

## DATA ARTICLE

# A Comprehensive Video Dataset for Multi-Modal Recognition Systems

Anand Handa<sup>1</sup>, Rashi Agarwal<sup>2</sup> and Narendra Kohli<sup>3</sup>

<sup>1</sup> Dr. APJ Abdul Kalam Technical University, Kanpur, IN

<sup>2</sup> Department of IT, University Institute of Engineering and Technology, Kanpur, IN

<sup>3</sup> Department of CSE, Harcourt Butler Technical University, Kanpur, IN

Corresponding author: Anand Handa (anandhanda1986@gmail.com)

This paper presents a comprehensive, highly defined and fully labelled video dataset. This dataset consists of videos related to 67 different subjects. The videos contain similar text and the text contains digits from 1 to 20 recited by 67 different subjects using the same experimental setup. This dataset can be used as a unique resource for researchers and analysts for training deep neural networks to build highly efficient and accurate recognition models in various domains of computer vision such as face recognition model, expression recognition model, speech recognition model, text recognition, etc. In this paper, we also train models related to face recognition and speech recognition on our dataset and also compare the results with the publically available datasets to show the effectiveness of our dataset. The experimental results show that our comprehensive dataset is more accurate than other dataset on which the models are tested.

**Keywords:** Machine leaning; Deep learning; video datasets; Convolutional Neural Network

## 1. Specifications Table

Subject area	<i>Image Processing, Computer Vision, Machine learning and Deep learning.</i>
More specific subject area	<i>Feature Extraction, Speech recognition and Text Recognition.</i>
Type of data	<i>Images, Audio files, Tables and Figures.</i>
How data was acquired (Experimental Setup)	<i>Original videos were captured at University Institute of Engineering Technology, Kanpur using a Canon Eos 1200D 18MP Digital SLR Camera with 18–55 mm and 55–250 mm lens in a highly sophisticated and noise free experimental laboratory.</i>
Data format	<i>Videos are in .MOV format, Frames are in .jpg format, audio files are in .wav format, Wave graphs for are in .png format.</i>
Experimental factors	<i>The video samples that have been generated for various subjects are De-noised by using Neat Video (Other, 2019).</i>
Experimental features	<i>Extract various biometric traits for every subject such as frames, boundary box coordinates, audio of the entire video of a subject, the audio wave signal for entire video length, split audio of text spoken by subject, and split audio waveform.</i>
Data source location	<i>University Institute of Engineering Technology, Kanpur, India.</i>
Data accessibility	<i>The dataset is accessible and it is publicly and freely available for any research, educational, and purposes.</i>

## 2. Value of the Data

- Today Human-Machine Intelligence (Lai., 2012) and computer vision (Frizzell et al., 2018) become pervasive because of their variety of applications ranging from medical informatics (Goodfellow et al., 2016) to face recognition (Sun, Wu, and Hoi, 2018) and building of smart surveillance system (Memos et al., 2018). The core to many of such applications is image classification and recognition

(Simonyan and Zisserman, 2014). Hence, our dataset (Handa, Agarwal, and Kohli, 2018) is a valuable resource for use by vision and learning community.

- Our dataset contains the videos, frames of a video subject, its boundary coordinates, audio format file of entire video subject, text (digits 1 to 20) in an audio format file and its wave format.
- Our dataset allows researchers to apply various neural network models, machine learning algorithms and deep neural network models in various domains like face recognition, expression recognition, and text and speech recognition (Shekar Naganna et al., 2018) for building robust recognition models.
- The various features mentioned in the Table 1.1 are extracted for a particular subject as a demonstration and the scripts by which the similar data can be generated for each subject is also provided in our dataset.

### 3. Data

A high definition video in .MOV format for each subject is captured using an experimental setup as mentioned in Table 1.1. Every subject in a video recites similar text starting from digits 1 to 20 to maintain uniformity in the data. The following are the various data values in multiple formats that can be used for further modeling:

#### 3.1. Frames

We collect the frames of a video corresponding to each subject. The frames are stored in .jpg file format. The code is customized in such a way that it generates frames for each video by giving a suitable path. We generate the same for a sample video and **Figure 1** shows the frames of a sample video *DSC\_0020.MOV* from our dataset. We show some example frames out of the total 1011 frames generated for the mentioned video *DSC\_0020.MOV*.

#### 3.2. Boundary Coordinates

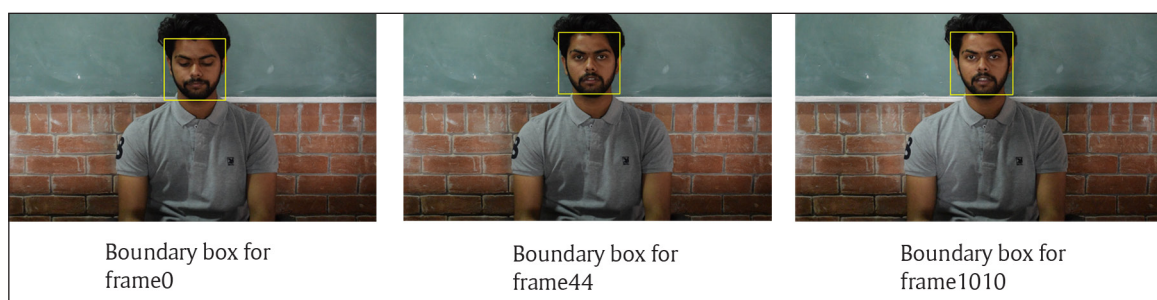
The frames from the sample video are used to find the boundary coordinates of the face corresponding to a subject in our dataset. We create .csv files for each frame and .csv file contains coordinates for each boundary box. The fields of a boundary box are LowerLeft(X), LowerLeft(Y), UpperLeft(X), UpperLeft(Y), UpperRight(X), UpperRight(Y), LowerRight(X), LowerRight(Y) corresponding to each frame. **Figure 2** shows some samples of the frames with boundary boxes. **Table 1** shows the sample .csv format for *DSC\_0020.MOV* video containing boundary box coordinates.

#### 3.3. Audio file in .wav format

In this module, we generate a complete audio file for every subject in our dataset which can be further used by the researchers for audio recognition models or for some audio detection related learning models. The



**Figure 1:** Frames generated for a sample video *DSC\_0020.MOV*.



**Figure 2:** Boundary box for the frames generated for a sample video.

files are stored in the *.wav* format for a sample video, and the script is customized to find out the same for every subject video in our dataset.

### 3.4. Wave Forms of Audio files

The audio files in the *.wav* format are generated for each video, and then we create the waveforms. This is done to help researchers in getting the peak amplitudes so that they can compare these amplitudes with the individual peak amplitudes of the text that is recited in the video. This is an intermediate step for text and voice recognition models. The graphs corresponding to the full-length videos are available in a *.jpeg* format. **Figure 3** shows a sample waveform for a sample video.

### 3.5. Split text files

In this step, we generate the data corresponding to the text that is recited by each subject in our dataset. The dataset consists of digits 1 to 20 recited by each subject. This data is of importance in designing learning models and architectures for text recognition or to identify that who has spoken a particular digit? And precisely which digit, out of 1 to 20 in the entire video length? These files in the *.wav* format for a sample video *DSC\_0020.MOV* and the customized scripts are available in our dataset.

### 3.6. Split text waveforms

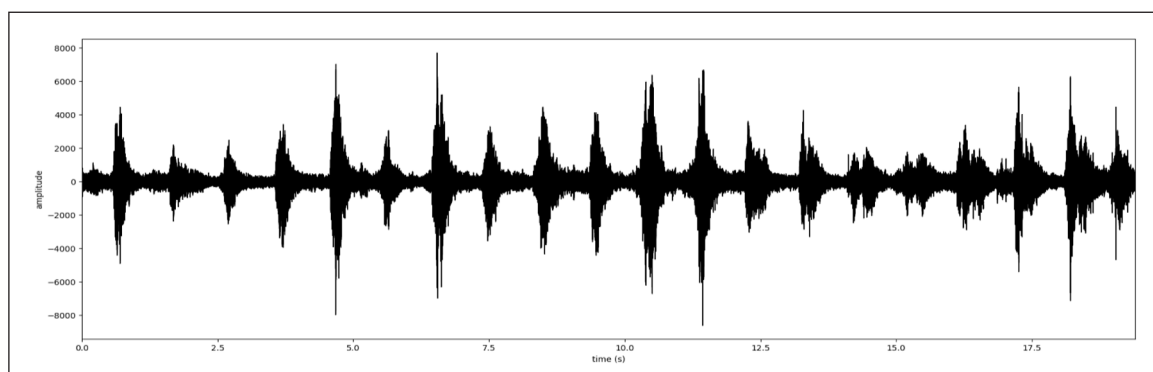
In this step, we generate graphs corresponding to each text present in the video, i.e., digits from 1 to 20 in the above step. For building more accurate and efficient text or speech recognition models, we need to compare the waveforms of the full-length video with the waveforms of split text waveforms. It helps us in gathering more information about the individual amplitudes of the text present in the video and it also help to identify any noise if present in any video file. The graphs are generated for each split text file and stored in *.jpeg* format. The X-axis represents the time and Y-axis represents the amplitude. **Figures 4** and **5** illustrates a sample graph for digits 1 and 2 respectively from the video sample *DSC\_0020.MOV*.

## 4. Experimental Design and Analysis of Our Dataset

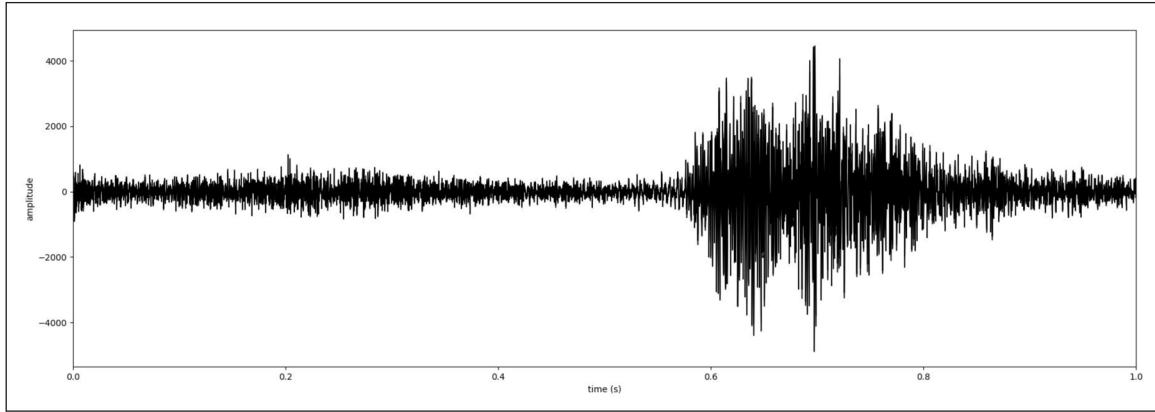
To analyze the effectiveness of our video dataset, we utilize the fundamentals of convolutional neural network (CNN) (Krizhevsky, Sutskever, and Hinton 2012) model for face recognition (Acharya et al. 2018) and speech recognition. We train a CNN model using high-resolution images present as frames in our dataset. **Table 2** shows the configuration of the CNN whose output later is variable and depends on the task to be performed. We then select 70% of our dataset randomly and use it as a training set. Further, to evaluate the performance, rest of the dataset is used. For speech recognition model, we train the CNN using waveforms. There are a total of 1340 waveforms related to 67 different subjects and it corresponds to digits 1 to 20 in

**Table 1:** *.csv* format for the boundary box coordinates of each frame for sample video *DSC\_0020.MOV*.

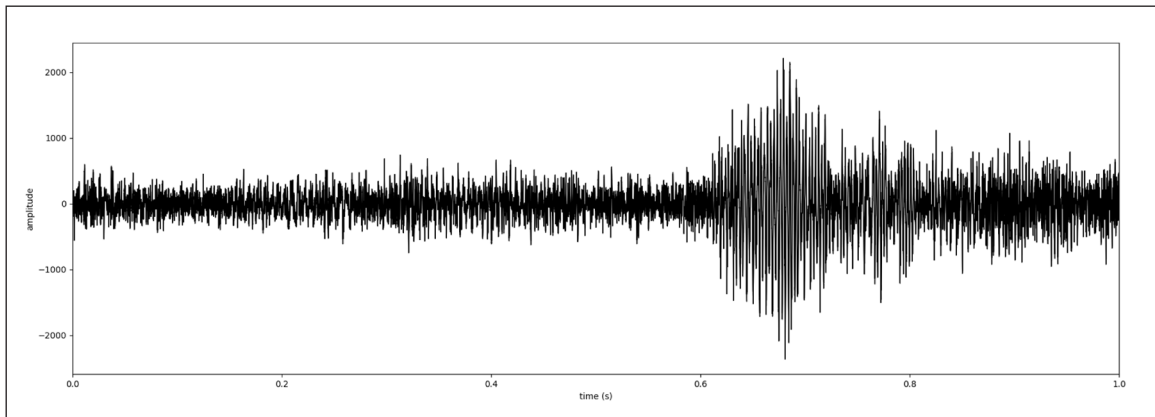
Frames	Lower Left (X)	Lower Left (Y)	Upper Left (X)	Upper Left (Y)	Upper Right (X)	Upper Right (Y)	Lower Right (X)	Lower Right (Y)
frame0.jpg	821	134	821	450	1137	450	1137	134
frame1.jpg	822	135	822	448	1135	448	1135	135
frame1010.jpg	811	108	811	421	1124	421	1124	108



**Figure 3:** Wave form of a sample video *DSC\_0020.MOV*.



**Figure 4:** Wave form for digit 1 recited in *DSC\_0020.MOV*.

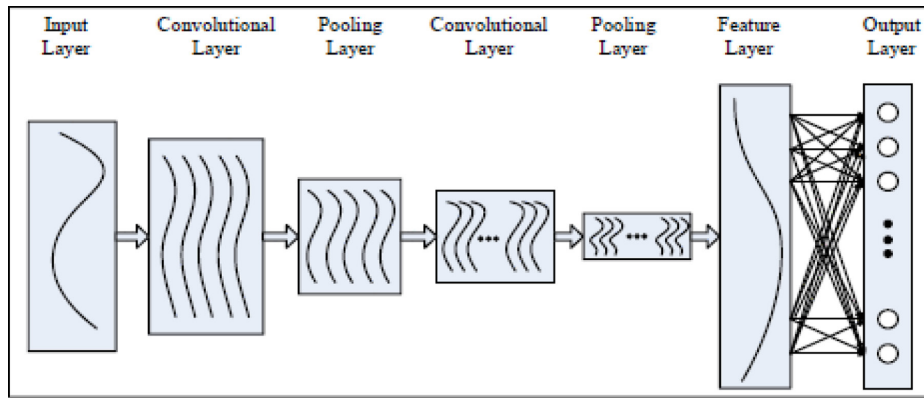


**Figure 5:** Wave form for digit 2 recited in *DSC\_0020.MOV*.

**Table 2:** Configuration of Convolutional Neural Network.

Layers	Filter Size	Strides	No. of filters
Convolution Layer 1	$5 \times 5$	1	32
Pooling Layer 1	$2 \times 2$	2	–
Convolution Layer 2a	$1 \times 1$	2	64
Convolution Layer 2a_1	$3 \times 3$	1	64
Convolution Layer 2b	$3 \times 1$	1	64
Convolution Layer 2b_1	$1 \times 3$	1	64
Pool 2b	$2 \times 2$	2	–
Convolution Layer 2c	$1 \times 1$	2	64
Concatenate	–	–	192
Pool 2	$2 \times 2$	2	–
Fully Connected Layer 1	–	–	1024
Fully Connected Layer 2	–	–	1024

our dataset. We split the dataset into a ratio of 70:30 for training and testing our model. **Figure 6** shows the architecture of our CNN model for speech recognition. The training of the model needs a high computational power and support. Hence, we use NVidia Tesla K80 GPU for our evaluations. **Table 3** shows the accuracy and training loss for the face recognition model on our dataset. Similarly, **Table 4** shows the accuracy and training loss for the speech recognition model on our dataset. **Figure 7** and **Figure 8** show the results for accuracy and training loss graph for both the recognition models on our datasets over 500 epochs. The x-axis and the y-axis correspondingly denote the number of epochs and accuracy/training loss.



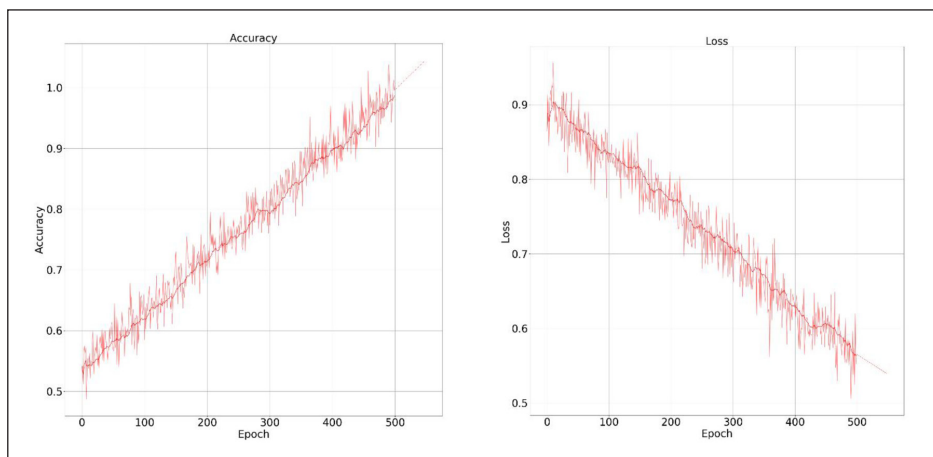
**Figure 6:** Architecture of CNN for speech recognition model (Zhao et al. 2017).

**Table 3:** Face Recognition Model Results on Our Dataset.

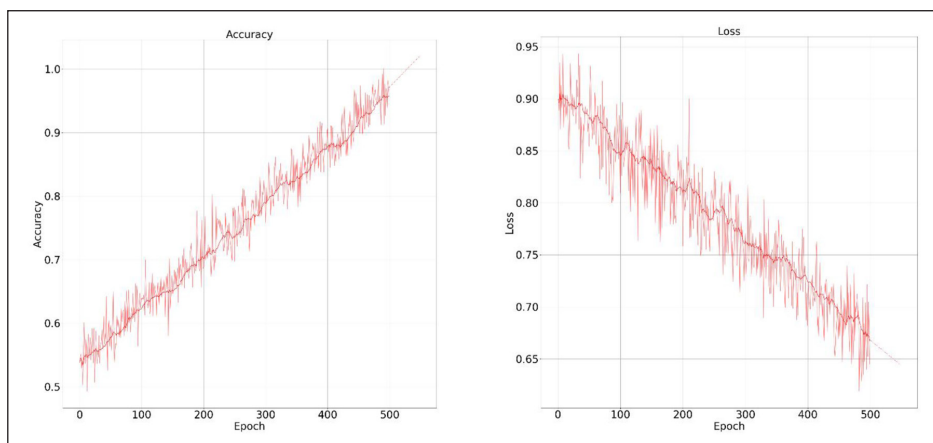
Dataset	Training/Testing Percentage	Accuracy	Training Loss
Our Dataset (Handa, Agarwal, and Kohli, 2018)	70% and 30%	99.14%	0.56%

**Table 4:** Speech Recognition Model Results on Our Dataset.

Dataset	Training/Testing Percentage	Accuracy	Training Loss
Our Dataset (Handa, Agarwal, and Kohli, 2018)	70% and 30%	96.42%	0.67%



**Figure 7:** Accuracy and training loss results graph on our dataset for face recognition.



**Figure 8:** Accuracy and training loss results graph on our dataset for speech recognition.



### 5. Effectiveness Comparison of Our Dataset with Open Source Datasets

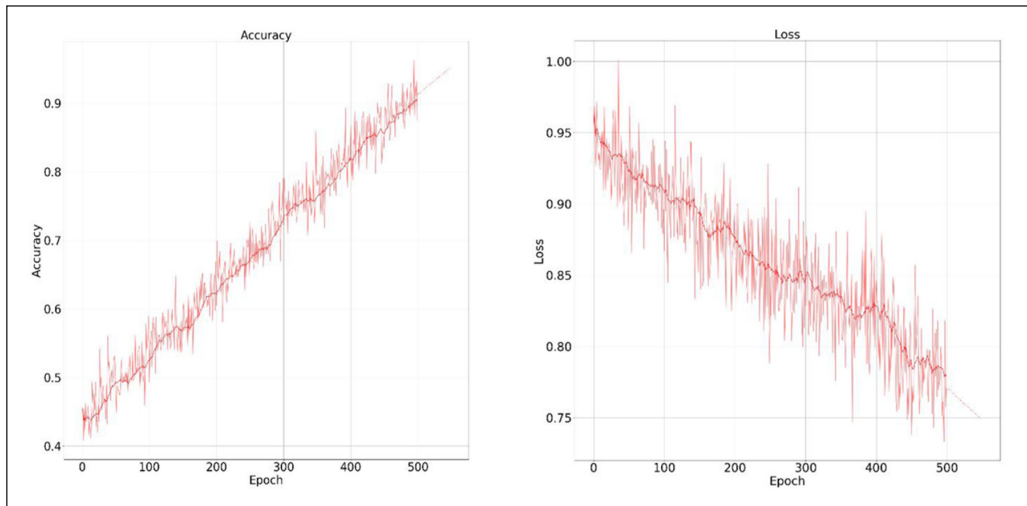
In this paper, we also compare the effectiveness of our dataset with other publicly available datasets. For face recognition comparison, we use the JAFFE dataset (Lyons et al., 1998). The dataset consists of 213 images posed by 10 Japanese models. Similarly, for speech recognition, we use Free Spoken Digit Dataset (FSDD) (Jackson et al., 2018), which is an audio/speech dataset that contains recordings of digits in .wav format. The recordings are done at 8kHz and do not contain noise. The total number of recordings is 1500, corresponding to 3 subjects of spoken digits from 0 to 9. Each digit is recited 50 times by a single subject. **Table 5** and **Table 6** show the results of accuracy and training loss on the JAFFE dataset and FSDD dataset for face recognition and speech recognition models, respectively. **Figures 9** and **10** shows the corresponding results for accuracy and training loss on JAFFE and FSDD datasets. To evaluate the performance of our

**Table 5:** Face Recognition Model Results on JAFFE Dataset.

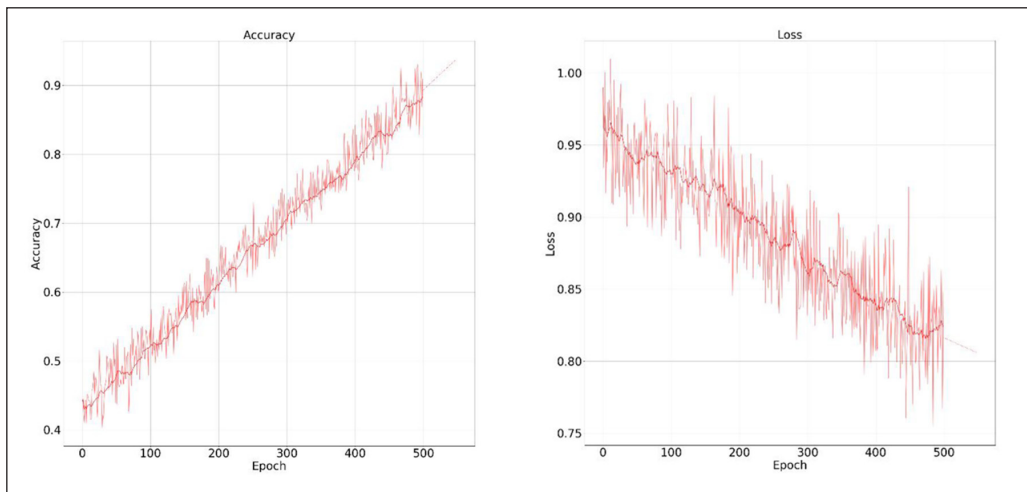
Dataset	Training/Testing Percentage	Accuracy	Training Loss
JAFFE Dataset (Lyons et al., 1998)	70% and 30%	92.1%	0.78%

**Table 6:** Speech Recognition Model Results on FSDD Dataset.

Dataset	Training/Testing Percentage	Accuracy	Training Loss
FSDD Dataset Jackson et al., 2018)	70% and 30%	89.2%	0.81%



**Figure 9:** Accuracy and training loss results graph on JAFFE dataset for face recognition.



**Figure 10:** Accuracy and training loss results graph on FSDD dataset for speech recognition.

face and speech recognition models which are trained on our dataset, we test the face recognition model using JAFFE dataset and achieve an accuracy of 93.04% as shown in **Table 7**. Similarly, for the speech recognition model, we test using the FSDD dataset and achieve an accuracy of 90.11%, as shown in **Table 8**. **Figure 11** shows the test accuracy versus the number of epoch graphs for face recognition and speech recognition models.

## 6. Conclusion

This paper presents a video dataset of 67 subjects in which all subjects recite same text, i.e. digits from 1 to 20. We present the ways to extract useful information such as video frames in *.jpeg* format, full length audio of the corresponding video in *.wav* format, spoken digits (1–20) audio in *.wav* format, the waveforms for full length video and for the spoken digits. To show the effectiveness of data, we trained the CNN models related to face recognition and speech recognition to test the accuracy. The results show that the dataset is more accurate as compared to the results for two publically available datasets, JAFFE dataset is used to show the effectiveness for face recognition and FSDD dataset is used to show the effectiveness for speech recognition. This is a comprehensive video dataset and to best of our knowledge there is no publically available dataset that provides all the data values related to a video that we presented in our paper.

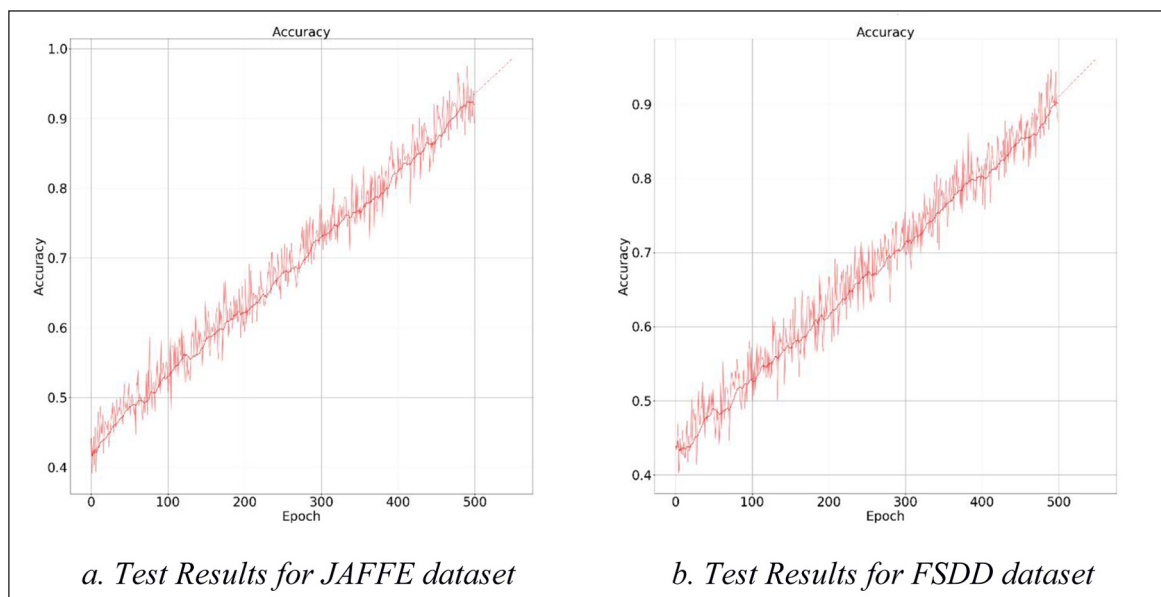
To mention here, there are a few limitations. One is that, video dataset consists of text information for digits 1 to 20. More text can be incorporated in future work. Another limitation is that the video contains only two class of emotions – happy and neutral. More emotions like anger, disgust, fear, sad may be added in future.

**Table 7:** Face Recognition test results of our trained model for JAFFE dataset.

Dataset		Training/Testing Percentage	Accuracy
Training	Testing		
Our Dataset (Handa, Agarwal, and Kohli, 2018)	JAFFE Dataset (Lyons et al., 1998)	70% and 30%	93.04%

**Table 8:** Speech Recognition test results of our trained model for FSDD dataset.

Dataset		Training/Testing Percentage	Accuracy
Training	Testing		
Our Dataset (Handa, Agarwal, and Kohli, 2018)	FSDD Dataset Jackson et al., 2018)	70% and 30%	90.11%



**Figure 11:** Test accuracy of face and speech recognition model trained on our dataset.

## Data Accessibility Statement

Anand Handa, Dr. Rashi Agarwal, and Prof. Narendra Kohli (2018). A comprehensive video dataset for Multi-Modal Recognition Systems [Data set]. Zenodo <http://doi.org/10.5281/zenodo.1492227>

## Ethics and Consent

All procedures performed in studies involving human participants were in accordance with the ethical standards of the institutional committee. A copy of consent from the institutional academic council of UIET, CSJM University, Kanpur is attached for reference.

## Acknowledgements

I want to acknowledge the students, laboratory technical staff, and summer interns who helped us in collecting the dataset at computer laboratory of University Institute of Engineering Technology, Kanpur, India where the experimental results are generated. We acknowledge the Google Inc. for providing NVidia Tesla K80 GPU support. NVidia Tesla K80 GPU has provided us with an environment to perform large computations in relatively less time.

## Competing Interests

The authors have no competing interests to declare.

## References


- Acharya, D**, et al. 2018. "Covariance pooling for facial expression recognition". In: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops*, 367–374. DOI: <https://doi.org/10.1109/CVPRW.2018.00077>
- Frizzell, K**, et al. 2018. "Modifiable Intuitive Robot Controller: Computer Vision-Based Controller for Various Robotic Designs". In: *SoutheastCon2018. IEEE*, 1–7. DOI: <https://doi.org/10.1109/SECON.2018.8479064>
- Goodfellow, I**, et al. 2016. *Deep learning*. vol. 1. Cambridge: MIT press.
- Handa, A, Agarwal, R, Dr and Kohli, N, Prof.** Dec. 2018. "A comprehensive video dataset for Multi-Modal Recognition Systems". DOI: <https://doi.org/10.5281/zenodo.1492227>
- Jackson, Z**, et al. 2018. Jakobovski/free-spoken-digit-dataset v1. 0.7.
- Krizhevsky, A, Sutskever, I and Hinton, GE.** 2012. "Imagenet classification with deep convolutional neural networks". In: *Advances in neural information processing systems*, 1097–1105.
- Lai, Y.** 2012. Human-machine interaction system. *US Patent App. 13/086,394*.
- Lyons, M**, et al. 1998. "Coding facial expressions with gabor wavelets". In: *Proceedings Third IEEE international conference on automatic face and gesture recognition. IEEE*, 200–205. DOI: <https://doi.org/10.1109/AFGR.1998.670949>
- Memos, VA.** et al. 2018. "An efficient algorithm for media-based surveillance system (EAMSuS) in IoT smart city framework". In: *Future Generation Computer Systems*, 83: 619–628. DOI: <https://doi.org/10.1016/j.future.2017.04.039>
- Other.** 2019. NeatVideo. <https://www.neatvideo.com/>.
- Shekar Naganna, S, Seth, A, Tomar, V and Yellareddy, SR.** 2018. Face recognition in big data ecosystem using multiple recognition models. *U.S. Patent Application 15/957,884*.
- Simonyan, K and Zisserman, A.** 2014. "Very deep convolutional networks for large-scale image recognition". In: *arXiv preprint arXiv*, 1409.1556
- Sun, X, Wu, P and Hoi, SCH.** 2018. "Face detection using deep learning: An improved faster RCNN approach". In: *Neurocomputing*, 299: 42–50. DOI: <https://doi.org/10.1016/j.neucom.2018.03.030>
- Zhao, B**, et al. 2017. "Waveforms classification based on convolutional neural networks". In: *2017 IEEE 2nd Advanced Information Technology, Electronic and Automation Control Conference (IAEAC). IEEE*, 162–165. DOI: <https://doi.org/10.1109/IAEAC.2017.8053998>



**How to cite this article:** Handa, A, Agarwal, R and Kohli, N. 2019. A Comprehensive Video Dataset for Multi-Modal Recognition Systems. *Data Science Journal*, 18: 55, pp.1–9. DOI: <https://doi.org/10.5334/dsj-2019-055>

**Submitted:** 20 November 2018    **Accepted:** 21 October 2019    **Published:** 08 November 2019

**Copyright:** © 2019 The Author(s). This is an open-access article distributed under the terms of the Creative Commons Attribution 4.0 International License (CC-BY 4.0), which permits unrestricted use, distribution, and reproduction in any medium, provided the original author and source are credited. See <http://creativecommons.org/licenses/by/4.0/>.

 *Data Science Journal* is a peer-reviewed open access journal published by Ubiquity Press.

**OPEN ACCESS** 