.v11i9s.7420
***Article Received**: 08 May 2023 **Revised:** 29 June 2023 **Accepted:** 28 July 2023*

IV.

[26] F. Martinez, A. Carbone, and E. Pissaloux, "Combining first-person and third-person gaze for attention recognition," 2013 10th IEEE Int. Conf. Work. Autom. Face Gesture Recognition, FG 2013, 2013, doi: 10.1109/FG.2013.6553735.

[27] M. Dimiccoli, Computer Vision for Egocentric (First-Person) Vision. Elsevier Ltd, 2018.

[28] C. Tan, H. Goh, V. Chandrasekhar, L. Li, and J. H. Lim, "Understanding the nature of first-person videos: Characterization and classification using low-level features," IEEE Comput. Soc. Conf. Comput. Vis. Pattern Recognit. Work., pp. 549–556, 2014, doi: 10.1109/CVPRW.2014.85.

[29] G. A. Sigurdsson, G. Varol, X. Wang, A. Farhadi, I. Laptev, and A. Gupta, "Hollywood in homes: Crowdsourcing data collection for activity understanding," Lect. Notes Comput. Sci. (including Subser. Lect. Notes Artif. Intell. Lect. Notes Bioinformatics), vol. 9905 LNCS, pp. 510–526, 2016, doi: 10.1007/978-3-319-46448-0_31.

[30] L. Shu, H. Zhang, Y. You, Y. Cui, and W. Chen, "Towards fire prediction accuracy enhancements by leveraging an improved naïve bayes algorithm," Symmetry (Basel)., vol. 13, no. 4, pp. 1–14, 2021, doi: 10.3390/sym13040530.