Automated Engagement Recognition in E-Environments

[1]Ankita Mishra, [2]Uday Sai Savitha, [3]Jonnadula Narasimharao

[1,2]Student, Computer Science and Engineering, CMR Technical Campus, Medchal, Telangana, India.
[3]Associate Professor, Computer Science and Engineering, CMR Technical Campus, Medchal, Telangana, India.

Abstract: The gap between the actual and virtual worlds is closing at an incredible rate. Interaction with computers is becoming increasingly common as more people utilize them to complete a variety of jobs ranging from online learning to shopping. In such circumstances, identifying a user's level of involvement with the system with which he or she is engaging can modify how the system responds to the user. This will result in more engagement with the system as well as improved human-computer connection. In today's vision applications, including advertising, healthcare, autonomous vehicles, and e-learning, identifying user engagement might be critical. An automated engagement detection system that can analyze a person's engagement outcome with a certain object or an environment can be crucial to many organizations and businesses around the globe. Therefore, we employ cutting-edge algorithms in our project to recognize user engagement levels and divide them into two categories: positive and negative.

***Index Terms*- Computer Vision, SlowFast Network, Crime recognition, Crime Recognition, Action Recognition, Video Understanding, Neural Networks, Machine Learning, Convolutional Neural Networks**

1. INTRODUCTION

An automated engagement detection system that can analyze a person's engagement outcome with a certain object or an environment in an E-Environment. We built a computer vision model that takes input from a video, recognizes human emotions from face expressions and body behavior and categorizes them into either positive or negative.
This can be further developed to track and detect in real time and send the information to the backend for further use cases. The gap between the actual and virtual worlds is closing at an incredible rate. Interaction with computers is becoming increasingly common as more people utilize them to complete a variety of jobs ranging from online learning to shopping. In such circumstances, identifying a user's level of involvement with the system with which he or she is engaging can modify how the system responds to the user. This will result in more engagement with the system as well as improved human-computer connection. In today's vision applications, including advertising, healthcare, autonomous vehicles, and e-learning, identifying user engagement might be critical. We automate engagement level recognition for E-Environments using advanced computer vision techniques such as Slow Fast networks.
The main feature of this project is that the system will be capable of identifying the different states of emotions a user goes through in a E-setting and analyze it and categorize whether the response is either positive or negative without any human intervention.

2. MOTIVATION

Existing systems that are in place for recognizing human engagements are highly inefficient. The following are some ways in engagement is measured in an E-Setting.

1. Surveys

User Engagement can be measured by conducting surveys where users will fill in a survey form to give information regarding their engagement levels.

2. Manual Study from Videos:

User Engagement is measured by a person by going through multiple videos and identifying the affective states of person

There are numerous limitations such as

1. Inaccurate
2. Less credible
3. Herculean task for a human to do all the analyzing
4. Practically impossible for vast amounts of data

Therefore, we propose a solution where an automated engagement detection system that can analyze a person's engagement outcome with a certain object or an environment can be crucial to many organizations and businesses around the globe. Therefore, we employ cutting-edge algorithms in our project to recognize user engagement levels and divide them into two categories: positive and negative.

3. LITERATURE SURVEY

3.1 Artificial Neural Network:

Artificial neural network (ANN), usually called Neural Network (NN), is an algorithm that was originally motivated by the goal of having machines that can mimic the brain. A neural network consists of an interconnected group of artificial neurons. They are physical cellular systems capable of obtaining, storing information, and using experiential knowledge. Like the human brain, the ANN's knowledge comes from examples that they encounter. In the human neural system, the learning process includes the modifications to the synaptic connections between the neurons. In a similar way, ANNs adjust their structure based on output and input information that flows through the network during the learning phase. Data processing procedure in any typical neural network has two major steps: the learning and application step. As the first step, a training database or historical price data is needed to train the networks. This dataset includes an input vector and a known output vector. Each one of the inputs and outputs represents a node or neuron. In addition, there are one or more hidden layers. The objective of the learning phase is to adjust the weights of the 11 the resulting outputs will be compared with the known outputs. If the result and the unknown output are not equal, changing the weights of the connections will continue until the difference is minimized. After acquiring the desired convergence for the networks in the learning process, the validation dataset is applied to the network for the validating step.

3.2 Learning Paradigms in ANNs:

The ability to learn is a peculiar feature pertaining to intelligent systems, biological or otherwise. In artificial systems, learning (or training) is viewed as the process of updating the internal representation of the system in response to external stimuli so that it can perform a specific task. This Wavelet De- noising- based Back propagation (WDBP) neural network. For demonstrating superiority new model in predicting, the results of it is compared with Back Propagation neural network and the total results showed that the WDBP model for forecasting index is better than BP model. Putra and Kosala (2011) try to predict intraday trading Signals at Dixey used technical indicators - the Price Channel Indicator, the Adaptive Moving Averages, the Relative Strength Index, the Stochastic Oscillator, the Moving Average Convergence-Divergence, the Moving Averages Crossovers, and the Commodity Channel Index. The result of their experiments showed that the model performs better than the naïve strategy. Also Veri and Baba (2013) forecasting the next closing price at IDX, they used opening price, highest price, lowest price, closing price and volume of shares sold as experimental variables. The result showed that the most appropriate network architecture is 5-2-1 with dividing the data into two parts, with 40 training data with 95% accuracy of data and 20 test data with 85% accuracy of data. 2.3 Learning Paradigms in ANNs The ability to learn is a peculiar feature pertaining to intelligent systems 15 includes modifying the network architecture, which involves adjusting the weights of the links, pruning or creating some connection links, and/ or changing the firing rules of the individual neurons. ANN approach learning has demonstrated their capability in financial modelling and prediction as the network is presented with training examples, similar to the way we learn from experience. In this paper, a three-layered feed-forward ANN model was structured to predict stock price index movement is given in Fig. 2. This ANN model consists of an input layer, a hidden layer and an output layer, each of which is connected to the other. At least one neuron would be employed in each layer of the ANN model. Inputs for the network were twelve technical indicators which were represented by twelve neurons in the input layer. Each neuron (unit) in the network can receive input signals, to process them and to send an output signal. Each neuron is connected at least with one neuron, and each connection is evaluated by a real number, called the weight coefficient, that reflects the degree of importance of the given connection in the neural network.

4. METHODOLOGY

4.1 Dataset

Input Data: Input data is generally in video format where the data is read and described using graphs.

Reading Data: Pandas library is used to read the data from csv files.

Describing Data: In this following step we are going to describe the data in video file to know the number of rows and columns in the dataset.

Data Cleaning: It is a very important step while we are dealing with the large datasets. To achieve the efficiency in computation we are going to remove not related to crime videos.

Training and test data: Training data is passed to train the model. Test data is used to test the trained model whether it is making correct predictions or not.

DAiSEE, the first multi-label video classification dataset comprising of 9068 video snippets captured from 112 users for recognizing the user affective states of boredom, confusion, engagement, and frustration in the wild. The dataset has four levels of labels namely - very low, low, high, and very high for each of the affective states, which are crowd annotated and correlated with a gold standard annotation created using a team of expert psychologists. We have also established benchmark results on this dataset using state-of-the-art video classification methods that are available today. We believe that DAiSEE will provide the research community with challenges in feature extraction, context-based inference, and development of suitable machine learning methods for related tasks, thus providing a springboard for further research.

4.2 Implementation

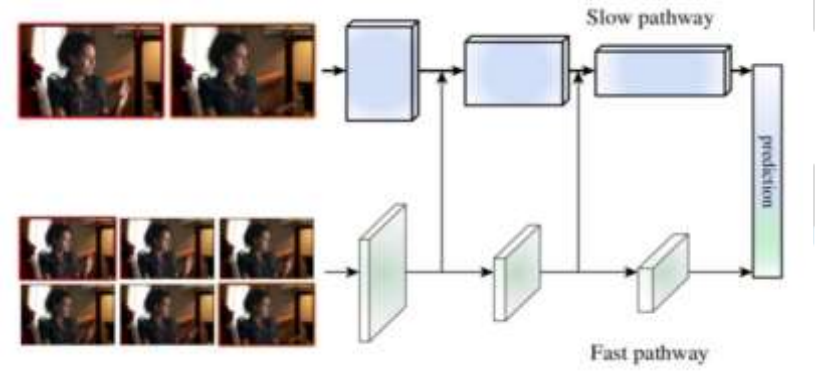
SlowFast, a new article from Facebook AI Research, proposes a whole new method for investigating the contents of a video segment, with state-of-the-art results on two famous video comprehension benchmarks — Kinetics-400 and AVA. The employment of two concurrent convolution neural networks (CNNs) on the same video segment — a Slow pathway and a Quick pathway — is at the heart of the strategy. The authors point out that frames in video scenes frequently have two separate parts: static sections within the frame that don't change much or change slowly, and dynamic areas that signify something vital that's happening right now. A video of a plane taking off, for example, will show a relatively static airport with a dynamic item (the jet) moving swiftly within it.

In a typical meeting between two people, the handshake is usually quick and lively, whereas the rest of the scene is motionless. SlowFast investigates the static content of a video using a slow, high-definition CNN (Fast pathway), while a quick, low-definition CNN (Slow pathway) investigates the dynamic content of a video. The approach is based on the retinal ganglion in primates, which has 80 percent P-cells that perceive minute details and 20 percent M-cells that are tuned in to rapid changes. Similarly, the compute cost of the Slow pathway is 4x that of the Fast pathway in SlowFast. The Slow and Fast paths both use a 3D ResNet model, which captures many frames and performs 3D convolution operations on them. The Slow pathway employs a large temporal stride (number of frames skipped per second) of 16, with approximately 2 sampled frames per second. The Fast route has a much shorter temporal stride, usually set to 8, with 15 frames per second. The Fast pathway is maintained light by using a smaller channel size (i.e. convolution width; number of filters employed) than the Slow pathway, which is commonly set at 1/8 of the Slow channel size. The Fast pathway's channel size is indicated as β. Despite having a higher temporal frequency, the Fast pathway requires 4x less compute than the Slow pathway due to the lower channel size. Lateral Relationships Data from the Fast pathway is sent into the Slow pathway via lateral connections across the network, allowing the Slow pathway to become aware of the Fast pathway's outcomes, as illustrated in the visual example. Because the shape of one data sample differs between the two paths (Fast is and Slow is), SlowFast must execute data transformation on the Fast pathway's results before summarising or concatenating them into the Slow pathway.

The paper suggests three techniques for data transformation, with the third one proving in practice to be the foremost effective:

1. Time-to-channel: Reshaping and transposing into , meaning packing all α frames into the channels of 1 frame.
2. Time-strided sampling: Simply sampling one out of each α frames, so becomes .
3. Time-strided convolution: Performing a 3D convolution of a 5×12 kernel with 2βC output channels and stride = α. Interestingly, the researchers found that bidirectional lateral connections, i.e. also feeding the Slow pathway into the Fast pathway, don't improve performance.

Putting the paths together SlowFast performs Global Average Pooling at the top of each pathway, which is a standard technique for reducing dimensionality. The results of the two routes are then concatenated and inserted into a completely connected classification layer, which utilises Softmax to classify which action is occurring within the image. On both datasets, SlowFast produces cutting-edge results. It outperforms the simplest top-1 score by 5.1 percent (79.0 percent vs. 73.9 percent) and the highest top-5 score by 2.7 percent in Kinetics-400 (93.6 percent vs 90.9 percent). It also delivers cutting-edge results on the new Kinetics-600 dataset, which is comparable to the Kinetics-400 dataset but includes 600 categories of human actions, each having at least 600 movies.



Input Data: Input data is generally in video format where the data is read and described using graphs.
Reading Data: Pandas library is used to read the data from csv files.
Describing Data: In this following step we are going to describe the data in video file to know the number of rows and columns in the dataset.
Data Cleaning: It is a very important step while we are dealing with the large datasets. To achieve the efficiency in computation we are going to remove not related to crime videos. Training and test data: Training data is passed to train the model. Test data is used to test the trained model whether it is making correct predictions or not.

5. RESULTS

5.1 UPLOADING IMAGES

The below table is a comparison of the results by the values passed in the model.

Test case ID	Test case name	Purpose	Test Case	Output
1	User uploads videos	Use it for identification	The user uploads the positive engagement video for analysis	Uploaded successfully and positive analysis is generated
2	User uploads 2nd video	Use it for identification	The user uploads the negative engagement video for analysis	Uploaded successfully and negative analysis is generated

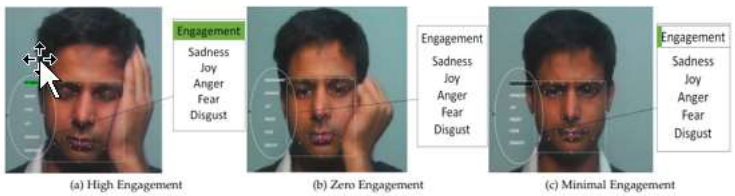
5.2 CLASSIFICATION

Test case ID	Test case name	Purpose	Input	Output
1	Classification test 1	To check if the classifier performs its task	Positive engagement video	Positive analysis is generated.
2	Classification test 2	To check if the classifier performs its task	Negative engagement video	Negative analysis is generated.

5.3 Accuracy

Most of the employed machine learning algorithms performed good with the accuracy of 60%.

Performance Screenshots:



Screenshot 5.1: Engagement Analysis

Screenshot 5.2: Engagement level increasing from left to right

Screenshot 5.3: Variety in Dataset

6. CONCLUSION

Our work can accelerate the entire process of crime detection using computer vision which will result in better security and response time. Evaluation results also indicate that the proposed implementation is effective in feature selection and prediction. This method can also be applied in other related research fields by fine tuning this existing method.

7. ACKNOWLEDGMENT

We thank CMR Technical Campus for supporting this paper titled with "Automated Engagement Recognition in E-Environments", which provided good facilities and support to accomplish our work. Sincerely thank to our Chairman, Director, Deans, Head of the Department, Department of Computer Science and Engineering, Guide and Teaching and Non- Teaching faculty members for giving valuable suggestions and guidance in every aspect of our work.

REFERENCES

1. SlowFast networks: https://arxiv.org/abs/1812.03982
2. Action Recognition: https://uwaterloo.ca/vision-image-processing-lab/research- demos/action-recognition-video
3. Human action recognition methods: https://www.frontiersin.org/articles/10.3389/frobt.2015.00028/full
4. Crime in India statistics: https://ncrb.gov.in/en/crime-india
5. Crime in India statistics: http://mospi.nic.in/sites/default/files/Statistical_year_book_india_chapters/ch37.pdf
6. Image and Video Understanding: https://medium.com/stradigiai/image-and-video- understanding-an-introduction-to-computer-vision-5d83f8fa63f5
7. Image/Video Understanding and Analysis: https://www.microsoft.com/en- us/research/project/image2text/
8. CNN for Deep Learning: https://www.analyticsvidhya.com/blog/2021/05/convolutional-neural-networks-cnn/