Scene Intensity Estimation and Ranking for Movie Scenes Through Direct Content Analysis

Saurabh Kataria (12807637)
Department of Electrical Engineering
Indian Institute of Technology, Kanpur
Email: saurabhk@iitk.ac.in

Abhay Kumar (12011)
Department of Electrical Engineering
Indian Institute of Technology, Kanpur
Email: abhayk@iitk.ac.in

***Abstract*—In this project, an approach for scene intensity estimation and subsequently, ranking of movie scenes based on extracted features such as scene length, harmonicity, and motion energy is implemented and experimented with. Such a ranking can be used for automatic trailer generation, movie summarization and characterizing the emotional content of multimedia content. The ranking can be used to learn the intrinsic parameters of a user by processing the ratings he/ she has provided to multimedia contents on website like netflix. This in turn allows us to index and retrieve content according to a given user's profile. A dataset of 3 movies was constructed where the movie was broken into "scenes" manually and top 10 critical scenes were marked. Experimentation of including facial emotion response is also addressed. Results show that a simple combination of audio-visual features, either individually or combined, can fairly reliably be used to predict the intensity of a scene. The validation is done by comparing its performance with the manually annotated ground truth.**

I. Introduction

Scene intensity estimation is a sub-task in automated multimedia content analysis. It is also closely related to emotion prediction in movies. As the name suggests, it aims at estimating the intensity of scenes/ shots in a given multimedia content (for example, commercial movies). Since scene intensity is a subjective quantity, it can be quantified by various low-level multimodal features. One particular application in the field of computational social science and media informatics is that of estimating gender representation in a movie.

A. Motivation

Can we predict how intense a scene is in a movie? Can we improve the existing models for that task? While intensity can be understood as a measure of excitement or activity in a scene, there are several questions that are much harder to answer. (a) Can we come up with a computational model which can well approximate the actual intensity humans feel after watching the scene? (b) What will be the factors required for estimating that? For example, emotional and music intensive scenes will increase the intensity. Based on some psychological findings, several attempts have been made to predict how interesting a video is [1]–[3]. While they have come up with promising results in gender representation estimation task, the work leaves several open questions.

B. State of the Art and Preliminary Work

Rapid growth of video contents online have accelerated the current research in video content analysis based on multimodal features. Several affective content-based video scene extraction schemes have been studied to map low-level features of video data to high-level emotional events. Multiple media modalities including audio and visual cues are exploited in [4] for detection of semantically meaningful scenes from feature films. Hidden Markov Models (HMM) based Video Affective Content and Audio Emotional Event have been explored in [5], [6]. Valence features together with emotion intensity are used for HMM-based emotion type identification in [7]. Detection of attention-invoking audiovisual segments is formulated in [7] on the basis of saliency models for the audio, visual, and textual information conveyed in a video stream. Analyzing a multimedia content at an affective level reveals information that describes its emotional value or scene intensity. Computable video features namely, average shot length, color variance, motion content and lighting key are exploited in [8]. One of the most recent work [3] considers three factors: shot length, motion energy, and harmonicity. Significant exploitation of cinematic principles and video features is

must for robust video content analysis or scene intensity estimation.

The organization of the rest of the paper is as follows. Section II describes the low level features used in direct content analysis and the framework adopted for scene intensity estimation. Section III presents the experimentation and results. Some of the challenges faced are listed in Section IV. A brief conclusion is presented in Section V and Section VI gives the scope of future work in this respect.

II. Direct Content Analysis

The general structure of a movie (feature film) is illustrated in Figure 1. We have assumed this hierarchical structure for all analysis throughout the report.

Fig. 1: General structure of a movie

The features used in the framework for scene intensity estimation are described in the following sub-sections.

A. Scene Length

Scene is a subjectively defined concept that depicts and conveys a high-level story. A scene is collection of semantically related shots, which may or may not be physically close to each other. Thus, we no longer benefit from the tradition causal processing approach where shots are processed sequentially and only the knowledge of the past is available. Generally speaking, a scene contains a collection of semantically related and temporally adjacent shots that have the following three features [9].

- **Visual Similarity** Similar visual contents, such as similar color layout and continuous object activities, could be observed in these shots, especially in movies because of one of the filming techniques called montage.
- **Audio similarity:** The similarity of audio contents is manifested as similar background noise that exists in these shots. In addition, if the same person is talking in several shots, his speeches in all these shots should present similar acoustic characteristics.
- **Time locality** Shots are temporally close to each other if they are within the same scene. For instance, given two shots of the same person, if they are juxtaposed together, they are more likely to be in the same scene than the case where they are placed far apart from each other.

B. Motion Energy

Motion Energy is defined as the measure of motion activity in a particular scene. Optical flow-based motion vectors for a large number of keypoints/interest points between each pair of consecutive frames are computed based on the standard Lucas-Kanade algorithm [10]. An overview of the methodology to compute motion energy is:

Algorithm 1 : Motion Energy Computation

Input : Movie Scenes. Let the number of frames in the scene is S

Keypoints/Interest Points Detection : Detect keypoints (good features to track) [11] using *cv2.goodFeaturesToTrack*. Let the number of keypoints detected is N

Optical Flow Computation : Compute optic flow motion vector (u, v) magnitude for all detected keypoints between each pair of consecutive frames.

Averaging over all frames :

$$\mathcal{M} = \frac{1}{N * S} \sum_{i=1}^{S} \sum_{j=1}^{N} \sqrt{u_i^2 + v_i^2} \tag{1}$$

Fig. 2: Figure illustrating the optic flow of Keypoints detected

C. Harmonicity

Music is characterized by a temporal structure and a note (pitch and overtones) structure. Temporal structure is analyzed based on amplitude statistics. But, spectrum analysis is necessary for better characterisation of music features. For example, a segmentation of musical harmony (chords) can be performed by analyzing the spectrum and retrieving regularities. Typical music consists of a series of chords which are frequently changed. These chords are visible in the spectrum as a group of frequencies simultaneously present for a longer time. We can perform a fundamental frequency determination on the chords as a first step toward note analysis. The sequence of fundamental frequencies in music segment is very important for the human attribution of content. It determines the perception of melody and the structure of a piece of music. The harmonicity H of a signal at time t measures the average periodicity in a long term neighborhood using a window. Pitch is undefined in a non-periodic audio sample. We exploit a standard pitch detection algorithm (aubiopitch) for this purpose. It gives additional information about pitch (p_i) in non-periodic regions with a value of 0 or -1. The absence of a pitch period is indicated by 1. The expression of harmonicity is:

$$\mathcal{H}(t) = \frac{\sum_{i=t/t_s}^{(t+w)/t_s}(1-\delta(p_i+1))}{w/t_s} \tag{2}$$

Basically, it is the ratio of the number of frames that contain a pitch period to the total number of frames in the scene. The details behind the expression can be found in [12].

D. Facial Emotion Estimation

Emotional factors directly reflect audiences attention, evaluation and memory. Affective contents analysis not only create an index for users to access their interested movie segments, but also provide feasible entry for video highlights. Content-based audio and video features can be exploited to classify basic emotions elicited by movie scenes.

We have tried to quantify the emotion factor from the facial expression. Two broad approaches have been widely popular in the literature:

- Emotion Prediction on static face detected.
- Emotion prediction from optical flow of facial keypoints over few consecutive frames

Second approach also account for the temporal variation, which is evident in the movie scenes. We implemented this approach for facial emotion detection. An overview of the methodology [13], [14] for facial emotion detection has been presented in Algorithm 2.

Algorithm 2 : Facial Emotion Detection

Face Detection :

- Used Haar feature-based cascade classifiers
- Adaboost (Final classifier is a weighted sum of these weak classifiers)
- Implement Cascade of Classifiers
- Instead of applying all the 6000(say) features on a window, group the features into different stages of classifiers and apply one-by-one

Training Phase :

- Calculate Optic Flow for all interest points on the detected face.
- Concatenated optic flow values is the feature used for training SVM classifier.

Emotion Detection : Used the trained SVM classifier to predict the emotion for a given frame sequence as Smiling, Angry, Shocked, Neutral.

Output :Decision value for each emotion state $[Smiling, Angry, Shocked, Neutral]$ Sample output would be like $[76.194, 14.624, 1.376, 7.806]$

1) Performance: We have tested the trained classifier for 80 test videos of each emotion state. The confusion probability matrix is shown below. The (i,j) entry of the confusion probability matrix is defined as the ratio of number of samples belonging to class "i" to that have been classifed as class "j".

	Angry	Smiling	Shocked	Neutral
Angry	0.625	0	0	0.375
Smiling	0	0.5	0	0.5
Shocked	0	0	0.875	0.125
Neutral	0.0125	0	0.0125	0.975

TABLE I: Confusion Probability Matrix for the trained SVM emotion classifier

2) Reasons for not incorporating in the framework for scene intensity estimation: The facial emotion detector was pre-trained only on frontal faces due to which inconsistent results were obtained on movie scenes. Because in movies, number of non-frontal faces forms a large proportion. Moreover, many frames don't even have faces

in it. We need profile face optical flow vectors for training and robust face recognition for incorporating it in the framework.

E. Framework for Direct Content Analysis

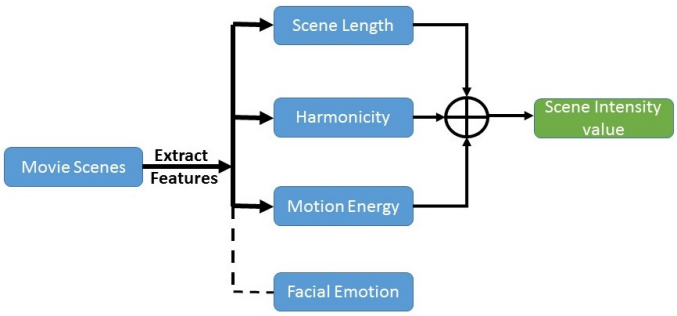

Fig. 3: Figure illustrating framework adopted for Direct Content Analysis. (Note: We wished to incorporate facial emotion in the framework, but it couldn't be incorporated because of the lack of profile view data-set available)

III. Experimentation and results

The problem statement is to devise a scene intensity estimator using features of movie (direct content analysis paradigm) and use it to rank scenes and compare the performance with the ground truth (manual annotations).

A. Dataset description

The dataset constructed for experimental purposes consists of following three feature films: *The Godfather* (1972; Francis Ford Coppola), *The Pursuit of Happyness* (2006; Gabriele Muccino), and *A Beautiful Mind* (2001; Ron Howard). We manually noted the the starting and ending timestamp of all the scenes in these movies by carefully observing them. The number and length of scenes in these movies is slightly subjective. For reducing the bias, we noted the chapter names from the CD covers. The complete list of scenes and their corresponding timestamps of all three movies are provided as separate files in the zip archive.

Total number of scenes in first, second, and third movie were found to be 76, 41, and 45 respectively. A total of 10 scenes were marked as critical for every movie without looking at the performance of our framework. In subsequent subsections, we have reported the performance of our framework in identifying those critical scenes.

B. Platform used

- Video analysis - OpenCV [15], FFmpeg [16]
- Audio analysis - Aubiopitch [17]
- Common tasks - MATLAB, python 3.5.1, 2.8.3
- OS - Ubuntu 14.04

C. Methodology

For each scene of a movie, its scene length, harmonicity, and motion energy is calculated. Then the three vectors (length = number of scenes in movie x 1) corresponding to the above mentioned factors are normalised and added together. Top 10 scenes according to the scores are noted. The size of intersection of this list with the manually marked critical scene list is reported as final score $\in [0, 10]$.

D. Why manual annotation of scenes?

A legitimate question can arise: Why the scenes were marked manually and not by using some software? We tried using the popular shot detection software [18] for identifying the scenes by increasing the threshold but did not achieve desirable results. In Figure 4, it can be observed that the results from the above mentioned approach gave skewed and unacceptable results. In some chapters, there were 0 or 1 scenes, in other there were around 20 scenes, which was incorrect. The reason is that scene boundary decision is an intelligent decision. On the other hand, shot boundary can be estimated by sudden change or fading transition. In Table II, demonstration of poor performance of shot detect code for determining scene boundaries is provided.

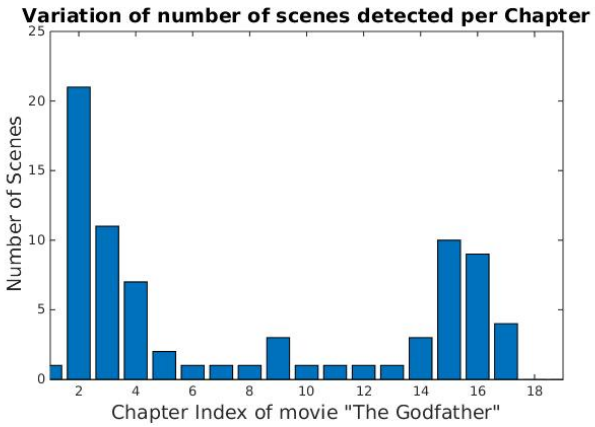

Fig. 4: Number of scenes versus chapters

Threshold for shot detect code	Number of scenes
0.5	350
0.6	208
0.7	89

TABLE II

E. Experiment on motion energy

In the implementation of motion energy, there is a user parameter which specifies the maximum number of keypoints detected for tracking. In Figure 5, average motion energy for all the 45 scenes for movie 3 (*A Beautiful Mind*) is presented as variation in graph. Note that the variation in motion energy is amplified when max_keypt parameter is increased from 500 to 1000. But the relative ordering of these scenes based on the motion energy is still almost the same.

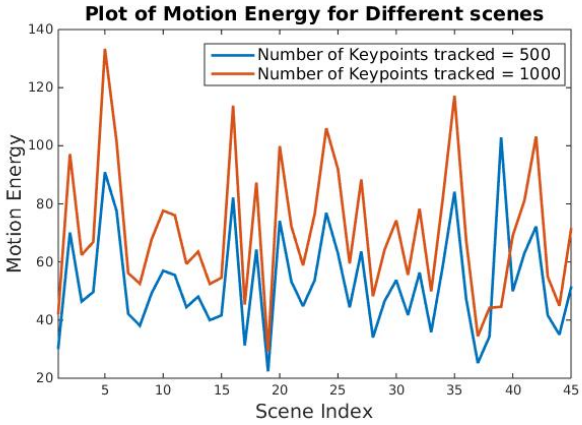

Fig. 5: Figure illustrating variation of motion energy for different scenes for varying total number of keypoints

F. Main results

In Table III, we have tabulated the number of critical scenes (out of 10 for each movie) correctly identified by three parameters individually and then by the combination of those. It can be observed that individually these factors are not giving high performance. But when the output of all the factors are considered, acceptable results are obtained. Approximately, for each movie 5 out of 10 critical scenes are detected by this combination of simple factors.

Parameter used	Movie1	Movie2	Movie3
Scene Length	3	4	1
Harmonicity	3	2	1
Motion Energy	2	0	1
All three parameters	5	5	6

TABLE III: Number of critical scenes identified by different parameters for all movies

In order to reduce the bias in selecting scenes as critical, we have restricted the number to 10 for each movie. We provide 5 critical scenes below for the first movie for reference:

- Vito Corleone is shot
- Sonny is murdered
- Baptism happens in parallel to murders
- Vito's speech to all dons for peace
- Vito sees the dead body of his son Sonny

Brief explanation on results: The actual critical scenes in movie1 were long, had music and motion which is reflected from the results. All three factors are working well individually. In movie2, actual critical scenes had no motion which is reflected from the result that motion energy is not able to identify any critical scene. The surprising result is observed from movie3 where all factors are not good individually but their combination is coming out to be a strong identifier of critical scenes.

IV. Challenges

- One technical difficulty we faced was that ffmpeg was not converting mp4 to wav properly. RIFF header was not getting added because of bitrate issues. So, we used the free software [19].
- Dataset construction (manual annotation) is a tedious task. Since, all the results directly depend on the accuracy with which this construction is done.
- The facial emotion detector was pre-trained only on frontal faces due to which inconsistent results were obtained on movie scenes. Because in movies, number of non-frontal faces forms a large proportion. Moreover, many frames don't even have faces in it.
- Manual annotation can be subjective therefore, number of critical scenes were restricted to 10 only.

V. Conclusion

- Scene boundary has to be decided manually. It can't be done by using the shot detect code by simply tuning the threshold or any other user parameter.
- A combination of simple features of movie like scene length, harmonicity, and motion energy is sufficient to identify approximately 5 out of 10 critical scenes.

VI. Future work

- The analysis presented in this report can be validated further by increasing the number of movies, maybe of different genres too.
- The manual decision of critical scenes in a movie can be done by doing a survey, rather than making personal decision. This step will reduce personal biases.
- If factors such as facial emotion, speech prosody is also considered, then the framework being used in this project can become a very efficient movie summarizer/ trailer maker/ scene intensity estimator.
- The factors being considered in this project for determining the intensity of a scene are currently given a uniform weight. We can assign different weight to those factors to make some inferences.

References

[1] S. Dhar, V. Ordonez, and T. L. Berg, "High level describable attributes for predicting aesthetics and interestingness," in *Computer Vision and Pattern Recognition (CVPR), 2011 IEEE Conference on*. IEEE, 2011, pp. 1657–1664.

[2] T. Schaul, L. Pape, T. Glasmachers, V. Graziano, and J. Schmidhuber, "Coherence progress: A measure of interestingness based on fixed compressors," in *Proceedings of the 4th International Conference on Artificial General Intelligence*, ser. AGI'11. Berlin, Heidelberg: Springer-Verlag, 2011, pp. 21–30. [Online]. Available: http://dl.acm.org/citation.cfm?id=2032873.2032877

[3] T. Guha, C.-W. Huang, N. Kumar, Y. Zhu, and S. S. Narayanan, "Gender representation in cinematic content: A multimodal approach," in *Proceedings of the 2015 ACM on International Conference on Multimodal Interaction*, ser. ICMI '15. New York, NY, USA: ACM, 2015, pp. 31–34. [Online]. Available: http://doi.acm.org/10.1145/2818346.2820778

[4] M. Xu, J. S. Jin, S. Luo, and L. Duan, "Hierarchical movie affective content analysis based on arousal and valence features," in *Proceedings of the 16th ACM International Conference on Multimedia*, ser. MM '08. New York, NY, USA: ACM, 2008, pp. 677–680. [Online]. Available: http://doi.acm.org/10.1145/1459359.1459457

[5] H.-B. Kang, "Affective content detection using hmms," in *Proceedings of the Eleventh ACM International Conference on Multimedia*, ser. MULTIMEDIA '03. New York, NY, USA: ACM, 2003, pp. 259–262. [Online]. Available: http://doi.acm.org/10.1145/957013.957066

[6] M. Xu, L.-T. Chia, and J. Jin, "Affective content analysis in comedy and horror videos by audio emotional event detection," in *Multimedia and Expo, 2005. ICME 2005. IEEE International Conference on*, July 2005, pp. 4 pp.–.

[7] G. Evangelopoulos, A. Zlatintsi, A. Potamianos, P. Maragos, K. Rapantzikos, G. Skoumas, and Y. Avrithis, "Multimodal saliency and fusion for movie summarization based on aural, visual, and textual attention," *Multimedia, IEEE Transactions on*, vol. 15, no. 7, pp. 1553–1568, Nov 2013.

[8] Z. Rasheed, Y. Sheikh, and M. Shah, "On the use of computable features for film classification," *Circuits and Systems for Video Technology, IEEE Transactions on*, vol. 15, no. 1, pp. 52–64, Jan 2005.

[9] Y. Li and C.-C. J. Kuo, "A robust video scene extraction approach to movie content abstraction," *International Journal of Imaging Systems and Technology*, vol. 13, no. 5, pp. 236–244, 2003. [Online]. Available: http://dx.doi.org/10.1002/ima.10063

[10] J.-Y. Bouguet, "Pyramidal implementation of the affine lucas kanade feature tracker description of the algorithm."

[11] J. Shi and C. Tomasi, "Good features to track," in *Computer Vision and Pattern Recognition, 1994. Proceedings CVPR'94., 1994 IEEE Computer Society Conference on*. IEEE, 1994, pp. 593–600.

[12] T. Guha, N. Kumar, S. S. Narayanan, and S. L. Smith, "Computationally deconstructing movie narratives: an informatics approach," in *Acoustics, Speech and Signal Processing (ICASSP), 2015 IEEE International Conference on*. IEEE, 2015, pp. 2264–2268.

[13] A. Lowhur and M. C. Chuah, "Dense optical flow based emotion recognition classifier," in *Mobile Ad Hoc and Sensor Systems (MASS), 2015 IEEE 12th International Conference on*. IEEE, 2015, pp. 573–578.

[14] Emotion-recognition-dof. [Online]. Available: https://github.com/vanstorm9/Emotion-Recognition-DOF

[15] Opencv. [Online]. Available: http://opencv.org/

[16] Ffmpeg. [Online]. Available: https://www.ffmpeg.org/

[17] Aubiopitch. [Online]. Available: http://aubio.org/

[18] T. Guha. (2014) Shot detect. [Online]. Available: http://tmguha.blogspot.in/2014/03/shot-detection.html

[19] Free mp4 to wav converter. [Online]. Available: http://download.cnet.com/Free-MP4-to-WAV-Converter/3000-2140_4 – 76169127.*html*