Affective Video Tagging

Mentor : Professor US Tiwari

Members:

Swapnaneel Nandy (IIT2014111)
Utkarsh Srivastava (IIT2014507)
Mohd Abdullah (ISM2014004)
Amit Vijay (IIT2014110)
Shivam Beri (IIT2014159)

Motivation

- As digital media world is evolving the multimedia content on the web is becoming larger and larger. Since the rate of increase of data on such systems is exponential, it becomes challenging to retrieve contents that would be appealing based on emotional disposition of the users.

- A system that retrieves multimedia contents according to the user's current mood is of great interest. There are a lot of everyday applications of such systems.

  For example : enabling video streaming websites like YouTube and Netflix to show more convincing recommendations, better parental control by allowing parents to choose what their children watch by knowing the emotional content of the video.

- As manual accurate glossary of the huge amount of multimedia content on web is obviously impractical, a huge chunk of those content remain untagged, which hinders the search process and reduces the performance of search systems.
- An automated system which tags videos based on the emotion elicited by the videos can be of paramount use and importance. This has been our main goal for the project.
- This topic is trending and a huge amount of research and experiments are being done on it at the moment.

Objective

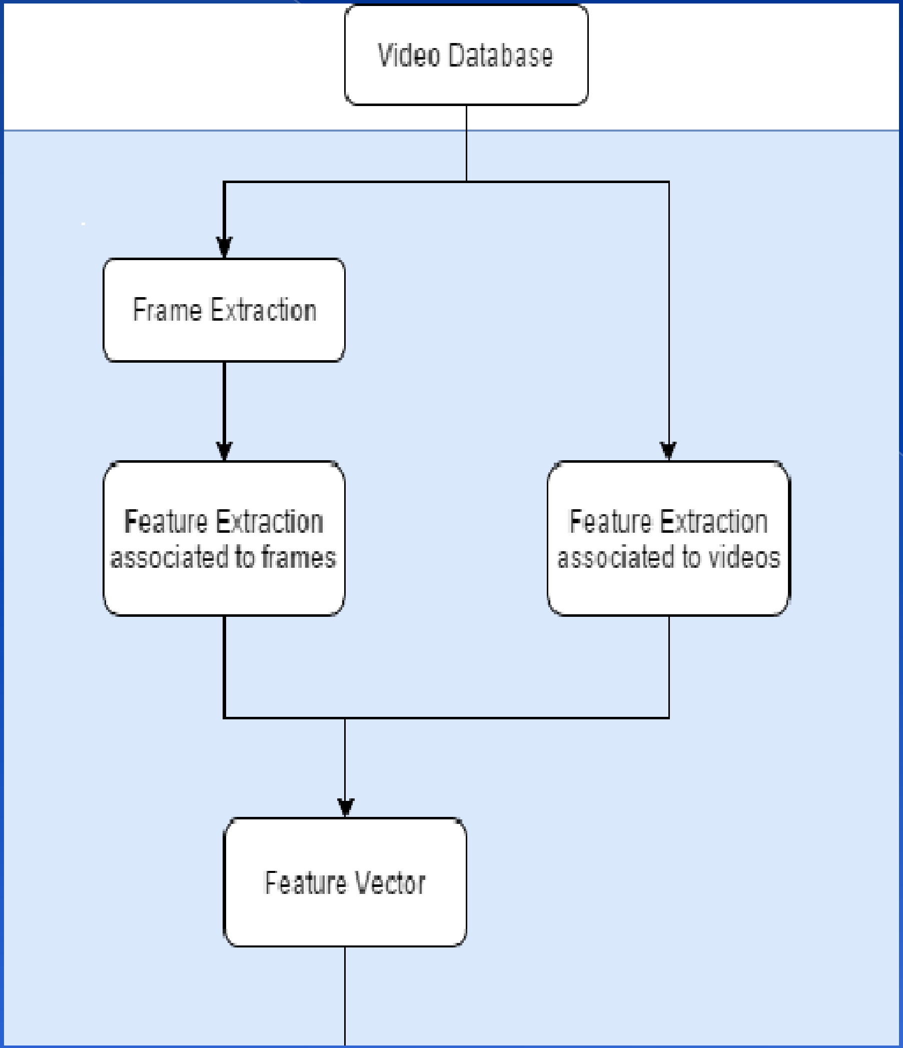
- Affective video content analysis aims at automatic recognition of emotions elicited by videos. We aim to implement an efficient affective video analysis system that can be used to tag the multimedia content .

- We planned to do this through feature extraction and using those features to create a classification model using machine learning algorithms.

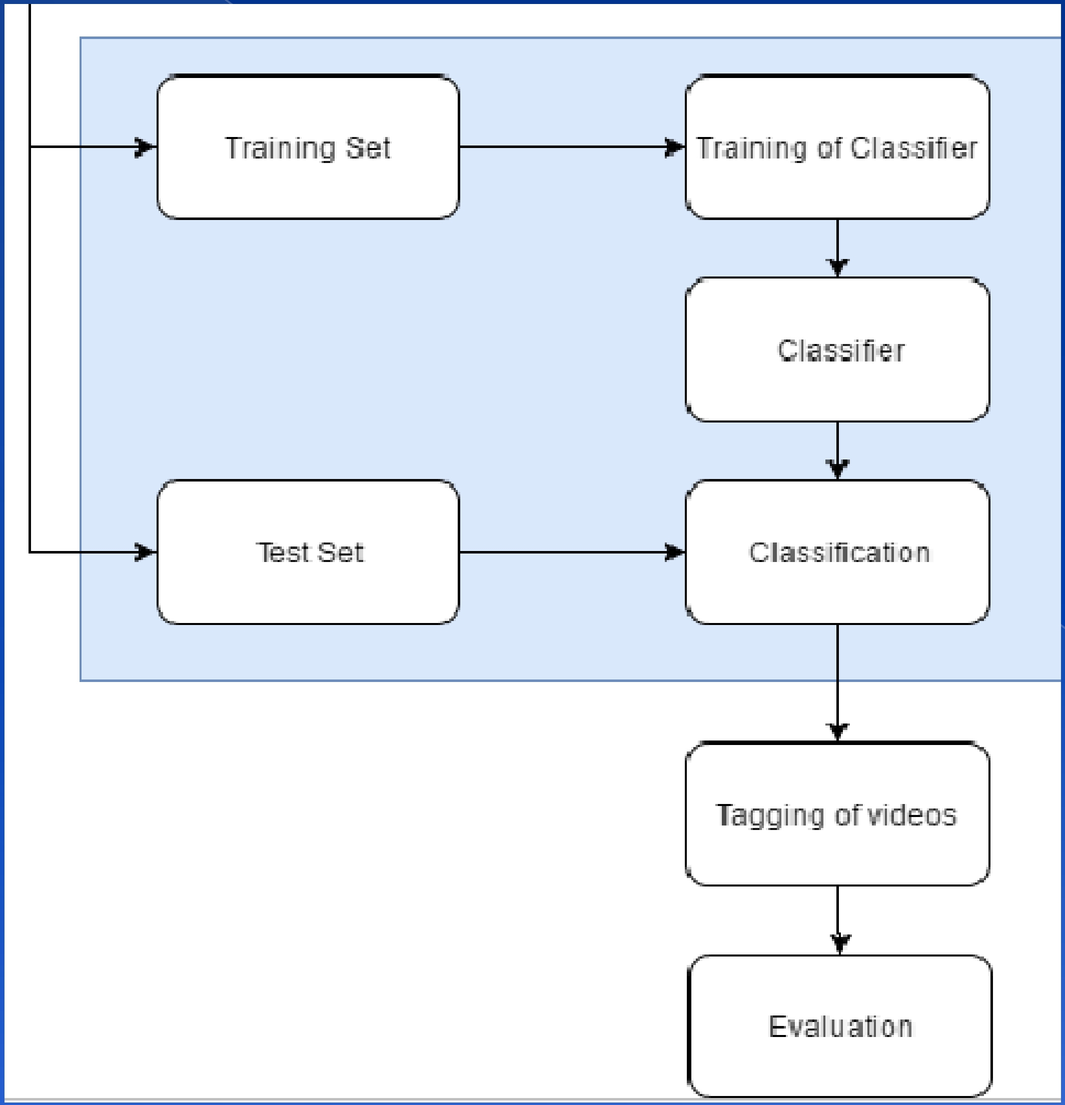
- This model can then be used to tag online multimedia videos for better searching and indexing.

- The dataset used for training the system is LIRIS-ACCEDE, which has over 9800 videos.

Methodology


Video Database
Frame Extraction
Feature Extraction associated to frames
Feature Extraction associated to videos
Feature Vector

- Various audio and visual features are extracted.
- Using these features, a feature vector is generated for training the classifier.
- The data set is divided in a ratio of 70 and 30 for training and testing respectively.



- Training and classification is done using KNN (K nearest neighbours) algorithm.
- From the output of the classifier, we calculate the accuracy of our result on various scales.
- Further we did extensive experiments on the arousal and valence values given in the data set by clustering data using K-means and then using KNN for classification.

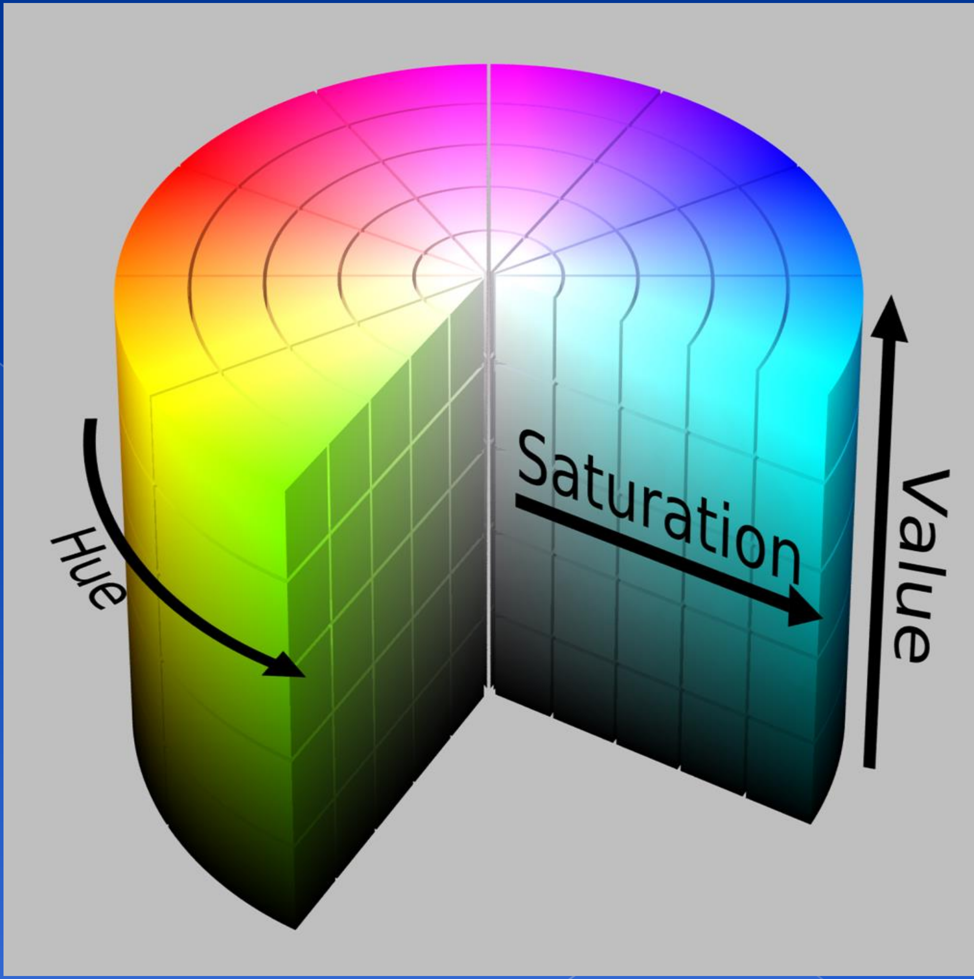
Feature List

The features we have decided to extract for our model are:

- **Global Activity**: Measures apparent motion in the video.
- **Number of scene cuts per frame**: Measures shot cut rate of the video.
- **Spectral Slope**: It is the measure of inclination of power spectrum.
- **Lighting**: Measures average luminescence of the video.
- **Colorfulness**: Measures average amount of color in the video.
- **Zero Crossing Rate** : Measures the **rate** of changes in sign of signal. Used to classify sounds.
- **Spectral Flatness** : Used in digital signal processing, it is used to measure the audio or signal spectrum.
- **Length of scene cuts** : Analyses the length of each scene in the video
- **Hue count**: Measures the number of unique hues in the video.
- **Depth of field**: is the measure of distance between closest and farthest object in an image.
- **Compositional balance**: is measure of how well visual weight is balanced around the mid vertical axis.

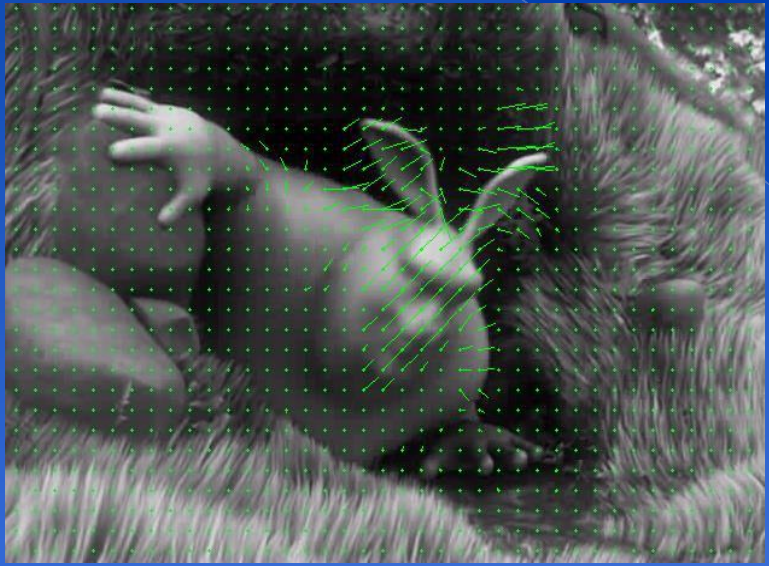
Hue Count

- Hue count is a feature which counts the number of unique hues in the video. It gives a measure of colour vibrancy in the video. A higher number of hues generally implies that the video elicits a positive emotion. In other words higher hue count symbolises positive valence.



Global Activity

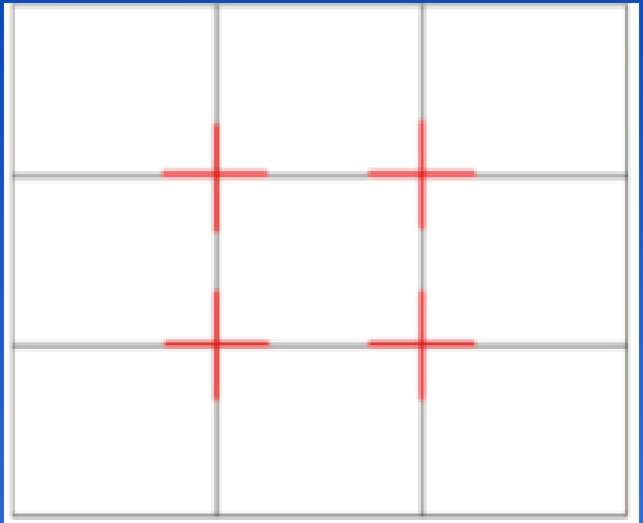
- Global Activity is a feature which captures the average amount of motion in a video. The apparent motion in a video can be captured using motion vectors. These motion vectors can be approximated by calculating the optical flow in a video.
- Optical flow or optic flow represents the apparent motion of objects in a visual scene which occurs due to the relative motion between an observer (the camera) and the scene.

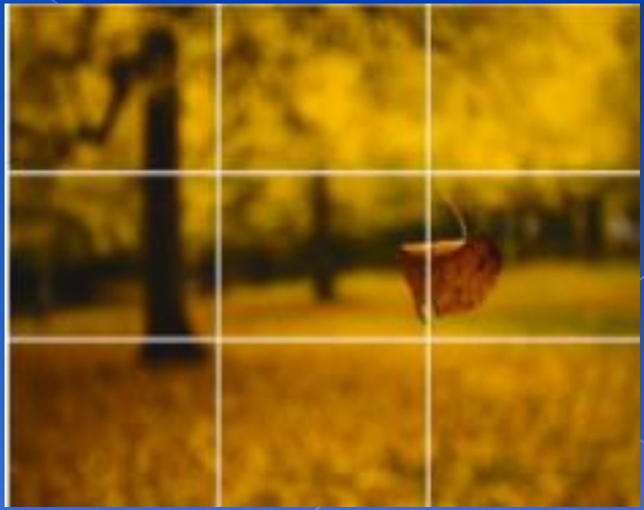


Compostional Geometry

- Compositional balance is a crucial requirement for fine quality of photos. When different parts of an image get equal attention of viewer, we say that image is perfectly balanced in terms of composition. The more balanced an image is the more pleasing it will be to the eyes.

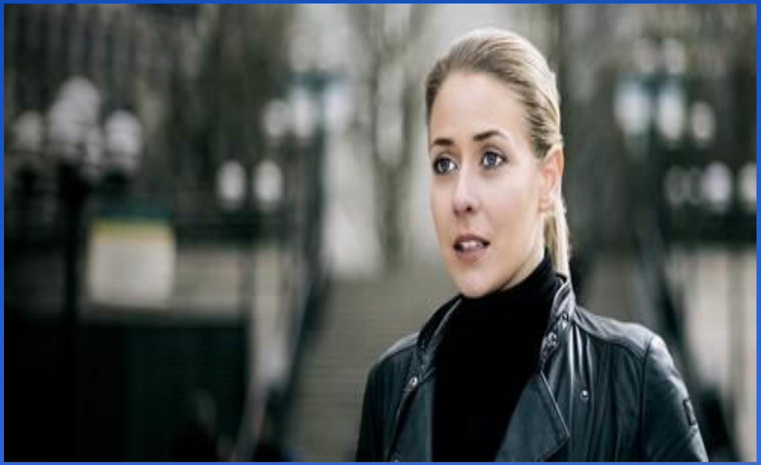
  Almost each photo has its lead character, the subject, around which the image is created and this subject is the region of interest for viewer. This subject acts as an anchor of that image.

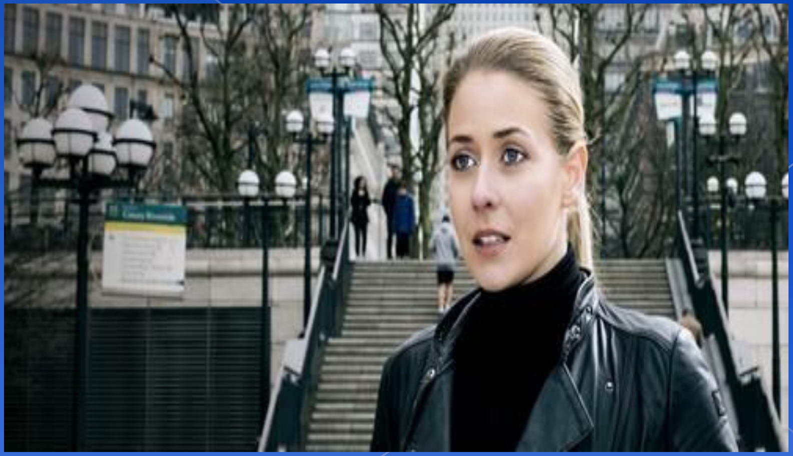




Depth of Field

- Theoretically depth of field is a parameter which represents the distance between the farthest and nearest objects in a normal sharp image. It is a perfect tool to direct viewer's attention to any specific portion of the scene. In this way cinematographers assures that he doesn't let the viewer miss any minute detail such as some deep emotions of face,emphasizing or de-emphasizing on background ,etc. Shallow DOF signifies negative valence (like tension,sad,stressed) whereas deep DOF signifies positive valence (like calm and relaxed).

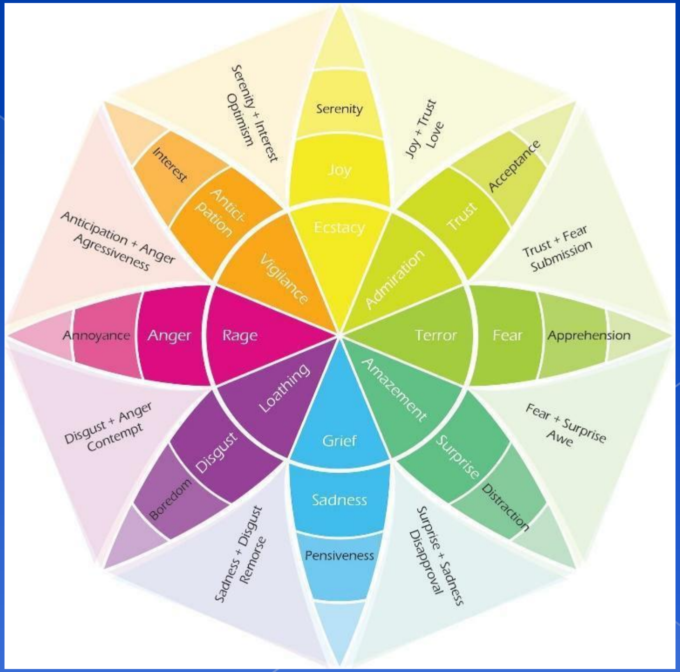




Colorfulness

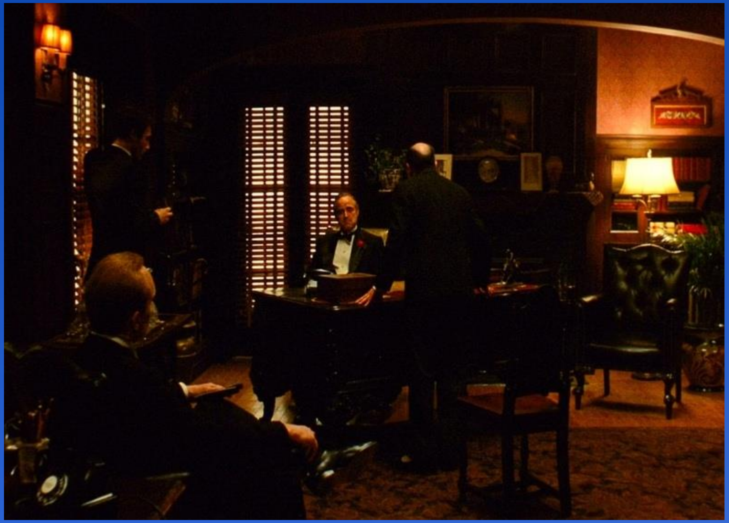
- The choice, contribution and contrast play a very important role in determining the affection conveyed by the video. The recent researches in psychology validate this fact, that color is an effective and precise non-verbal code of communication. Each color provokes specific subconscious reaction, thus evoking emotions.

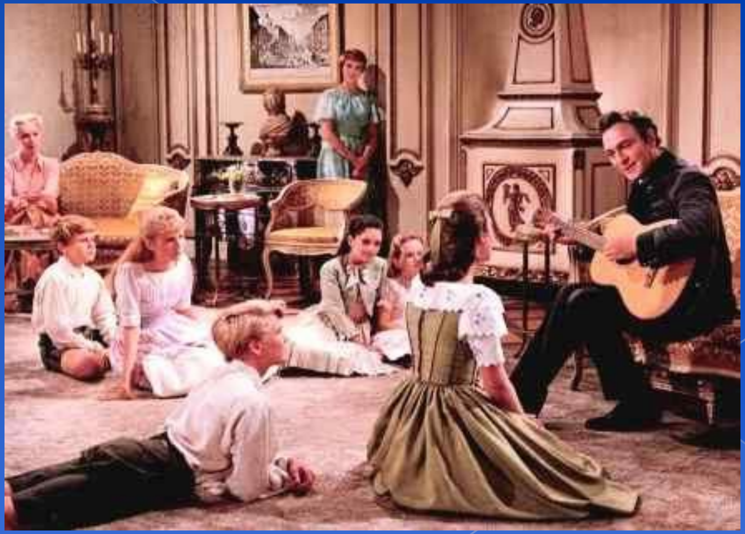
Eg : Orange and Red are considered as "warm" colors, and they correspond to energetic and vigorous emotions such as 'anger' or 'fright'.



Lighting

- Lighting is a metric to measure the contrast in the luminescence of an image. It is a important feature to determine the affection of the scene. For an example, the scenes with positive emotions, that is involving pleasantness or joy are most often depicted with high luminescence, whereas those with negative feeling like sadness or horror contain low lighting value. It is a vital tool used by artists and film creators. Dramatic effects of scenes are enhanced by manipulating the intensity of light.





Length of Scene Cuts

Whether the scene stays on for long or ends Shortly, length of a scene cut conveys something to the viewer.

‘Holding Long’ : Longer scene builds tension, suspence and anticipation.

Long Scene ➔ Low arousal

‘Cutting Short’ : Length is proportional to pacing of the scene. Shorter the scene, faster the pacing. And faster the pacing, more energy is into that scene. So anger, excitement, happy can be judged on the basis of the length.

Shorter Scene ➔ High Arousal

Zero Crossing Rate

- The **zero-crossing rate** is the **rate** of changes in sign of signal. The frequency or how fast the signal changes to and fro from positive to negetive and back. In Music information extraction and Speech recognition, ZCR has been used consistently. It can also be used to classify sounds.

- A reasonable generalization is that if the ZCR value is high the speech signal is unvoiced. While if the ZCR is low the speech signal is voiced. This is mainly because high frequencies imply high ZCR and there is a strong correlation between ZCR and energy distribution of the signal.

- It can also be also used to detect weather a a human is present in the scene or not.

Spectral Flatness

- Spectral flatness is a quantitative measure of how noisy or tonal a sound is.

- It has been observed that the tonal would resound to the number of peaks in the spectrum whereas noisy corresponds to a flat spectrum.

- So from the shape of the power spectrum, we can deduce how tonal (tending to 0.0) or noise (tending to 1.0) the sound is. This can be used to detect if someone is speaking in the clip. If the spectral flatness is between 0 and 0.25, we can detect someone speaking, rest is noise.

More flat the signal, more tonal the sound is and more calm and sad it will be perceived to be.

More peaks and crests mean more noise, or excitement or anger.

Number of Scenes cuts per frame

- The number of scene cuts per frame can also be said as the shot cut rate of a video. It is one of the low level features for arousal in video indexing along with Color and Motion. This is generally proportional to the arousal of the video. The more the scene cuts are in the video, the more will be the arousal of the video. In case if two video have same number of scene cuts, the one having less number of frames have more value for number of scene cuts per frame.

Eg :Whenever the director wants to get the viewer excited, he increases the scene cut rate. Similarly whenever the scene cut rate is low, the arousal induced is low.

Spectral Slope

- Spectral tilt, or spectral slope, is an important parameter in voice synthesis and voice perception. It is a measure of voice quality.Those voice quality include harsh, tense, breathy etc.

- The value of the spectral slope represents the amount of decrease of the spectralamplitude. It shows how rapidly the amplitudes of successive partials (component frequencies) decrease as the value of the frequency increases.

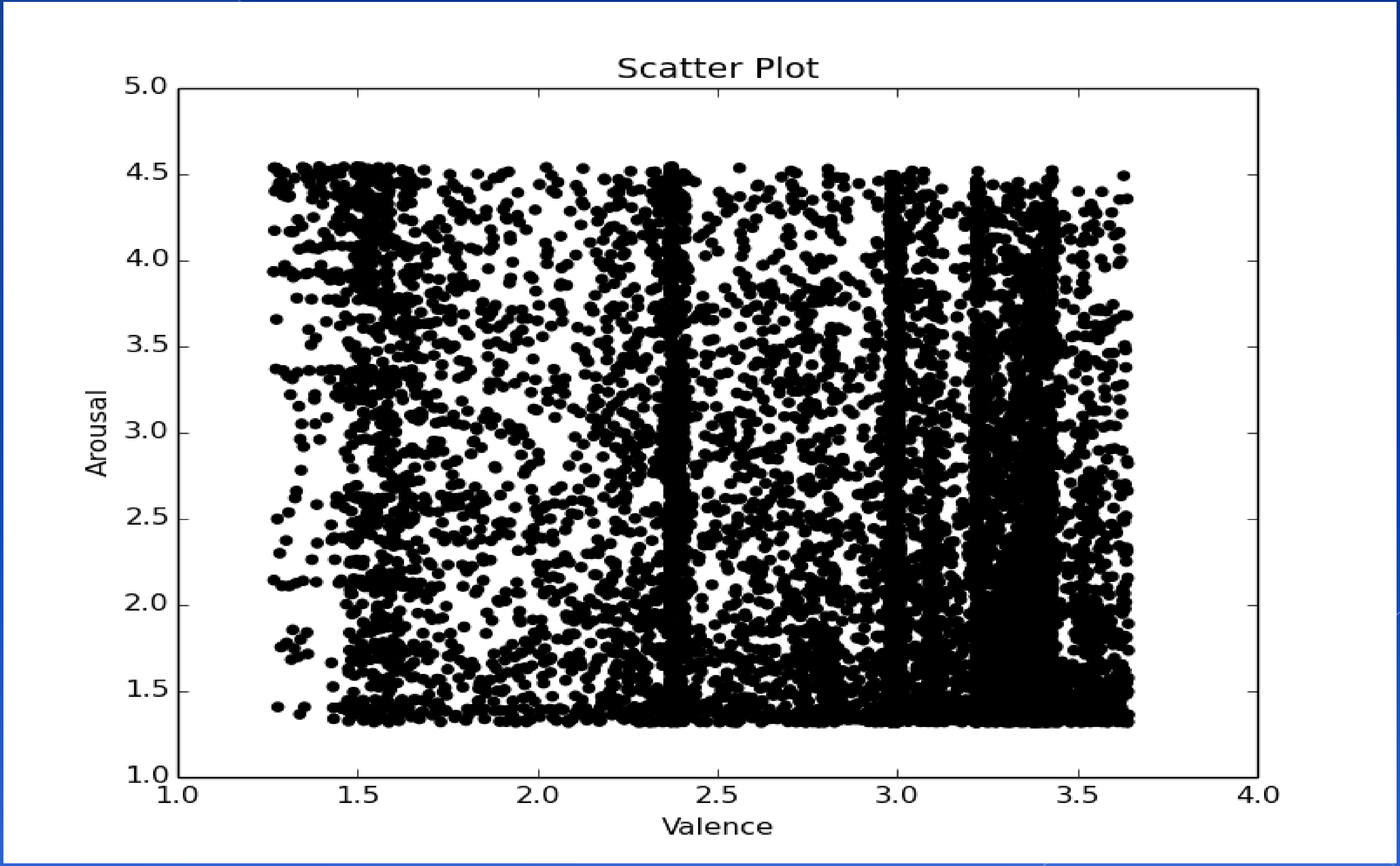
- A less steep spectral slope represents to a louder voice and when the loudness of the audio decreases, the spectral slope increases.

Classification and Clustering Algorithms

- We have used the KNN (K Nearest Neighbor) algorithm to classify our data, based on the tagging given in dataset.

- We have also used K-means algorithm to cluster the data based on arousal and valence values and then applied KNN on the new clusters obtained.

Valence – Arousal Scatter Plot

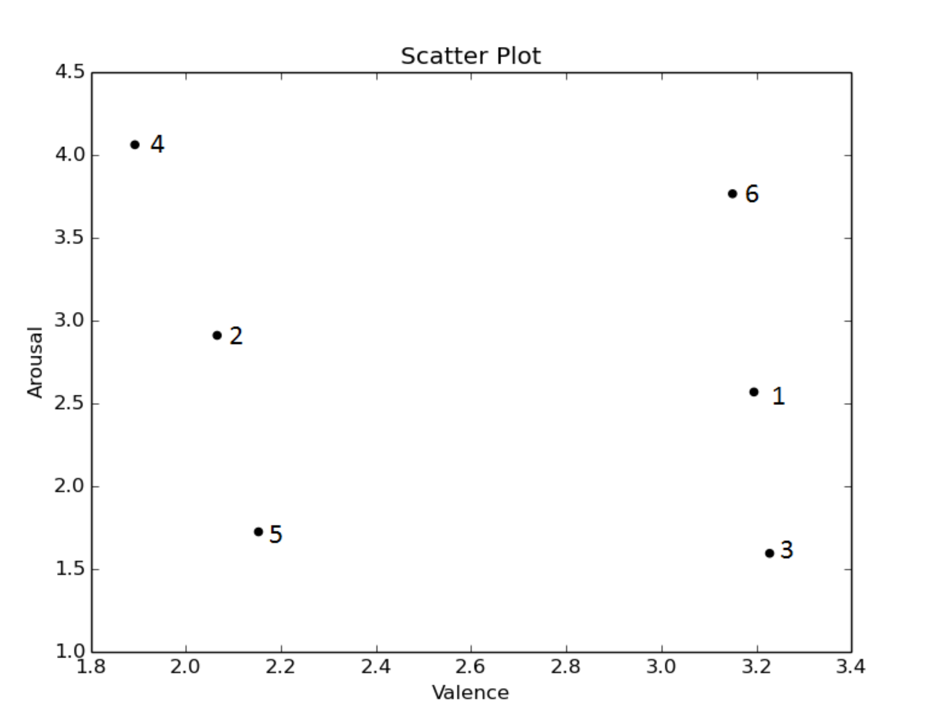


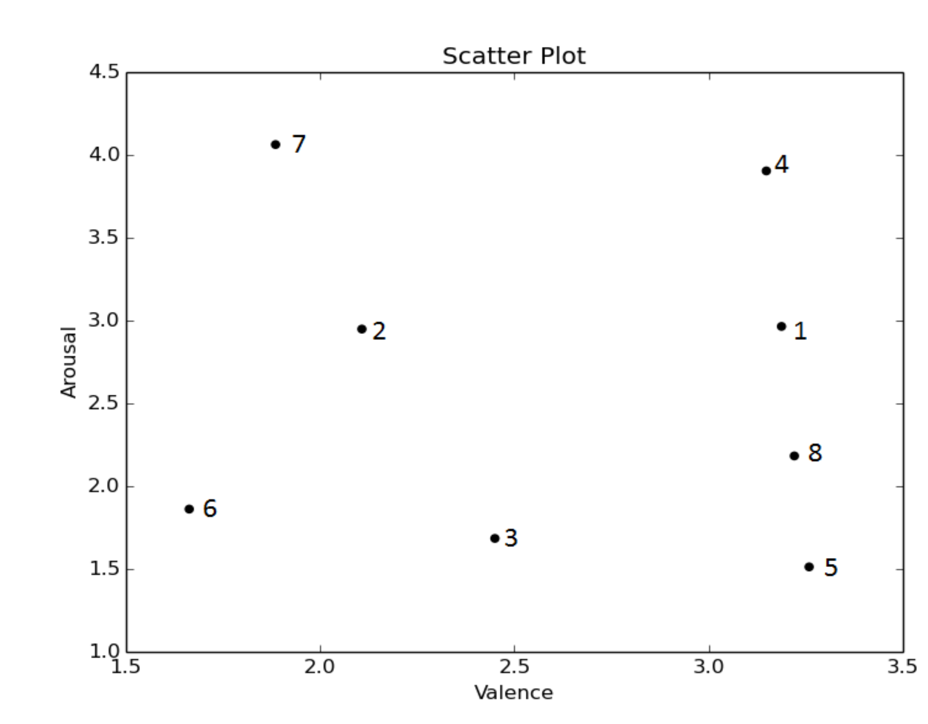
Confusion Table for Labelled Data

	Happy / Excited	Sad /Tense	Upset/Distressed	Calm / Relaxed	Accuracy (%)
Happy / Excited	243	31	49	273	40.77
Sad /Tense	34	214	16	130	54.04
Upset/Distressed	36	9	229	148	54.26
Calm / Relaxed	271	91	180	846	60.95

Analysis of Clustering

- On applying K means on the valence arousal model we get following centroids for the classes.





Confusion Table for No. of clusters = 6

	1	2	3	4	5	6	Accuracy (%)
1	122	318	57	32	10	19	21.86
2	107	718	41	22	20	23	77.12
3	63	263	37	7	4	8	9.68
4	21	188	16	19	8	18	7.03
5	46	181	12	11	15	14	5.37
6	57	264	11	12	7	23	6.14

Confusion Table for No. of clusters = 8

	1	2	3	4	5	6	7	8	Accuracy (%)
1	18	7	79	5	0	3	5	5	14.75
2	15	71	336	7	8	12	3	5	15.53
3	53	81	470	61	15	11	10	19	65.20
4	16	21	175	37	7	5	7	7	13.45
5	17	24	223	7	24	10	6	10	7.47
6	21	42	235	10	12	17	9	8	4.90
7	8	21	190	15	5	4	23	7	8.27
8	7	29	186	12	4	7	3	29	10.46

Result

- We achieved an accuracy of 54.71% on supervised learning on the tagged data.
- We clustered the data on the basis of Valence and Arousal values of the crowd sourced data, for cluster size of 6 and 8. On applying classification on these models, we achieved lesser accuracy of 33.35% and 24.60% respectively.

Conclusion

We tried to extract visual and audio features from videos from LIRIS dataset, which were short excerpts of movie scenes. We managed to extract 11 features and then used the tagging available to create a Machine Learning model. We got an accuracy of 54.71% on the tagged data. We then tried to analyze different models based on different cluster numbers to determine the results we can obtain on the crowd sourced data. We then achieved an accuracy of 33.35% and 24.60% on cluster size of 6 and 8. Our hypothesis on our result is that small cluster size will give us better results since most of the video in the data set belong to calm and happy classes.

Future Work

- Find the goodness of the features.
- Optimise the features extracted and incorporate more features helping us improve the efficiency and accuracy.
- Apply better classification algorithms for the affectie analysis of the data.
- Develop an application based on our project for day to day use and develop systems to use in online video websites and search and retrieval systems.