



How does hand gestures in videos impact social media engagement - Insights based on deep learning

Kartik Anand^a, Siddhaling Urolagin^{a,b,*}, Ram Krishn Mishra^a

^a Department of Computer Science, BITS Pilani, Dubai Campus, Dubai International Academic City, PO Box 345055, Dubai

^b APP Centre for AI Research (APPCAIR), BITS Pilani, Dubai Campus, Dubai International Academic City, PO Box 345055, Dubai



ARTICLE INFO

Keywords:

Computer vision
Hand gesture
Deep learning
Hand gesture recognition
Neural networks
Convolutional neural network

ABSTRACT

With the fast improvement and development in deep learning and computer vision, the interaction between humans and computers is becoming immense. This research work aims at recognizing hand gestures performed in TEDx videos and analyse the relationship between gestures and user engagement. Here we propose a technique based on deep learning, Convolutional Neural Network (CNN), to recognize hand gestures from a video or image input. ResNeXt-101 model is used for the classification of hand gestures. Here we also used images from the Twenty Bn dataset for some gestures while we also collected gesture images from TEDx videos. Our experiments obtained high accuracy for gesture recognition on training as 95% to 99% and 94.35% for testing. Firstly a gesture is identified in each frame and then classified. We also created a count function that helped us count the number of times a gesture was performed and help us analyse these talks in a better way. Two experimental studies were carried out to analyse viewer engagement: one, the effect of suitable gestures on the viewer count, and secondly, suitable gestures on the sentiment of the viewer. Interesting results are observed that suitable gestures from talkers have an impact on increasing positive review and viewer count.

Introduction

Due to the advancements in deep learning and expanded accessibility of video data with an extension of worldwide video networks, video analytics has changed conventional computer vision to utilize powerful deep learning techniques. Regarding video data analysis, deep learning techniques are used to teach the system to distinguish individuals and objects in a video. By doing so, a video intelligence arrangement can empower end-users to speed up investigations and examinations via looking and sifting videos based on specific criteria. The intelligent video analysis provides a situational analysis capability by considering principles-based alerts used on videos and images. Moreover, it allows inferring operational knowledge by envisioning the video information into dashboards, graphs, identifying patterns, and mathematically analyzing results. Deep learning has been applied to numerous research areas such as time series forecasting, image segmentation, video analysis, image classification, etc. Video analytics has brought out many real-time applications. Video analytics usually involves object detection, object identification, face detection, face recognition, and gesture recognition. In this research work, we performed the recognizing of hand gestures using deep learning algorithms and then analyzed the results. A gesture is characterized as the actual motion of the hands, fingers, arms, and different parts of the human body along which a human can

pass on significance for cooperation with one another. There are two unique methodologies for human gesture recognition, the information gloves approach (Lu et al., 2016) and the vision-based methodology. The gloves approach comes with the obligation of using an apparatus with masses of cables, although it still presents good accuracy and speed of gesture recognition. The vision-based methodology was researched in the accompanying investigations including, the recognition and order of hand signals. This method can be considered the most practical option as this avoids the usage of additional equipment. It is critical for the infrastructure of any gesture recognition gadget to be practically applied to real-life scenarios.

In this work, we have developed a vision-based recognition system using deep learning to recognize the most common hand gestures used in TEDx talks and the number of times the gesture was used in the video. Thus, in this methodology, we show the application of convolutional neural networks (CNN) to TEDx video analytics. CNN models have proved to be the best performing networks not only in gesture recognition but also in object detection, activity recognition, and localization (Kopuklu et al., 2018) (3). Many different algorithms and methods have been used for classification and detection of hand, such as in Nguyen et al. (2015) Principal Component Analysis (PCA) is used to select attributes, and neural networks are used for classification. In Simonyan and Zisserman (2014) and Karpathy et al. (2014), many video

* Corresponding author.

E-mail address: siddhaling@dubai.bits-pilani.ac.in (S. Urolagin).

frames are used as inputs to 2D CNNs. Many other methods are used, which will be discussed in the literature sections. In our work, image datasets such as TwentyBN were considered. Furthermore, several images were added to the image dataset by capturing frames of TEDx videos using the OpenCV library in Python. To boost the accuracy, we used the data augmentation technique to increase the number of images to be used for training deep learning algorithms. For data augmentation, we have used Keras "ImageDataGenerator" to generate several new images using the captured images at different angles, rotations, pixels, zoom range, flips, etc. Here the parameters for flipping, rotations, and others were kept in a small range to avoid extra loss. For our proposed approach, we used 3D CNN and ResNeXt-101 (Hara et al., 2018). In this paper, we have created a system that uses these approaches to work on a custom-created dataset with newly added gestures performed using both hands. Gestures performed using two hands have been challenging to recognize.

With our proposed approach and workflow, we have enabled our model to simultaneously identify gestures with two different hands and identify gestures performed most commonly in TEDx talks. The paper focuses more on identifying hand gestures in TEDx videos that can help TEDx talkers improve their talks. The hand gesture performed during these talks has been given more importance as they contribute to communicating speakers ideas and effectively convincing the audience. Thereby suitable gestures have a significant role in making TEDx videos popular among the audience. Hand gestures serve as a neurological link between one's thoughts and their activity. These gestures help us visualize the words and also help in increasing meaning. While speaking, making hand gestures helps others remember what you say and helps you speak more quickly and efficiently. These gestures also help gain the listener's attention. They are a visual aid that is unmatched in their usefulness. The most popular TED Talkers used both their words and their hands to communicate. In this study, we analyze TEDx talks using our architecture to identify hand gestures and find how these hand gestures have helped TEDx talkers succeed. In the following sections, we covered the literature survey and works done in gesture recognition. The proposed methodology is given in section III, which elaborates on training and other processes. Sections IV and V elaborate on the experimental setup, results, and statistical analysis, and finally, the conclusion is covered in section VI.

Literature review

A massive amount of user-generated data in social media platforms are being used for many applications. Researchers in different areas are trying to utilize these data and build a solid big data theory model to solve the multiple issues and deal with various data (Kar & Dwivedi, 2020). Twitter has attracted many users to share their opinions and provided a centralized space to collect different global emotional data. These data have been used in a specific domain to solve real issues such as sentiment analysis and classification (Neogi et al., 2021) and handling the misinformation (Aswani et al., 2019) and personality dimensions (Lakhiwal & Kar, 2016).

Hand gesture recognition is a propitious subject, and it is applied to several sensible implementations and practical works (Kojima et al., 2000). A sample hand gesture is ascertained and acknowledged by police work and securities to forestall criminal behavior (Cohen et al., 2008). Hand gesture recognition has been used to build several applications such as sign language recognition (Mitra & Acharya, 2007) and build lie detection applications (Bond et al., 1990). For an image-based human hand gesture recognition system, since the quantity of variables of a picture area is wide, it is crucial to extract the essential options of the image. Recently, with the progress of CNN, abundant analysis concerning document images has found success in Simard et al. (2003). The CNN in image classification has fortunately resolved many issues (Krizhevsky et al., 2012). Using the application of NLP and CNNLSTM architecture sing language translator is developed in Agrawal and Uro-

lagin (2020). Nasir et al. (2021), introduce a novel hybrid deep learning model that blends CNN and RNNs together for false news classification. The model was effectively verified, yielding detection results that were much superior to non-hybrid approaches. Moreover, 3D-CNN extracted features in the field of CNN by classifying pictures and combining them with temporal options. The research work has been projected the combination of exploring 3D-CNN and RNN connected (Molchanov et al., 2016).

A strategy that utilizes the skin color model and AdaBoost classification for hand gestures recognition have been proposed (Sun et al., 2018). A study that orders computers using static and dynamic hand gestures with three primary advances are presented in Plouffe and Cretu (2015). Researchers have developed RGB-D static gesture recognition using fine turned Inception V3 model in Xie et al. (2018). In Liu et al. (2015), a profound CNN model is utilized to handle profundity assessment from single monocular picture issues. It likewise means to investigate the limit of profound CNN and persistent Restrictive Random Field (CRF). The proposed plot learns the unary and pairwise possibilities of nonstop CRF. Additionally, a model dependent on convolutional networks and a novel super-pixel pooling technique is proposed, which is around multiple times quicker. This proficient model is a superior performing CNN design. The authors have used a deep residual network for image recognition in He et al. (2016). Temporal Segment Network (TSM) methodology is used in Wang et al. (2016), where this divides the video into several small parts and uses 2D CNNs for action recognition. Tran et al. (2017) used 3D CNNs and 3D pooling to captures features in different dimensions. The 3D CNNs, instead of dividing videos into different frames, takes in a sequence of the frame. Different new architectures such as Covnet (Tran et al., 2017), GTN (LeCun et al., 1998), Very Deep Convnet (Simonyan & Zisserman, 2014), ResC3D (Miao et al., 2017), FOANet (Narayana et al., 2018) and fused model (Lin et al., 2018) have already been utilized for detection and recognition applications. Ensemble algorithms are also broadly used (He et al., 2016; Huang et al., 2017; Szegedy et al., 2015) to coordinate data from various modalities and further develop execution, which prompts unsuitable training and deduction in time in practice. In Donahue et al. (2015), the authors have proposed highlighting features from CNN frames and applying LSTM for global temporal modeling. A comparable methodology is proposed in Molchanov et al. (2016) where a recurrent CNN for hand gesture recognition, which again used 3D CNN to extract features of the hand. Many other methods used Markov models (Vieriu et al., 2011) for real-time static gesture recognition. Improving the classification performance of the nearest neighbor classifier is given (Athitsos & Sclaroff, 2005). Using both the depth and intensity information given by time of light static hand gesture recognition presented in Oprisescu et al. (2012). Real-time hand gesture recognition using ResNext-101 is used in Köpüklü et al. (2019). Optimizing and fine-tuning of the deep network is carried out in Chauhan et al. (2021) for Covid-19 medical images. Hand gestures are applied for human-computer interaction in 7th International Conference on Advances in Computing and Communications (2017). Based on multi-features fusion and template matching, gesture recognition is carried out in 2012 International Workshop on Information and Electronics Engineering. Lu (2019) carried out the analysis on techniques to improve the viewers' engagement and communication efficiency in the live stream. Several databases of videos containing videos on demand, long video and live content are analyzed to understand the quality of the videos on user engagement (Dobrian et al., 2011). Viewer engagement is studied on videos related to the gift-giving behavior in Yu et al. (2018) and found that engagement is positive with such videos. The researchers in Chen et al. (2015) present a new model using computer networks for predicting individual viewer involvement. They used self-deployed routers on the university campus to collect engagement information from online video traffic. YouTube data of 6 million videos spread over 25 thousand channels are used in Hoiles et al. (2017) and studies social dynamics and viewers engagement. Garg et al. (2021) use an Artificial Intelligence-based method, i-Pulse, that can review and evaluate thou-

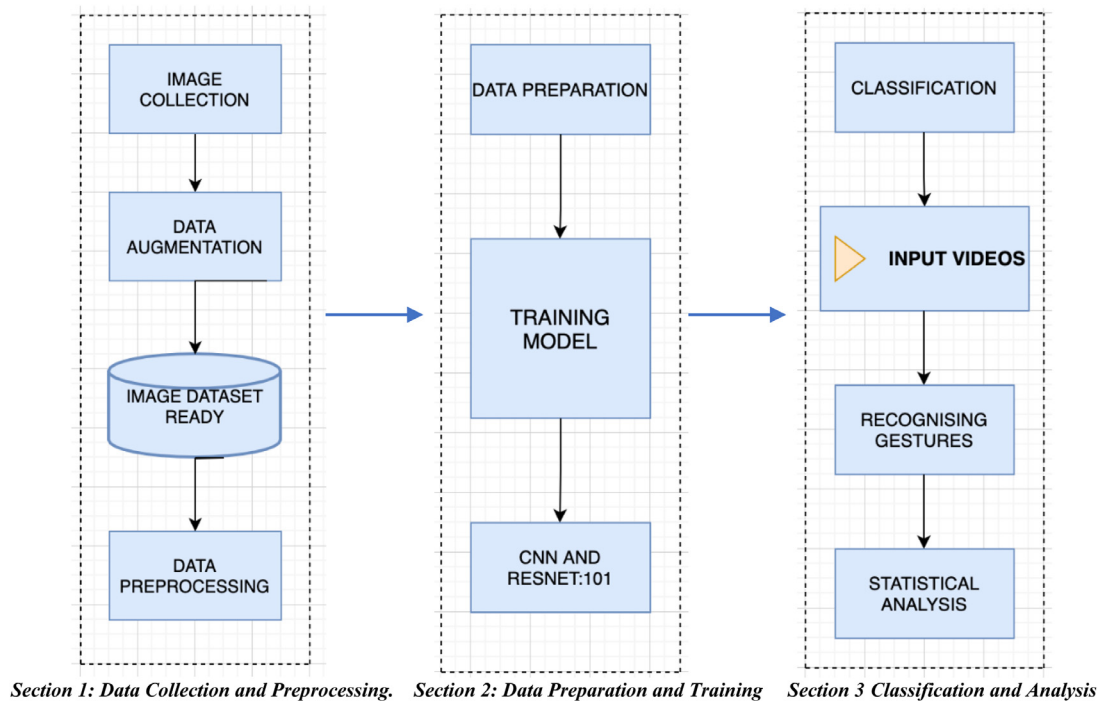


Fig. 1. Overall workflow.

sands of surveys. This can help large organizations to gain insights to improve their work culture. The (Rautaray & Agrawal, 2015) research aims to gain a deeper understanding of how streaming can be improved using HCI techniques. In Tran et al. (2015), techniques are presented for learning spatiotemporal features using 3D CNN. The main difficulty in egocentric vision is due to global camera movement, which is addressed in Cao et al. (2017) using recurrent 3D CNN. Doubly deep Long-term Recurrent Convolutional Networks are presented in Donahue et al. (2015), and they have the capability to extract space and time information. In Wang et al. (2016), researchers extracted feature point matches between frames to represent motion videos. A two-stream inflated 3D CNN was used to learn Spatio-temporal features on the Kinetic Human Action video dataset in Carreira and Zisserman (2017). The gesture recognition in a complex environment is carried out in Liang et al. (2015) using the multi-feature fusion technique.

Proposed method

This section explains our two architectures for real-time hand gesture recognition in TEDx videos using CNN and ResNeXt-101. Further, we explain the basic hierarchical architecture for detection in videos. With the accessibility of enormous datasets, CNN-based models have demonstrated their capacity in hand gesture recognition tasks. The 3D CNNs designs particularly stand out for video examination since they utilize worldly relations between outlines. Fig. 1 shows the overall workflow of the model. At first, we collect images from different datasets and TEDx videos. Moreover, we are then using the data augmentation technique to increase the dataset size, which helps increase the accuracy of the model. Pandas(1.3.1) library is used for pre-processing of images. We have used ResNeXt-101 for training our model. We need to prepare data according to the predefined methods. Here ResNeXt-101 takes in a target size of 64 by 96 size frames from videos, and we also predefine the number of frames per second to be 15. A skip value of one second for the trained model to work on a single frame. The Stochastic Gradient Descent (SGD) is an optimization algorithm when training the model using ResNeXt-101. Finally, the classification step is performed to clas-

sify gesture output. Further, we explain the entire functioning of the workflow in more detail.

The processing steps followed during real-time hand gesture recognition in our architecture is shown in Fig. 2. Since the algorithm does not know when the gesture is being performed in each video, we used frame-by-frame passing of images using a video to get the most efficient output. Each frame is passed into the detection algorithm, which further sends the frame to the classifier to classify the hand gesture performed. The primary purpose of the detection algorithm is to identify whether a gesture is performed or not and look for a similar gesture using the trained model. If no gesture is performed, the frame is passed, and no output is presented. Then the next frame is given as input to the detection algorithm. The overall performance of this model highly depends upon the detection algorithm. Hence this detection algorithm must be very accurate. That is the crucial reason for giving every frame as input to the algorithm. The detector is also trained to reduce false positives values as much as possible. If a gesture is recognized, a number is assigned to the frame passed to the classifier. The classifier has been predefined with gestures and their corresponding numbers. As the classifier receives the frame with a particular number, it then classifies the frame into a gesture and moves to the next frame. For gesture recognition, ResNeXt-101 (Hara et al., 2018) architecture is used. The authors of the paper had reformulated the entire network to learn residual functions concerning the layer inputs. Residual networks tackle the problem of gradient becoming infinitely small. Whenever a neural network has many layers, the gradient is pushed back to the previous layer—eventually bringing the gradient to a minimal value that further affects the network's performance. However, with the introduction of residual networks, this problem has been tackled efficiently. The counting function is an additional part of our algorithm that helps us count the number of hand gestures performed in a particular video. We used Pandas Library (1.3.1). We used the Series and value counts function in the library that took in output from the classifier and added up all of them separately. In the end, it printed the result in column format where the number of hand gestures is mentioned against the name of the gesture. In Fig. 2, the dotted red frame is currently in the detection algorithm, while the dotted yellow frame is a waiting queue and is the next frame to go inside the detection

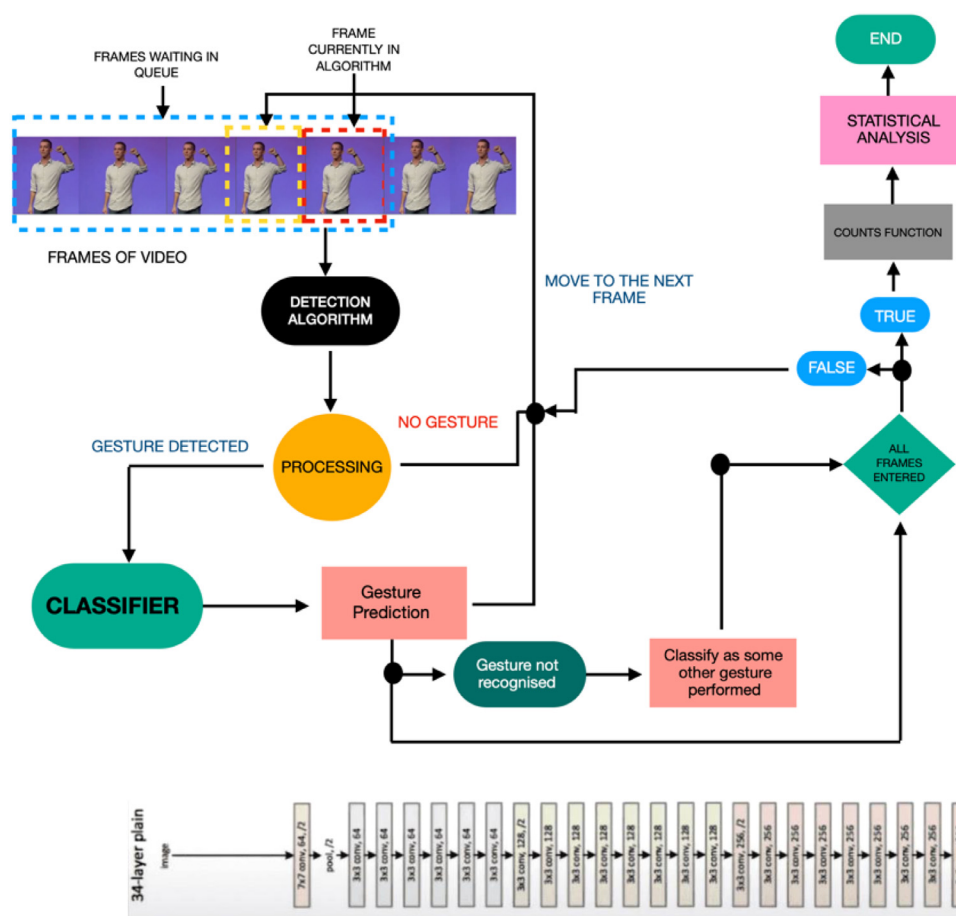


Fig. 2. General architecture of workflow.

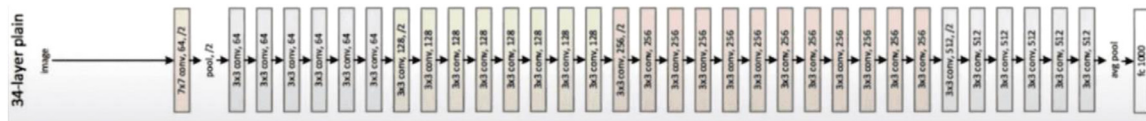


Fig. 3. 34-layer plain network.

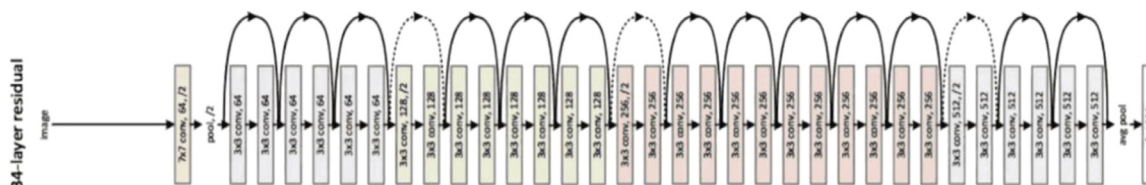


Fig. 4. 34-layer residual.

algorithm. Dotted blue frames show the total frames left. The classifier stays activated according to the number of frames in the blue dotted lines.

ResNet architecture is built using the concept of residual blocks. Suppose a residual block starts with some activation. The first step is to apply the linear operator by multiplying the weight metrics and adding the bias to the initial activation. Then the activation passes through the ReLU layer. ReLU gives the rectified feature map of an image. Since Resnets involve multilayer networks, we pass the activation again through the second layer of a linear operator and similarly again through the ReLU layer. Whereas in Resnets, we skip the first and the second step. We skip connections to avoid gradient loss. Here a skip connection is established. With the introduction to skip connection, we can avoid plain networks and instead use Resnet as these have given minor training error compared to theoretical error compared to other networks (Hara et al., 2018). Figs. 3 and 4 show the comparison between plain networks and residual networks. In Hara et al. (2018), the working of these networks and even a 1000 layered network can be trained with minor training errors. We use this architecture and the ResNeXt-101 defined in Köpckli et al. (2019) as our base for the training on our dataset.

Experimental setup

The results of the experiments were obtained using real-time videos given as input. First, the results were tested using our self-created image dataset. The dataset was built by collecting videos from YouTube and different TEDx talks and capturing frames of hand gestures to be recognized using OpenCV (4.5.3). The images were captured in 400×400 pixel resolution as shown in Fig. 5. Our final image dataset contained 17 hand gestures and a total of 130000 sample images. Many skin tones and a variety of shapes helped us build a general model.

Experimental results

The accuracy and loss for the proposed model during training and testing are shown in Fig. 6. The proposed architecture produces the results on test data which are shown in Table 1. An accuracy of 94.35% and an F1-score of 94.23% are obtained. The recall is 93.73%, while precision is 97.83% on the proposed model. The training and testing of our model are carried on a custom-built dataset. Our dataset included some new gestures that were added and have shown less accuracy than



Fig. 5. Sample collected gesture images.

Table 1
Detection results.

Model	Recall	Precision	Accuracy	F1-score
Resnet	93.73	97.83	94.35	94.23

more common gestures but still have an overall high accuracy value. After performing the experiment and observing the results, it was evident that ResNeXt-101 is one of the best performing CNN.

Fig. 7 shows us some of the live results recorded while testing our model. TEDx videos were given as input, and gestures were recognized live by the model very accurately. However, the model is not perfect but has performed well. With the recognition of live gestures, we also counted the number of gestures performed in every video given as input, and the detected gesture for every frame is saved. The frame with the highest accuracy of a particular gesture performed within the time limit is printed. If the same gesture is performed after a certain amount of time, it is then saved as another gesture and added to the count of that gesture. Hence, we were able to count the number of times a gesture is performed. With the help of the count function, we were able to detect the most common gestures performed in the TEDx videos.

Figs. 8 and 9 show us the most common gestures performed in our testing videos. We tested around 25 videos in total using our model—results and analysis of 4 such videos in the Fig. 8(a), (b), (c), and (d). The count function helped us analyze the results. As shown, the

most common gesture used was handed apart. This gesture has proven to improve the TEDx talks experience by making them more interactive and help communicate the idea or the story in a better way. Other gestures were also used in videos. Some other gesture is recognized as a category where the gesture performed cannot be classified as one of the gestures in the training set and hence is classified as some other gesture to avoid confusion and reduce false positives. In the future, we can improve the accuracy and performance using different detection methods, adding new gestures to the dataset, and finding new ways to identify and locate these gestures in images. Figure showing the various hand gesture made by TEDx speakers during his/her talk.

The above results show an analysis of some videos which were used for testing purposes. Now we analyze how performing more hand gestures has led to better engagement with the audience. Using our architecture, we were able to derive a relation between hand gestures and successful TEDx talks. In an experimental setup, we mentioned we gave the most famous TEDx talks as inputs. Fig. 9 shows the output of this relation. We noticed that we could recognize around 60 hand gestures in a 5-minute-long video with our architecture. Now we compared this result with the less popular TEDx videos. The less popular TEDx talks had an average number of 26 gestures in a minute video. This led us to conclude that meaningful hand gestures are to engage a more significant number of audiences. We saw that famous TEDx talkers used almost twice the number of hand gestures than in the other TEDx talks.

Tables 2 and 3 here shows the count function results represented in the form of a table. This table is showing the results of the model on ten videos. The testing videos were selected carefully as TEDx videos involve various camera angles, leading to false positives and bad results. Hence the videos selected for testing were first selected. The video duration was also reduced to the point where the camera angle in the video is suited for the model. Fig. 10 depicts the effect of performing hand gestures on views. Fig. 10(a) graph shows that as the number of hand gestures increases, so does the number of views. The first graph depicts the points of view in exponential terms: 0.25 equals 0.25×10^7 views. The Fig. 10(b) graph shows the same comparison, but with less popular videos and fewer views. We can see that the hand gestures performed are twice as few as those seen in popular videos. Even in less popular videos, we can see that increasing the number of hand gestures increases views. This demonstrates how effective hand gestures are in TEDx talks. The following two tables show how many gestures are performed in the most popular and least popular videos and how many different gestures are performed.

Fig. 11 compares the different hand gestures used in Popular Vs. Less Popular videos. The figure indicates how important using hand gestures is. Hand gestures are frequently used in Popular videos to highlight specific points of speech and solidify the speaker's message.

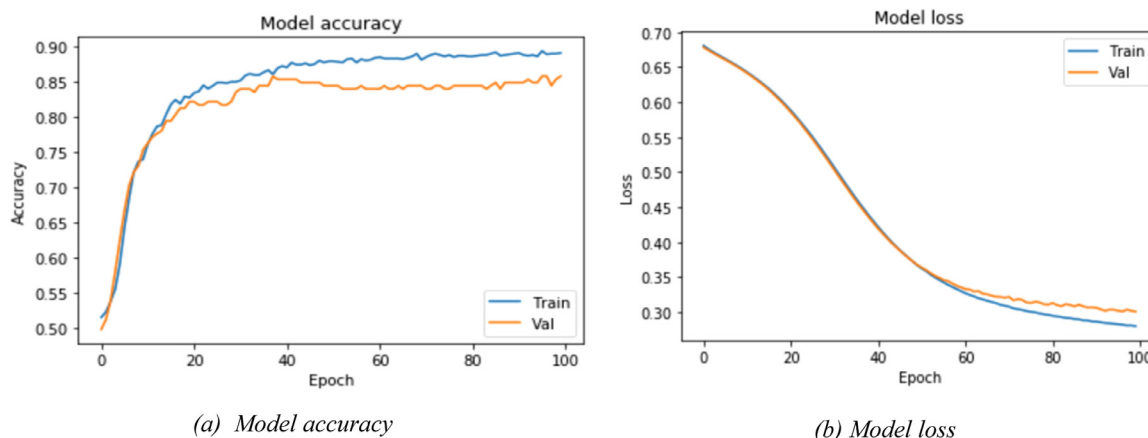
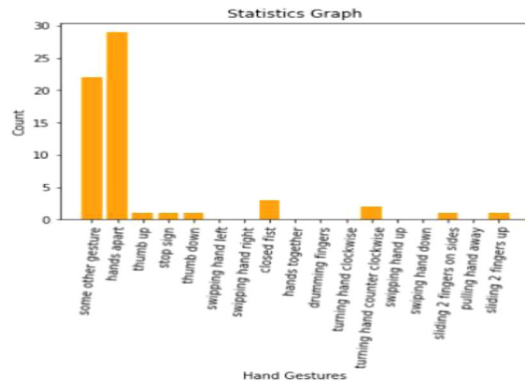


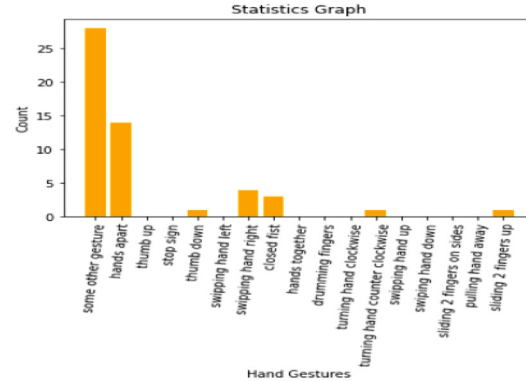
Fig. 6. Model accuracy and loss.



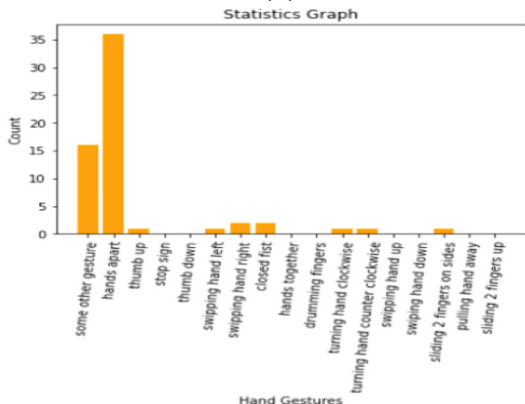
Fig. 7. Live results.



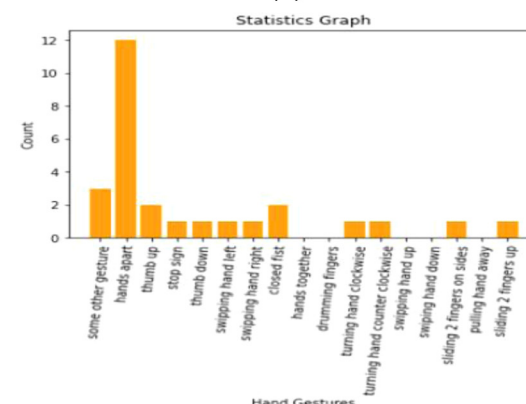
8 (a)



8(b)



8 (c)



8 (d)

Fig. 8. Analysis of TEDx video.

Table 2

Count function results for popular videos in %.

Video Number	Hands Apart	Stop Sign	Clenched hand	Hands Together	Pulling hand away
1	40.2	4.16	8.33	2.77	2.77
2	27.6	6.15	4.61	6.15	6.15
3	38.8	1	2.98	2.98	2.98
4	32.3	7.6	6.15	7.6	4.61
5	34.9	1.5	14.2	11.11	3.17
6	24	7	3	4	3
7	34.7	7.2	11.59	13.04	1.4
8	44.6	6.3	6.3	6.3	4.2
9	42.8	0	3.8	3.8	0
10	37.9	8.6	1.7	8.6	6.8
TOTAL	35.6	5.1	6.4	6.7	3.5

At the same time, analyzing the results, we also realized that people tend to like TEDx talks more based on a speaker's body language and gestures used more than their actual words. We also found that Temple Grandin, Simon Sinek, and Jane McGonigal top the hand gesture charts.

Fig. 12 shows us the most common hand gestures recognized using our architecture in the most popular and viewed TEDx talks. We see that hands apart have been the most frequently used hand gesture. This char-

acteristic gesture where hands are held some inches apart with palms facing each other has been a hallmark move for these talkers. When discussing a topic with much depth, great presenters tend to do this. Because the palms usually are in an open stance, it is commanding but candid. We noticed that when discussing an important topic, this was done. It psychologically prepares the listeners to know the importance of what is being expressed in the talk. Hence hands apart are found to be the most crucial gesture, which should be more commonly used

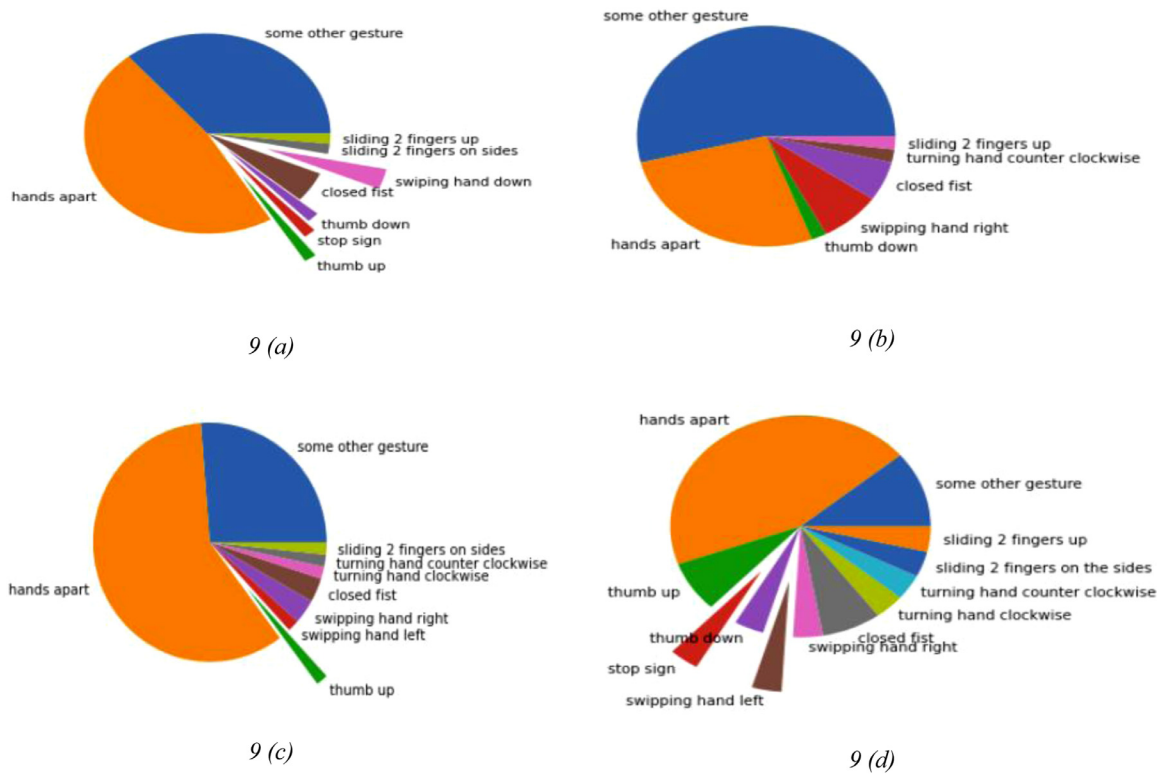


Fig. 9. Commonly performed gestures in tested videos.

Table 3

Count function results for less popular videos in %.

Video Number	Hands Apart	Stop Sign	Clenched hand	Hands Together	Pulling hand away
1	21.05	5.2	0	10.5	10.5
2	16.6	5.5	0	5.5	5.5
3	30	5	1	5	5.1
4	9	4.54	13.6	9.09	4.54
5	5.88	5.88	5.88	17.64	11.96
6	15	0	5	20	0
7	33.3	12.5	4.16	4.16	4.16
8	20	15	15	10	0
9	23.7	0	7.69	7.69	15.38
10	23.6	0	5.88	5.88	0
TOTAL	19.89	11	6.2	8.9	4.7

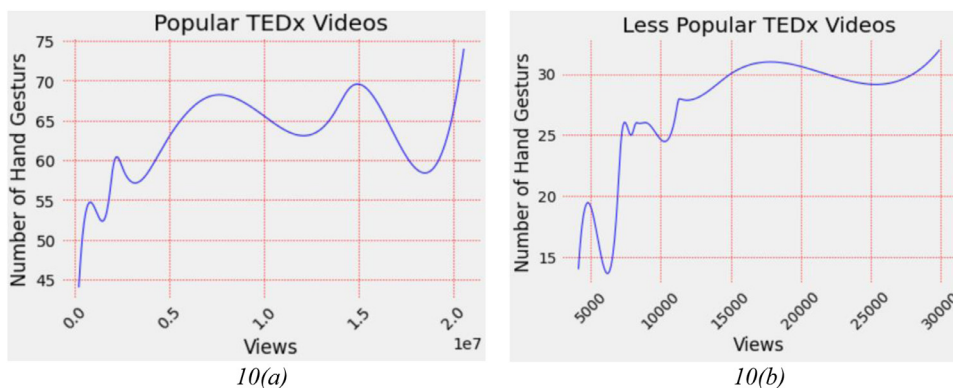


Fig. 10. Number of gestures and views (Most Popular Vs Less Popular TEDx Talks).

to engage with the audience and have a more impactful talk. The next most used hand gesture is hands together. The second most common gesture found using our study is combining hands, implying two separating forces coming to join and achieving a connection. The stop sign was also often seen used during experiments. Stop sign indicates to slow down or stop when someone's actions.

While we see turning hand clockwise and counterclockwise was also quite often used. This gesture has an entirely different meaning as the talker is probing some ideas. This is very similar to the hands apart gesture, while this was done when the TEDx talker usually asked a question like What? How? Swiping hands-on left, right, and in other directions was a grandiose gesture. It seemed as if one is sweeping through all the

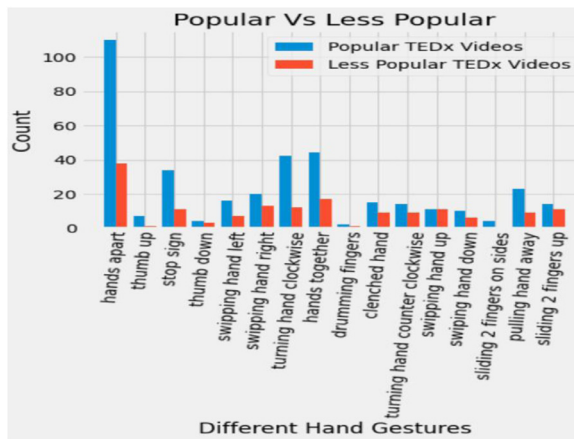


Fig. 11. Count function results for less popular videos.

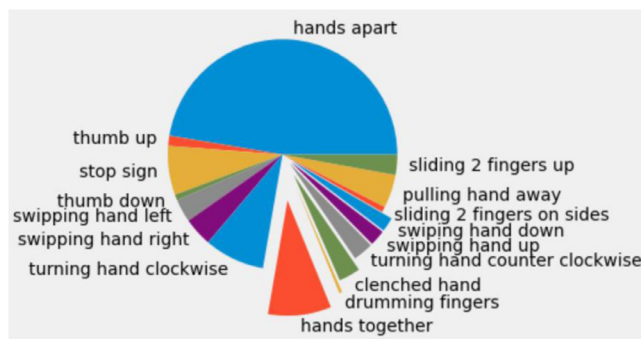


Fig. 12. Most common hand gestures in TEDx videos.

ideas with the ability to be supportive. This expression can also mean "cleaning the slate" or "moving something far away." Thumbs up and down were also used many times. Thumbs up and down imply confidence and cooperation. While it is also used for expressing gratitude for a favor received, asking authorization to perform a task. While thumbs down is usually used to deny something. A clenched was also often used. TEDx talkers exhibited intensity by making a solid fist, shaking it at someone, or punching it in the air. It was primarily used in conjunction with a crucial point. Although they were cautious when utilizing this motion as sometimes this might come across as angry. This shows how interrelated successful TEDx talks and hand gestures are. An energetic speaker makes a more substantial impact on the viewers rather than the passive one. Without gestures, public speaking will always be considered boring and stale, no matter how great the speech is. Also, talking about engagement with audience and viewers online, we went through and used Selenium (3.141.0) and BeautifulSoup (4.9.3) to scrape comments on Popular TEDx talks and classified them using text mining. We classified comments as positive, neutral, or negative. We discovered that popular TEDx talks with more hand gestures received more positive and neutral feedback. It may be said that having more views leads to more comments, both positive and negative. It can also be said that feedback/views of the audience depend mostly on content and topic of speech. Yes, the positive response mostly depends on the topic, but we also completely believe that gestures play an essential role in making the talk more stirring. Hence, a greater audience would be able to connect themselves with the topic. The content of Talk is what attracts people, but the narration might or might not receive good feedback. It entirely depends upon how the speaker puts his/her narration to the audience. Using our study, however, we have already discovered a link between views and hand motions. As a result, we deduce that employing more hand gestures can lead to more successful TEDx lectures with more views and more positive comments and thoughts on the recordings.

Fig. 13(a) illustrates how having more views has resulted in more positive feedback. As a result, we were intrigued to learn more about the link between hand gestures and positive comments and emotions. Fig. 13(b) depicts the direct relationship between more gestures and the number of positive comments received. We used the classified outputs from our classifier algorithm and compared them with the number of hand gestures used. It showed us a direct and conclusive correlation.

Discussion

Computer vision has accelerated the work in many different areas. Medical domain data has vastly been used by the deep learning approaches (Chen et al., 2015). Recent advancements in the deep learning approached provided a more effortless and faster approach to deal with video data.

Contribution to the theory

Artificial intelligence has provided new research and approach to solve real-time problems. A detailed algorithms review has been done on swarm intelligence which gives a new direction to suggest a better algorithm than the existing one (Chakraborty & Kar, 2017). Textual information has given more opportunities to build a theory for benefiting the end-users, such as sentiment analysis for recommending the people to new hotels and tourist spots (Kar, 2020) (Mishra et al., 2020). Based on user-generated content on mobile payment systems, the satisfaction model has been proposed to enhance digital services (Mishra et al., 2019). Many of organizations are facing a big issue of providing permanent storage of user-generated data, and in such cases, a flexible pricing model is required to use the cloud services (Kar & Rakshit, 2015). The cloud services also provide the host with real-time applications and applying modified machine learning methods such as spam detection in email using proper optimized techniques (Batra et al., 2021).

Gesture recognition from video data has become an essential task in the computer vision domain. The main objective of gesture recognition is to identify the action performed by a person using his hand movements. These hand movements convey subtle information, especially during a presentation. In our work, we focused on TEDx video analysis. The TEDx videos are collected to identify the gesture of a person during presentation time. The deep learning model Resnet is developed by taking the pre-trained model and then training it on 130000 frame images of TEDx videos. The Resnet model has given an accuracy of 94.35% on the test set. Our model can identify 17 different gestures. The Resnet profound learning model results are used to recognize different gestures performed by experts during the TEDx presentations.

In Table 4 comparison of the most recent works on gesture recognition are summarized. Köpüklü et al. (2019) gave outstanding results using ResNet and Resnet-X as a classifier. They tested their work on publicly available datasets such as EgoGesture Dataset and NVIDIA Dynamic Gesture Dataset. We used a similar approach but in more complex environments and using custom datasets. The above table also shows other studies that have been carried out. While looking at the results, we can be assured that Resnet has outperformed all other algorithms.

We see that our approach had the highest accuracy. The accuracy is heavily reliant on the features of the datasets used and particularly the duration, quality, and angle of each video. While creating the dataset, we considered all these important parameters. Also, one of the key factors that contributed to gaining a higher accuracy is the way our algorithm is working. We found that a frame-by-frame input offered a superior outcome in real-time evaluation because the system could detect the gesture more accurately. This helped the algorithm to detect the start and end of the gesture more robustly. As a result of the dataset and the addition of the sliding window concept for frame-by-frame detection, we were able to achieve a greater level of accuracy.

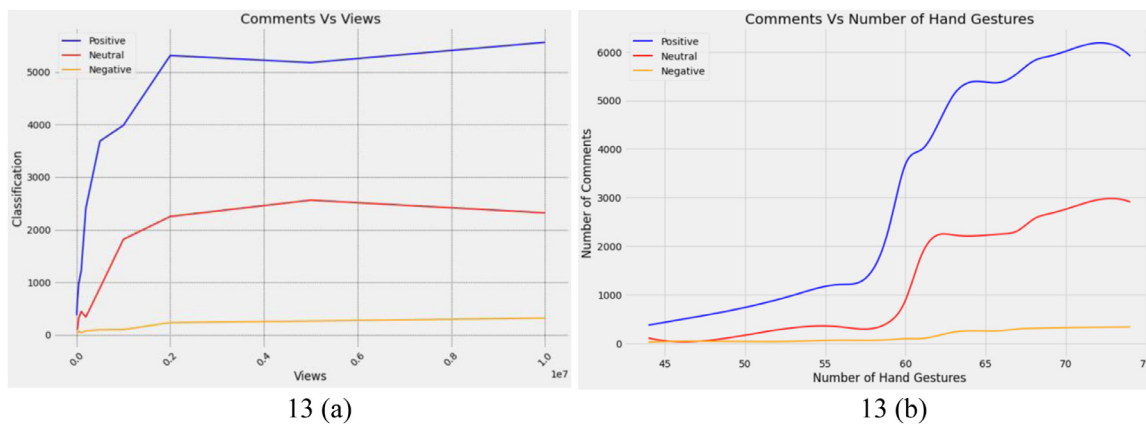


Fig. 13. (a) Comments vs. Views, (b) Comments Vs. Number of Hand Gestures.

Table 4
Comparison of results.

Refs.	Classify Algorithm	Network Used	Number of Gestures/ Dataset	Accuracy
(Rautaray & Agrawal, 2015)	CNN	VGG16	Ego Gesture	66.5%
(Tran et al., 2015)	C3D	C3D	Ego Gesture	89.7%
(Cao et al., 2017)	C3D	C3D+LSTM+RSTM	Ego Gesture	92.2%
(Donahue et al., 2015)	CNN	VGG16 + LSTM	Ego Gesture	81.4%
(Wang et al., 2016)	CNN	MTUF	Multiple Datasets	93.87%
(Carreira & Zisserman, 2017)	CNN	I3DF	Ego Gesture	92.78%
(Köpküklü et al., 2019)	3DCNN	ResNet and ResNeXt	NVIDIA Dynamic Gesture Dataset	94.03%
Ours	3DCNN	ResNeXt-101	Custom Dataset	94.35%

Implication to practice

In gesture recognition, human gestures are recognized using computational techniques such as machine learning and deep learning. The video analytics using gesture recognition reveals the most crucial information about gesture movements. These gesture movements can help to convey ideas and information to the viewer. TEDx is one of the major platforms for sharing information across audiences by experts in that field. The gesture analysis of TEDx videos has two crucial implications. One the gesture analytics will help to understand and study how experts are using gestures to convey information. Secondly, the results from gesture analysis can be utilized by the presentation to improve his presentation skills. In this study, the statistical analysis of gestures is carried out on 25 TEDx videos, and various visualizations are given.

Conclusion and future scope

Hands gesture recognition system addresses various problems and has been an important topic in the computer vision field. Hand gesture has many applications in medical, industries, assistance to humans, virtual reality, crisis management, etc. Many algorithms and methods have been implemented and tested for gesture recognition. In our work, we have used Resnet-101 to build a gesture recognition system. We have trained the model on the customized data set created from TEDx videos. The dataset consists of 130000 samples, and the model is trained on 17 different gestures. A gesture recognition accuracy of 94.35% is observed on the test set. The system analyzes each frame at a time. Whenever a gesture is performed, it is identified and sent for further processing. During the next phase, the classification of the gesture takes place. The statistical analysis is then carried out on the video, and various gesture frequencies are collected. Our count function method helped us analyze the outputs of these videos. Different visualizations are prepared to represent the statistics collected on the TEDx videos graphically.

We discovered that expressing something was also essential to get impact and attract a larger audience.

While analyzing our results, we found that the greatest way to describe oneself or a topic is to use your hands as your narrative tool. This shows that TEDx talkers should pay attention to their nonverbal communication as much as their spoken communication. They should allow their hand to speak for themselves, which adds more value to a TEDx talk. As a result, we deduce that the topic/content is what draws the audience in, but how the speakers convey stories using gestures is what helps them succeed. Our model successfully detected hand gestures and helped us perform an analysis that future TEDx talkers can use to improve their speeches. The models we built were very accurate although, some extensions can still be done in complex environments. For example, if two people perform a gesture in the same frame, like two or more people present on the stage, which might confuse the model and lead to less accuracy. Therefore, the recognition of gestures in more complex environments can still be extended. Hence, this research can be further improved by creating more complex and congested environments and training the model using new techniques and architectures.

Declaration of Competing Interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

References

- 2012 International Workshop on Information and Electronics Engineering (IWIEE) A Hand Gesture Recognition Method Based on Multi-Feature Fusion and Template Matching Liu Yun, Zhang Lifeng, Zhang Shujun
- 7th International Conference on Advances in Computing & Communications, ICACC-2017, 22- 24 August (2017), Cochin, India Hand Gesture Recognition for Human Computer Interaction Aashni Hariia, Archanasri Subramaniana, Nivedhitha Asokkumara, Shristi Poddara, Jyothi S Nayaka
- Agrawal, T., & Urolagin, S. (2020). 2-way Arabic Sign Language translator using CNNLSTM architecture and NLP. In *ACM, Ei Compendex and Scopus, ISI web of science, international conference on big data engineering and technology BDET-20* (pp. 96–101).
- Aswani, R., Kar, A. K., & Ilavarasan, P. V. (2019). Experience: Managing misinformation in social media—insights for policymakers from Twitter analytics. *Journal of Data and Information Quality (JDIQ)*, 12(1), 1–18.

- Athitsos, V., & Sclaroff, S. (2005). Boosting nearest neighbor classifiers for multiclass recognition. *2005 IEEE computer society conference on computer vision and pattern recognition (CVPR'05)-workshops* 45-45.
- Batra, J., Jain, R., Tikkwal, V. A., & Chakraborty, A. (2021). A comprehensive study of spam detection in e-mails using bio-inspired optimization techniques. *International Journal of Information Management Data Insights*, 1(1), Article 100006.
- Bond, C. F., Omar, A., Mahmoud, A., & Bonser, R. N. (1990). Lie detection across cultures. *Journal of nonverbal behavior*, 14(3), 189-204.
- Cao, C., Zhang, Y., Wu, Y., Lu, H., & Cheng, J. (2017). Egocentric gesture recognition using recurrent 3d convolutional neural networks with spatiotemporal transformer modules. In *Proceedings of the IEEE international conference on computer vision* (pp. 3763-3771).
- Carreira, J., & Zisserman, A. (2017). Quo vadis, action recognition? a new model and the kinetics dataset. In *Proceedings of the IEEE conference on computer vision and pattern recognition* (pp. 6299-6308).
- Chakraborty, A., & Kar, A. K. (2017). Swarm intelligence: A review of algorithms. *Nature-Inspired Computing and Optimization*, 475-494.
- Chauhan, T., Palivela, H., & Tiwari, S. (2021). Optimization and Fine-Tuning of DenseNet model for classification of Covid-19 cases in Medical Imaging. *International Journal of Information Management Data Insights*, Article 100020.
- Chen, Y., Chen, Q., Zhang, F., Zhang, Q., Wu, K., Huang, R., & Zhou, L. (2015). Understanding viewer engagement of video service in wi-fi network. *Computer Networks*, 91, 101-116.
- Cohen, C. J., Morelli, F., & Scott, K. A. (2008). A surveillance system for the recognition of intent within individuals and crowds. In *2008 IEEE conference on technologies for homeland security* (pp. 559-565).
- Dobrian, F., Sekar, V., Awan, A., Stoica, I., Joseph, D., Ganjam, A., Zhan, J., & Zhang, H. (2011). Understanding the impact of video quality on user engagement. *ACM SIGCOMM Computer Communication Review*, 41(4), 362-373.
- Donahue, J., Anne Hendricks, L., Guadarrama, S., Rohrbach, M., Venugopalan, S., Saenko, K., & Darrell, T. (2015). Long-term recurrent convolutional networks for visual recognition and description. In *Proceedings of the IEEE conference on computer vision and pattern recognition* (pp. 2625-2634).
- Donahue, J., Anne Hendricks, L., Guadarrama, S., Rohrbach, M., Venugopalan, S., Saenko, K., & Darrell, T. (2015). Long-term recurrent convolutional networks for visual recognition and description. In *Proceedings of the IEEE conference on computer vision and pattern recognition* (pp. 2625-2634).
- Garg, R., Kivilekar, A. W., Netak, L. D., & Ghodake, A. (2021). i-Pulse: A NLP based novel approach for employee engagement in logistics organization. *International Journal of Information Management Data Insights*, 1(1), Article 100011.
- Hara, K., Kataoka, H., & Satoh, Y. (2018). Can spatiotemporal 3d cnns retrace the history of 2d cnns and imagenet? In *Proceedings of the IEEE conference on computer vision and pattern recognition* (pp. 6546-6555).
- He, K., Zhang, X., Ren, S., & Sun, J. (2016). Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition* (pp. 770-778).
- He, K., Zhang, X., Ren, S., & Sun, J. (2016). Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition* (pp. 770-778).
- Hoiles, William, Aprem, Anup, & Krishnamurthy, Vikram (2017). Engagement and popularity dynamics of youtube videos and sensitivity to meta-data. *IEEE Transactions On Knowledge And Data Engineering*, 29(7) July.
- Huang, G., Liu, Z., Van Der Maaten, L., & Weinberger, K. Q. (2017). Densely connected convolutional networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition* (pp. 4700-4708).
- Kar, A. K. (2020). What affects usage satisfaction in mobile payments? Modelling user generated content to develop the "digital service usage satisfaction model. *Information Systems Frontiers*, 1-21.
- Kar, A. K., & Dwivedi, Y. K. (2020). Theory building with big data-driven research—Moving away from the "What" towards the "Why. *International Journal of Information Management*, 54, Article 102205.
- Kar, A. K., & Rakshit, A. (2015). Flexible pricing models for cloud computing based on group decision making under consensus. *Global Journal of Flexible Systems Management*, 16(2), 191-204.
- Karpathy, A., Toderici, G., Shetty, S., Leung, T., Sukthar, R., & Fei-Fei, L. (2014). Large-scale video classification with convolutional neural networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition* (pp. 1725-1732).
- Kojima, A., Izumi, M., Tamura, T., & Fukunaga, K. (2000). Generating natural language description of human from video images. In *Proceedings of the IEEE conference on computer vision and pattern recognition* (pp. 728-731).
- Kopuklu, O., Kose, N., & Rigoll, G. (2018). Motion fused frames: Data level fusion strategy for hand gesture recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition workshops* (pp. 2103-2111).
- Köptüklü, Okan, Gunduz, Ahmet, Kose, Neslihan, & Rigoll, Gerhard (2019). Real-time Hand Gesture Detection and Classification Using Convolutional Neural Networks. *IEEE International Conference on Automatic Face and Gesture Recognition FG*.
- Krizhevsky, A., Sutskever, I., & Hinton, G. E. (2012). Imagenet classification with deep convolutional neural networks. *Advances in Neural Information Processing Systems*, 25, 1097-1105.
- Lakhiwal, A., & Kar, A. K. (2016). Insights from Twitter analytics: Modeling social media personality dimensions and impact of breakthrough events. In *Conference on e-Business, e-Services and e-Society* (pp. 533-544).
- LeCun, Y., Bottou, L., Bengio, Y., & Haffner, P. (1998). Gradient-based learning applied to document recognition. *Proceedings of the IEEE*, 86(11), 2278-2324.
- Liang, Wang, Gui-xi, Liu, & Hongyan, Du (2015). Dynamic and combined gestures recognition based on multi-feature fusion in a complex environment. *The Journal of China Universities of Posts and Telecommunications*.
- Lin, C., Wan, J., Liang, Y., & Li, S. Z. (2018). Large-scale isolated gesture recognition using a refined fused model based on masked res-c3d network and skeleton lstm. In *2018 13th IEEE international conference on automatic face & gesture recognition (FG 2018)* (pp. 52-58).
- Liu, F., Shen, C., Lin, G., & Reid, I. (2015). Learning depth from single monocular images using deep convolutional neural fields. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 38(10), 2024-2039.
- Lu, D., Yu, Y., & Liu, H. (2016). Gesture recognition using data glove: An extreme learning machine method. In *2016 IEEE international conference on robotics and biomimetics (ROBIO)* (pp. 1349-1354).
- Lu, Z. (2019). Improving viewer engagement and communication efficiency within non-entertainment live streaming. In *The adjunct publication of the 32nd annual ACM symposium on user interface software and technology* (pp. 162-165).
- Miao, Q., Li, Y., Ouyang, W., Ma, Z., Xu, X., Shi, W., & Cao, X. (2017). Multimodal gesture recognition based on the res3d network. In *Proceedings of the IEEE International Conference on Computer Vision Workshops* (pp. 3047-3055).
- Mishra, R. K., Urolagin, S., & Jothi, J. A. A. (2019). A Sentiment analysis-based hotel recommendation using TF-IDF Approach. In *2019 International Conference on Computational Intelligence and Knowledge Economy (ICCIKE)* (pp. 811-815). 10.1109/ICCIKE47802.2019.9004385.
- Mishra, R. K., Urolagin, S., & Jothi, J. A. A. (2020). Sentiment Analysis for POI Recommender Systems. In *2020 Seventh International Conference on Information Technology Trends (ITT)* (pp. 174-179). 10.1109/ITT51279.2020.9320885.
- Mitra, S., & Acharya, T. (2007). Gesture recognition: A survey. *IEEE Transactions on Systems, Man, and Cybernetics, Part C (Applications and Reviews)*, 37(3), 311-324.
- Molchanov, P., Yang, X., Gupta, S., Kim, K., Tyree, S., & Kautz, J. (2016). Online detection and classification of dynamic hand gestures with recurrent 3d convolutional neural network. In *Proceedings of the IEEE conference on computer vision and pattern recognition* (pp. 4207-4215).
- Narayana, P., Beveridge, R., & Draper, B. A. (2018). Gesture recognition: Focus on the hands. In *Proceedings of the IEEE conference on computer vision and pattern recognition* (pp. 5235-5244).
- Nasir, Jamal Abdul, Subhani Khan, Osama, & Varlamis, Iraklis (2021). Fake news detection: A hybrid CNN-RNN based deep learning approach. *International Journal of Information Management Data Insights*, 1(1), Article 100007 ISSN 2667-0968https://doi.org/10.1016/j.ijime.2020.100007.
- Neogi, Ashwin Sanjay, Garg, Kirti Anilkumar, Mishra, Ram Krishn, & Dwivedi, Yogesh K. (2021). Sentiment analysis and classification of Indian farmers' protest using twitter data. *International Journal of Information Management Data Insights*, 1(2), Article 100019 ISSN 2667-0968https://doi.org/10.1016/j.ijime.2021.100019.
- Nguyen, T.-N., Huynh, H.-H., & Meunier, J. (2015). Static hand gesture recognition using principal component analysis combined with the artificial neural network. *Journal of Automation and Control Engineering*, 3(1).
- Oprinescu, S., Rasche, C., & Su, B. (2012). Automatic static hand gesture recognition using ToF cameras. In *2012 Proceedings of the 20th European Signal Processing Conference (EUSIPCO)* (pp. 2748-2751).
- Plouffe, G., & Cretu, A. M. (2015). Static and dynamic hand gesture recognition in depth data using dynamic time warping. *IEEE Transactions on Instrumentation and Measurement*, 65(2), 305-316.
- Rautaray, S. S., & Agrawal, A. (2015). Vision based hand gesture recognition for human computer interaction: A survey. *Artificial Intelligence Review*, 43(1), 1-54.
- Simard, P. Y., Steinkraus, D., & Platt, J. C. (2003). Best practices for convolutional neural networks applied to visual document analysis. *Icdar*, 3(2003).
- Simonyan, K., & Zisserman, A. (2014). *Very deep convolutional networks for large-scale image recognition*. arXiv preprint arXiv:1409.1556.
- Simonyan, K., & Zisserman, A. (2014). *Two-stream convolutional networks for action recognition in videos*. arXiv preprint arXiv:1406.2199.
- Sun, J. H., Ji, T. T., Zhang, S. B., Yang, J. K., & Ji, G. R. (2018). Research on the hand gesture recognition based on deep learning. In *2018 12th International symposium on antennas, propagation and EM theory (ISAPE)* (pp. 1-4).
- Szegedy, C., Liu, W., Jia, Y., Sermanet, P., Reed, S., Anguelov, D., Erhan, D., Vanhoucke, V., & Rabinovich, A. (2015). Going deeper with convolutions. In *Proceedings of the IEEE conference on computer vision and pattern recognition* (pp. 1-9).
- Tran, D., Bourdev, L., Fergus, R., Torresani, L., & Paluri, M. (2015). Learning spatiotemporal features with 3d convolutional networks. In *Proceedings of the IEEE international conference on computer vision* (pp. 4489-4497).
- Tran, D., Ray, J., Shou, Z., Chang, S. F., & Paluri, M. (2017). *Convnet architecture search for spatiotemporal feature learning*. arXiv preprint arXiv:1708.05038.
- Vieriu, R. L., Goras, B., & Goras, L. (2011). On HMM static hand gesture recognition. In *ISSCS 2011-International symposium on signals, circuits and systems* (pp. 1-4).
- Wang, H., Oneata, D., Verbeek, J., & Schmid, C. (2016). A robust and efficient video representation for action recognition. *International Journal of Computer Vision*, 119(3), 219-238.
- Wang, L., Xiong, Y., Wang, Z., Qiao, Y., Lin, D., Tang, X., & Van Gool, L. (2016). Temporal segment networks: Towards good practices for deep action recognition. In *European conference on computer vision* (pp. 20-36).
- Xie, B., He, X., & Li, Y. (2018). RGB-D static gesture recognition based on convolutional neural network. *The Journal of Engineering*, 2018(16), 1515-1520.
- Yu, E., Jung, C., Kim, H., & Jung, J. (2018). Impact of viewer engagement on gift-giving in live video streaming. *Telematics and Informatics*, 35(5), 1450-1460.