MLG-42 Real Time Sentiment Analysis

1. Devashish N Mehta - dnmehta@iitk.ac.in
2. Nikita Awasthi - anikita@iitk.ac.in
3. Shivam Utreja - sutreja@iitk.ac.in
4. Mrinaal Dogra - mrinaald@iitk.ac.in
5. Mohd. Abbas Zaidi - mzaidi@iitk.ac.in

Problem Statement

We aim to develop a user-end application to get time-mapped viewer sentiment while watching a video from the following set of 6 different emotions:

$$E = \{\text{neutral, happy, sad, angry, surprised, fearful}\}$$

The input will be the real time viewer's webcam feed while watching the video and the output will be an emotion specified at regular time intervals throughout the video.

1 Problem motivation

- **Widespread need of genuine user feedback**
  Users are exposed to a lot of visual content online on a daily basis. This could be in the form of product advertisements, youtube videos, movie trailers etc. The creators of these will be benefited immensely if given the user's emotional feedback on being exposed to this content. The recommendation systems will also be able to use this data in order to make better recommendations in the future to the same user. This will hence, improve the experience of the consumers, and also provide with constructive feedback to the creators.
- **Drawbacks of the current user review methods**
  Presently, the only methods by which creators can get reviews of their content online are via comments, reaction buttons, reviews, feedback forms etc.. There are two major disadvantages of all these methods. Firstly, only a small percentage of users opt to fill these out (which itself is a biased set of users in the first place), of which, an even smaller percentage coincides well with the true feelings of the reviewer, since the reviews can also be affected by external factors[1]. Secondly, these responses are not real-time, but an overall summarized response of what the user thinks he felt of the content as a whole, after watching it completely. This provides a very crude feedback to the creators as well as the recommendation systems to be of any significant use.
- **Advantages of Realtime Sentiment Analysis**
  Ideally, the recommendation systems and creators will be benefitted the most if they could know the responses of the users exposed to a given piece of content. These are the pure sentiments that each specific user had exhibited throughout the duration of the video. This data set would be a lot richer, being the time-mapped review of the video. It would also be a lot larger, and less biased as it would be inclusive of all the users watching that specific content. It would also be a highly accurate estimate of the true user emotions, as opposed to what the user chooses to mention in his/her written review.

2 Previous Work

- A number of features are used to analyse user sentiment including voting, rating, etc. Comments are one of the most informative features and they try to analyse user comments on video to attach a sentiment to the video. [2]
- A framework for classifying images according to high-level sentiment which subdivides the task into three primary problems: emotion classification on faces, human pose estimation, and 3D estimation and clustering of groups of people. [3]
- A project was made using deep learning techniques to predict the emotion depicted by an image. [4]
- Currently, sentiment analysis on Youtube profiles of the person hosting the video or the comments by users to identify possible polarisation by the video. However as mentioned in [5], to gauge the extent of radicalization possible , the cue is basically text.
- For social media analysis using tweets for analysing user feedback and sentiment, the existing work predicts the emotion labels as positive, negative or neutral. This can limit the amount of information that can be extracted from the content for feedback to content creators (videos in case of Youtube). This might be disadvantageous for content creators who wish to maximise the reach of their content by analysing the highs and lows of the video mapped to time frame. [6] [7]

2.1 Further Discussion and Problems

- Most of the existing methods rely on some sort of feedback received from a user. As we discussed above it cannot be trusted as a reliable source of information about the actual feelings of an individual. Therefore using tweets, comments or reacts as the feedback may not actually serve the purpose when the analysis is based to gauge user feedback to a particular commodity/ product.

3 Novel Contribution

With a greater participation of individuals on social media, coporates have also shifted to social media for popularisation of their products. Revenue generation is directly linked to the reach of any product. With advertisements shifting to a video based platform, it is important that the content creators receive a more detailed feedback on their content. Our application proposes to bridge that gap between the producers and the consumers. With a time-mapped sentiment feedback to provide better analysis, the producers have an opportunity to identify their weak points. The current feedback systems lag behind in this regard. Most of the review systems rely on cumulative feedback at the end of the video in the form of comments or reactions. We create an end-to-end application that takes a viewers' video feed as input and gives the time mapped graph of the sentiments.

4 Methodology

4.1 Data

- Cohn-Kanade face database was compiled by researchers at Carnegie Mellon University, this database is easily available upon request using the EULA form. The dataset is available in two versions CK and CK+, Version 1, includes 486 sequences from 97 posers and is referred to as CK. It consists of a range of expression from neutral to peak. The peak expression is EMACS coded. However the given labels correspond to the requested expression and not the actual expression. Cohn Kannade 2 Version 2 of the data set includes both spontaneous as well as requested set of images. It also provides standards for facial feature tracking and sentiment analysis. The CK/ CK+ dataset provided a set of 120 images, we expected to extract around 100-150 more images using FACS to emotions conversion. This has been discussed in great detail later. [8]
- ISED Dataset or Indian Spontaneous Expression Database provides frontal images corresponding to video while watching emotion inducing clips. The labelling is done by 4 decoders and validated by the stimuli and the self-report of sentiments. It covers 4 emotions

which include include happiness, disgust, sadness, and surprise. The dataset comprises of around a set of 500 peak images.[9]

- FERC 2013 dataset has a collection of $28,709$ low resolution images. Each of these images is a 48×48 pixel grayscale image of the face, with the face being centered and occupying roughly the same area in each. The labels for this dataset are not posed unlike CK+, making them hard to classify, however owing to the large size of our data set this comes to benefit since the classifier becomes robust.

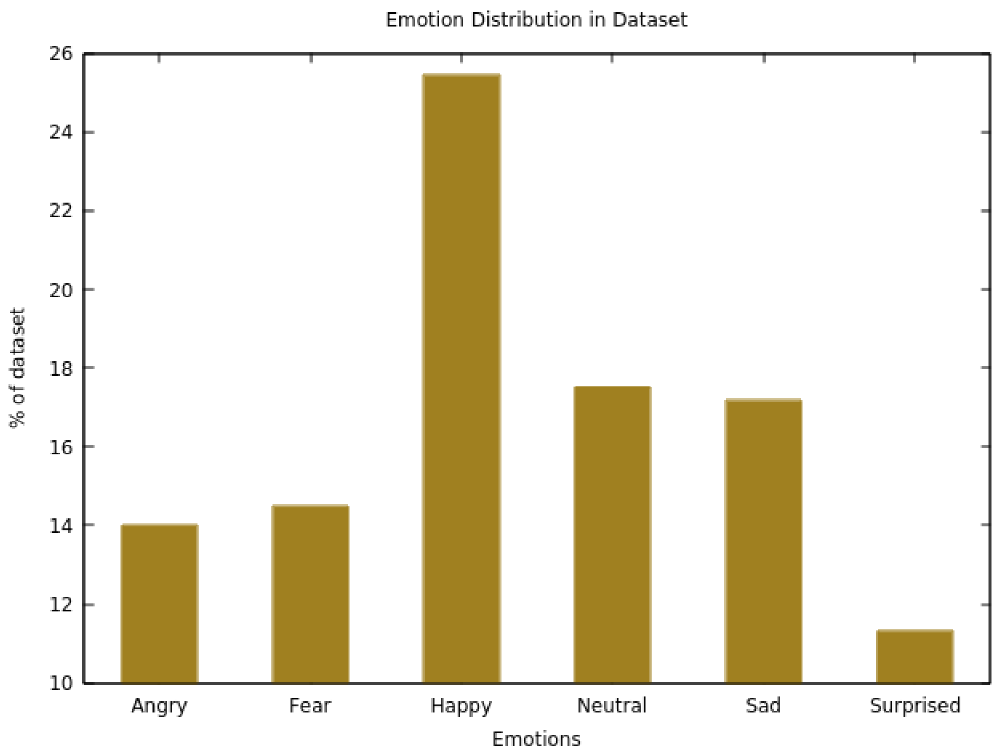

Figure 1: Emotion Distribution in Dataset

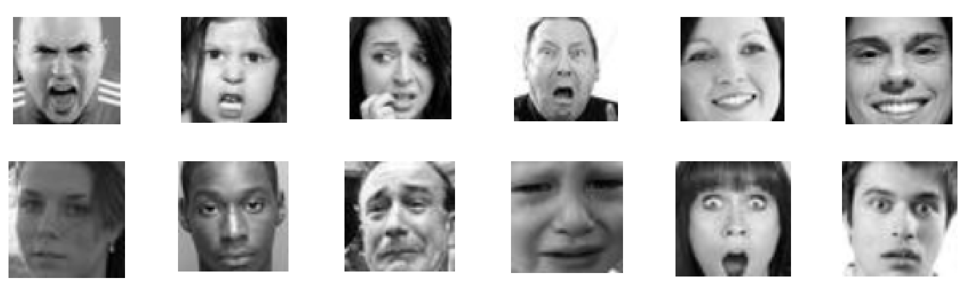

Figure 2: Sample Images from FERC Dataset

4.2 FACS vs EMFACS

FACS or Facial Action Coding System- It refers to the facial muscle movements which cause an emotion. The system arranges sentiments by assessing the movements of muscles at the face, this enables us to code any emotion/ expression. FACS is the standard for automated system which analyses faces in videos. FACS labels each component of observed facial movement in form of Action Units(AU). This is the only technique using which we can read emotions from an image/ video in real time. FACS eliminate the requirement of an individual to label a video or image as per the sentiments.[10]

It was planned to use EMFACS, because it seemed to be highly appropriate for sentiment based analysis, also it would save time and allow us to use a more complex architecture, however the problems have been discussed below. EMFACS or emotion FACS- Under EMFACS the FACS is applied selectively, only those images which are likely to have any emotional significance are coded. Prototypes which are of emotional significance, the application of FACS is decide by these prototypes. It is obvious that EMFACS saves on times as compared to normal FACS. However since no definite standard has been defined it is difficult to obtain consistency on EMFACS coding. Moreover, the set of instructions EMFACS delivers to selectively use FACS which are available only to those who obtain certifications on FACS(by Paul Ekman). [11]

4.3 Convolutional Neural Networks

In machine learning, a convolutional neural network(CNN) is a class of deep, feed-forward neural networks. CNNs use multiple hidden layers of nodes which, in an abstract sense transfer information on recieving a stimulus. As images have strong spacial structures, these types of neural networks are apt for the task at hand.

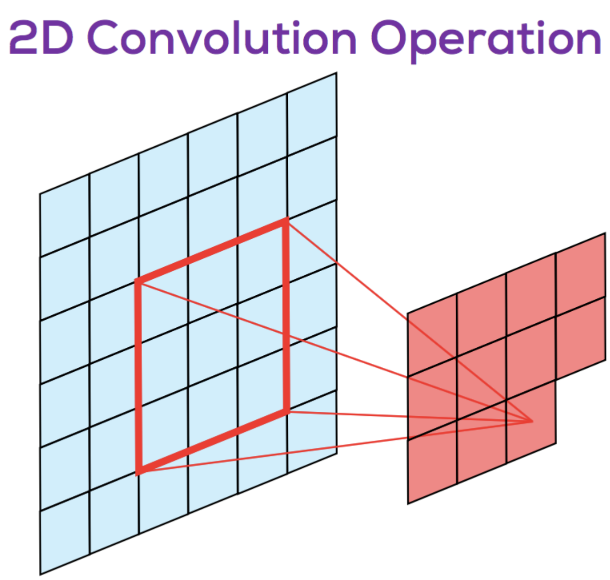

Figure 3: 2D Convolution[1]

As the images have many pixels, it is not possible to use fully connected layers and hence it is useful to use some modifications to the network architecture so that we are able to efficiently train the CNN on many images from the dataset.
In a convolutional neural network, layers are sparsely connected with parameter sharing which drastically reduces the number of parameters to be learnt and thus, reducing the training and testing time. These are the type of local operations which are desired to capture objects like images (finding patterns locally) and are called convolution operations. These operations determine features like

[1]Image courtesy: Purushottam Kar , Course CS771 IITK

edges,etc using kernels.
The second type of operation used is the pooling operations which aims to make the network insensitive to small/minor changes. Some nodes at fixed intervals (strides) are selected and a function is applied to them in order to create only one node representing all of them. The common type of functions used here are the max or average function.

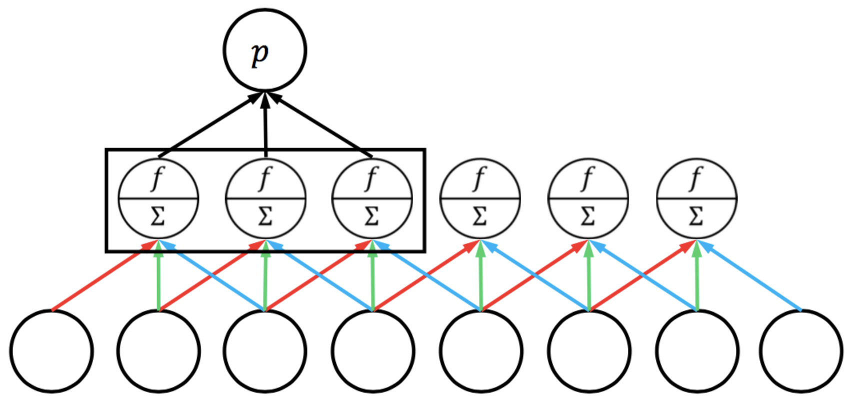

Figure 4: Pooling Operation[2]

4.3.1 Variants of CNNs tried

1. **LeNet**
   These are very basic multilayer neural networks which constitute the backpropagation algorithm. These Gradient based classifiers can be used to extract high level complex features depending on the network architecture. As discussed in [12] Lenets were initially used for the purpose of document extraction. Lenets are seen as learning relevant features in initial layers and training on the given set of features for the coming layers. Although artificial intelligence is useful but it is not possible to avoid the bias, which manifests in the bias on the architecture of the network, therefore the architecture should be specific to the problem at hand. The output of each layer is expressible as a function of inputs from previous layers and the edge weights, using which the gradients are found and optimum values of edge weights are found, The lenet used is basically a 5 layer neural network for the task, learning parameters using gradient descent. It consists of repeated convolutional followed by pooling(max pooling), followed by a final hidden layer and an output layer. The activation function is ReLU in hidden layers and finally softmax at the output. However the accuracy obtained through normal lenet is not satisfactory, hence we tried to modify it to improve accuracy.

2. **Modified LeNet**
   As the observed acuracy on lenet was very poor, we decided to tweek the network parameters and other hyper parameters with the aim to obtain better accuaracy. Deep learning networks are very hard to train (in general NP hard) and thus we have no other option than trying new parameters and functions . The activation function used was the same , ReLU . We now used the Adam Optimiser which is can used instead of the classical stochastic gradient descent procedure to update network weights iterative based in training data. We now have 7 layers in our architecture instead of 5. The training time increased as we are now using more layers than before. The perfomance increased a little than before but was still unsatisfactory. The performance for lenet and modified lenet are tabulated below.

[2]Image courtesy: Purushottam Kar , Course CS771 IITK

CNN Model Used	Top-1 Accuracy (%)
LeNet	29.6
Modified LeNet	36.3

Table 1: Top-1 Accuracy for Lenet and Modified Lenet

3. **MobileNets**[13]
   MobileNets are basically streamlined light weight network architectures for mobile and embedded applications. They are more efficient with respect to size and speed, while maintaining almost similar accuracies. They are able to achieve this by inculcating depthwise separable convolutions in their architecture, which basically breaks the interaction between the number of output channels and the size of the kernel. This convolution technique is explained below.
   It is made up of 2 layers, depthwise convolutions and pointwise convolutions. The depthwise convolutions apply a single filter to each of the input channels, but does not combine them further to create new features for the output layer. This is where the 1×1 pointwise convolutional layer comes in. It computes a linear combination of the output of the depthwise convolution layer, via 1×1 convolution, for each of the channels of the output to produce the final output's feature map.
   By using this, we get an overall complexity reduction of $\frac{1}{N} + \frac{1}{D_K^2}$, compared to the standard, fully connected CNN; where N is the number of output channels and $D_K \times D_K$ is the spatial dimension of the kernel.

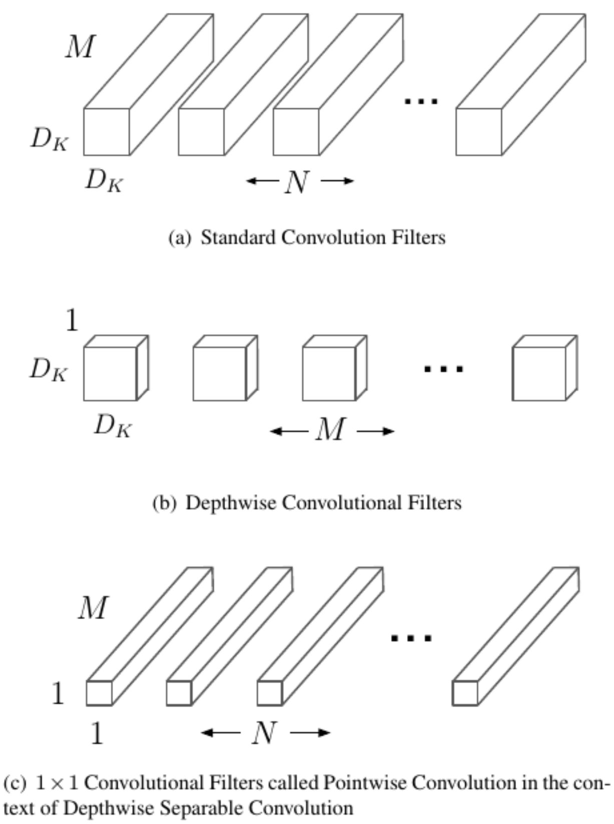

Figure 5: Depthwise Separable Convolutions[13]

In Figure 5 above, M is the no. of channels in the input layer, N is the number of channels in the output layer and $D_K \times D_K$ is the spatial dimension of the kernel.
Due to low levels of accuracy using the above methods, research material was referred which could help to modify the CNN architecture in a task-specific manner, as discussed in [14], the problem involves observing high level abstractions, hence a number of fully connected nodes were also employed. Many changes were applied to the existing CNN, however the improvements in accuracy were not significant,therefore we shifted to a larger CNN model

4. **AlexNet**[15]
   Some features of the AlexNet are:
   - 7 hidden weight layers
   - 650K neurons
   - 60 million parameters
   - 630 million connections

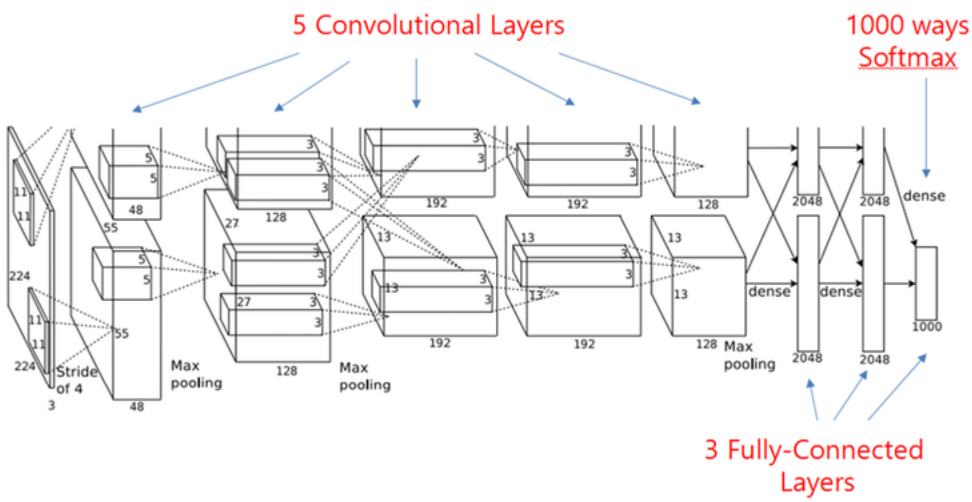

Figure 6: AlexNet Architecture[3]

AlexNet is a larger neural network with 7 hidden layers, the first 5 layers being convolutional layers and last 2 fully connected layers. But it contains around 60 million parameters, and is meant to classify among 1000 classes. Also, it requires a lot of computational resources to train an AlexNet from scratch. Since we only need to classify among the six emotion classes, we tried a smaller modified version, with only 3 convolutional layers and 2 fully connected layers along with reduced number of parameters for these layers.
Meanwhile, we researched as why the above mentioned neural network models were performing so poor. One fact that came out was we were training all these models for atmost 2000 steps, which is around 3 epochs for a batch size of 32 images and a very less quantity. So for this model, we gradually increased the number of epochs and found that the accuracy of our model improved with increase in the number of epochs.

4.4 Pre-processing of Data

Open CV Haar Cascade- It is a classifier for object detection, used to detect faces in our case. The number of neighbors in this classifier were tuned and finally set at 5. This helps to eliminate false positive detections from the image, it means that any bounding box needs to be surrounded by atleast 5 other bounding boxes for it to be classified as a face. This also ensured that images when being sent to the classifier during training and test time (consisting of 48 x 48 gray scale images) are centred at the face. [16]

[3]Image Source: `https://world4jason.gitbooks.io/research-log/content/deepLearning/CNN/Model%20&%20ImgNet/alexnet/img/alexnet2.png`

4.5 Results

The accuracies for various CNNs employed are tabulated below, different values for MobileNet correspond to different values of hyper-parameters α(the width parameter) and ρ(the resolution parameter).

CNN Model Used	Top-1 Accuracy (%)
LeNet	29.6
Modified LeNet	36.3
MobileNet_0.25_128	38.2
MobileNet_0.25_160	44.4
MobileNet_0.25_192	42.4
MobileNet_0.25_224	35.0
MobileNet_0.50_128	37.6
MobileNet_0.50_160	43.9
MobileNet_0.50_192	43.2
MobileNet_0.50_224	44.1
MobileNet_0.75_128	35.4
MobileNet_0.75_160	44.2
MobileNet_0.75_192	41.3
MobileNet_0.75_224	42.5
MobileNet_1.0_128	43.6
MobileNet_1.0_160	44.0
MobileNet_1.0_192	42.9
MobileNet_1.0_224	38.9
Smaller AlexNet	**74.7**

Table 2: Top-1 Validation Accuracies for various CNNs employed

The following graph shows the detected labels as a function of time, this is what a content creator would desire, a temporal feedback of the content viewed by the user.

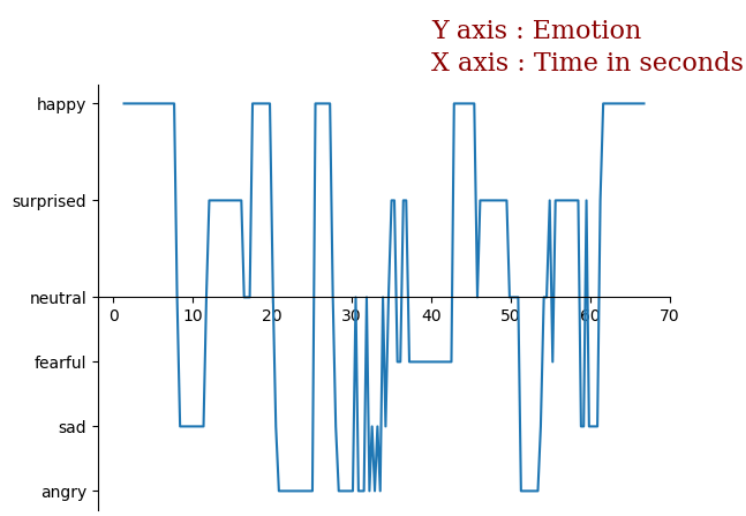

Figure 7: Peak Labelled Emotion during Testing for Final Classifier

The Recall and Precision for various labels and the overall accuracy for the final model has been tabulated below:

Emotion Label		Top-1 Accuracy (%)	Top-2 Accuracy
Angry	Recall	30.69	54.24
	Precision	37.01	65.42
Fearful	Recall	22.62	49.94
	Precision	31.78	70.18
Happy	Recall	70.52	80.72
	Precision	70.72	80.94
Sad	Recall	39.33	70.47
	Precision	28.38	50.85
Surprised	Recall	32.34	74.33
	Precision	61.73	73.60
Neutral	Recall	49.70	71.89
	Precision	44.84	64.86
Overall		53.29	71.16

Table 3: Final Model Results over Test Data

5 System limitations and Privacy Concern

With privacy being a major concern with our proposed system which claims to monitor viewer's activity through a webcam feed, we propose the following solutions to the posed problem.

- Ask permission of the user to allow sharing of his feed right before they start viewing your sentiment data
- Do not include this as a feature of the final application but use this system on a small set of volunteers to identify the strong and weak points

6 Possibilities for Future Work

The data generated by using this CNN is rich enough to be exploited in several applications to derive better insight about both,the users as well as the content.

- **Improvement of existing online recommendation systems**
  The users' emotional state while browsing the web can be monitored and this can be used to decide what content/advertisements should be recommended next to the user. This will help the recommendation system reach its predictions, using both, the past online behaviour of the user with the website/application, as well as his present emotional state.
- **Depression detection applications**
  Highly frequent, content independent instances of sad/disgusted/angry sentiments can be used by applications to detect possibilities of depression in a user well before the problem becomes chronic.
- **Exploiting the information along temporal dimension using RNNs**
  Since the data from a video, is in form of a stream or sequence of images, it would be much better if the algorithm could give feedback to the future evaluations, this is inherently the idea behind the working of an RNN. We can implement this to make the output robust, as it would take in feedback and thus it will be less prone to noise.
- **Improving Robot Interaction Experience**
  The responses of robots while interacting with humans can be made more emotionally informed by using this technique with a camera mounted on the head of the robot.

References

[1] "A discussion on youtube forum, are likes and dislikes really important." https://productforums.google.com/forum/#!topic/youtube/3jbHpNIZpT8.

[2] M. Z. Asghar, S. Ahmad, A. Marwat, and F. M. Kundi, "Sentiment Analysis on Youtube: A Brief Survey," *CoRR*, vol. abs/1511.09142, 2015.

[3] Z. Hussain, T. Patanam, and H. Cate, "Group Visual Sentiment Analysis," *ArXiv e-prints*, Jan. 2017.

[4] V. Gajarla and A. Gupta, "Emotion detection and sentiment analysis of images," in *Georgia Institute of Technology*, 2015.

[5] A. Bermingham, M. Conway, L. McInerney, N. O'Hare, and A. F. Smeaton, "Combining social network analysis and sentiment analysis to explore the potential for online radicalisation," in *2009 International Conference on Advances in Social Network Analysis and Mining*, pp. 231–236, July 2009.

[6] A. Go, L. Huang, and R. Bhayani, "Twitter sentiment analysis," *Entropy*, vol. 17, p. 252, 2009.

[7] H. Wang, D. Can, A. Kazemzadeh, F. Bar, and S. Narayanan, "A system for real-time twitter sentiment analysis of 2012 us presidential election cycle," in *Proceedings of the ACL 2012 System Demonstrations*, pp. 115–120, Association for Computational Linguistics, 2012.

[8] "Details of cohn kahnade dataset." http://www.pitt.edu/~emotion/ck-spread.htm.

[9] "Ised database." https://sites.google.com/site/iseddatabase/.

[10] "Facs." https://en.wikipedia.org/wiki/Facial_Action_Coding_System;https://imotions.com/blog/facial-action-coding-system/;https://www.paulekman.com/product-category/facs/.

[11] "Emfacs." https://www.paulekman.com/facs-faq/.

[12] Y. LeCun, L. Bottou, Y. Bengio, and P. Haffner, "Gradient-based learning applied to document recognition," *Proceedings of the IEEE*, vol. 86, no. 11, pp. 2278–2324, 1998.

[13] A. G. Howard, M. Zhu, B. Chen, D. Kalenichenko, W. Wang, T. Weyand, M. Andreetto, and H. Adam, "Mobilenets: Efficient convolutional neural networks for mobile vision applications," *CoRR*, vol. abs/1704.04861, 2017.

[14] Q. You, J. Luo, H. Jin, and J. Yang, "Robust image sentiment analysis using progressively trained and domain transferred deep networks.," in *AAAI*, pp. 381–388, 2015.

[15] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "Imagenet classification with deep convolutional neural networks," in *Advances in Neural Information Processing Systems 25* (F. Pereira, C. J. C. Burges, L. Bottou, and K. Q. Weinberger, eds.), pp. 1097–1105, Curran Associates, Inc., 2012.

[16] "Haar cascade." https://docs.opencv.org/3.3.0/d7/d8b/tutorial_py_face_detection.html.