Scene intensity estimation and ranking in movies

EE698M course project presentation

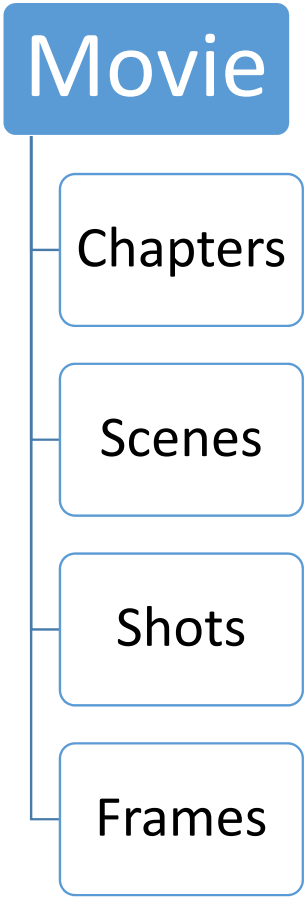
Saurabh Kataria (12807637); Abhay Kumar (12011)

Break-up of title

- Direct content analysis: Source of information for analysis is restricted to the media content only (here, movie). No auxiliary information from world is used.
- Scene intensity estimation: Every scene of the movie can be assigned an intensity value based on how “important” that scene is. The value can depends on the factors like emotion, music, etc.
- Movies/ feature films: Unlike documentaries or art films, movies have a particular structure and hence, more suitable for analysis

General structure of a movie

- Movie tells a story through a defined path
- We are interested in finding how “intense” each “scene” is.



Problem statement

- As hinted earlier, we are interested in assigning a intensity value to each scene of a movie.
- Implement a scene intensity estimator and rank scenes as per that.
- See how it performs when compared to ground truth.
- Ground truth refers to the scenes marked as “critical” by human

Motivation for problem

- Identifying important scenes can help in automatic content analysis on large scale.
- Possible extensions to machine vision, video surveillance, etc.
- One particular application in film analysis – gender estimation, where an objective understanding of gender portrayals and biases is studied [1]
- Learning the intrinsic interests of a viewer by analyzing the movie he/she has watched. Scope for making better recommender systems.

Methodology used

- Step 1: Dataset construction

Manually detected the timings of scenes in a movie, and segmented the movie

- Step 2: Calculating following three values for each scene

Scene length, Harmonicity value, Motion energy value

- Step 3: Combine the above results

Theory

- Harmonicity
  - Measures the average periodicity in a neighborhood using a window
  - Harmonicity = number of pitch frames/ total frames in that window, as expressed follows

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

- Motion energy
  - Measures motion activity averaged over consecutive frames
  - Detect keypoints (good features to track) [Tomasi (7)]
  - Lucas-Kanade optical flow algorithm Took average of absolute optic flow values for all interest points

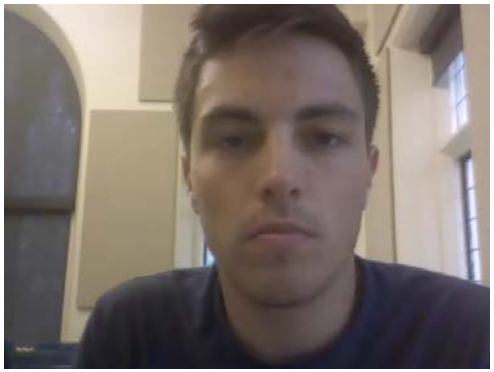
Demo of motion detection

Facial emotion detection [6]

- Face Detection
  - Haar feature-based cascade classifiers
  - Adaboost (Final classifier is a weighted sum of these weak classifiers)
  - 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
- Emotion detection
  - Calculate Optic Flow for all interest points on the detected face.
  - Concatenated optic flow values is the feature used.
  - Used the trained SVM classifier to predict the emotion as ‘Smiling’, ‘Angry’, ‘Shocked’, ‘Neutral’.

Probability values for different emotions [6]



NEUTRAL	SHOCKED	ANGRY	SMILING
Neutral: 73.376	Neutral: 5.854	Neutral: 23.955	Neutral: 2.56
Smiling: 16.205	Smiling: 7.825	Smiling: 24.09	Smiling: 87.241
Shocked: 1.599	Shocked: 81.692	Shocked: 1.449	Shocked: 9.024
Angry: 8.819	Angry: 4.629	Angry: 50.507	Angry: 1.175

Platform and dataset description

- Movie1 - The Godfather (1972)
- Movie2 - The Pursuit of Happyness (2006)
- Number of critical scenes per movie (hand labelled) = 10
- Video analysis – FFmpeg library, OpenCV
- Audio analysis – Aubio library

Screenshots of dataset

MOVIE1; number of actual scenes = 75

1. I believe in America * 00:01:20
   - i) Man_1 requests Vito to give his daughter justice by killing her boyfriend * 00:04:08
   - ii) Vito expresses dissent but agrees afterwards * 00:07:31
2. The wedding
   - i) Photography session * 00:08:30
   - ii) Dance session; sonny breaks camera * 00:11:48
   - iii) Nazorine talks with Vito * 00:12:49
   - iv) Michael arrives * 00:15:00
   - v) Luca Brasi meets Vito * 00:15:51
   - vi) Sonny talks to women in pink dress and take her upstairs * 00:18:06
3. Johnny Fontane
   - i) Johnny arrives and sings * 00:22:14
   - ii) Michael introduces his brother Fredo to Kay Adams * 00:22:37
   - iii) Johnny describes his problem with producer to Vito * 00:27:04
4. Tom Hagen goes to Hollywood
   - i) Tom Hagen lands and talks to Woltz * 00:29:35
   - ii) Tom has dinner with Woltz * 00:32:39
   - iii) Woltz finds head of a horse in his bed * 00:34:18
5. Meeting with Sollozo
   - i) Vito holds a pre meeting * 00:36:12
   - ii) Failed meeting * 00:40:18
   - iii) Luca Brasi goes to Bruno Tattaglia * 00:44:36
6. Shooting of Don Corleone
   - i) Vito is shot * 00:46:16
   - ii) Michael gets the news * 00:49:00

MOVIE2 ; number of actual scenes = 41

1. Riding the bus * 00:03:36
   - i) Going to school * 00:05:09
   - ii) Trying to sell the bone density scanner * 00:07:11
   - iii) Eating dinner * 00:08:19
   - iv) Solving Rubik's cube * 00:09:22
2. Two questions
   - i) Talking to the stockbroker * 00:10:28
   - ii) Heated argument with Linda * 00:12:36
3. Being stupid
   - i) Chasing the girl for scanner * 00:15:27
   - ii) Sending his kid to day care * 00:18:58
4. Making contacts
   - i) Handing over the application for trainee * 00:20:11
   - ii) Running after the thief * 00:22:13
   - iii) At home with Linda and his son * 00:24:22
5. Ride share
   - i) Ride with the person_1 * 00:28:20
6. Cab fare
   - i) Running away from the driver * 00:30:34
7. Go get happy
   - i) Family leaves * 00:32:30
   - ii) Got the number * 00:34:43
   - iii) Chris gets Christopher back * 00:36:51

Why manual annotation of movie scenes?

- Tried to get scenes from automatic shot detection code by increasing threshold but that gave undesirable results as tabulated.

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

- 0.7 threshold gave number of scenes close to actual (which is 75)
- But, it had other problems such as 1) Chapter combined into single scene, 2) Actual scene segmented into unnecessary number of scenes, 3) Sometimes chapter got combined into one scene!

Screenshot of poor results for scene detection via shot detect scheme

Chapter name	number of scenes detected	scene indices
1. I believe in America	1*	1**
2. The wedding	21	2-22
3. Johnny Fontane	11	23-33
4. Tom Hagen goes to Hollywood	7	34-40
5. Meeting with Sollozo	2	41-42
6. Shooting of Don Corleone	1	43
7. Luca Brasi Sleeps with the fishes	1	44
8. Michael at the hospital	1	45
9. It's strictly business	3	46-48
10. How's the Italian food in this restaurant	1	49
11. The don returns home	1	50
12. The thunderbolt	1	51
13. Sonny gives Carlo a warning	1	52
14. Michael marries Apollonia	3	53-55
15. I don't want his mother to see him this way	10	56-65
16. Apollonia's murder	9	66-74
17. We are all reasonable men here	4	75-78
18. The don puts Michael in charge	"	"
19. I'm Moe green	"	"

Results

- Movie1

Parameter used	Number of critical scenes identified (out of 10)
Scene length	3
Harmonicity	3
Motion energy	2
Scene length + Harmonicity + Motion energy	5

- Movie2

Parameter used	Number of critical scenes identified (out of 10)
Scene length	4
Harmonicity	2
Motion energy	0
Scene length + Harmonicity + Motion energy	5

Results (continued; added later)

- Movie3

Parameter used	Number of critical scenes identified (out of 10)
Scene length	1
Harmonicity	1
Motion energy	1
Scene length + Harmonicity + Motion energy	6

Comments and observations from results

- Combining all three factors surely gives improvement over using one of them.
- The critical scenes in Movie1 were long, had music in it, and action too. That’s why all three factors perform good individually
- The critical scenes in Movie2 were mostly silent and didn’t have much action. That’s why these factors doesn’t perform well individually.

Challenges

- FFmpeg does not convert mp4 to wav (s16le) properly. The RIFF header of wav is missing. So, used a windows software “free mp4 to wav converter” [5].
- Dataset construction is a tedious task
- Inconsistency in facial emotion detection due to lack of frontal face shots in the pre-trained model
- Labelling is subjective

Possible extensions

- Increase the dataset, currently working on third movie.
- Establish ground truth via a survey
- Account for features such as emotion from audio (speech prosody), emotion from text/subtitle (NLP), effective facial emotion extraction

References

1) Gender representation in cinematic content: A multimodal approach, **T Guha**, C Huang, N Kumar, Y Zhu, and S S Narayanan, ICMI 2015
2) Computationally deconstructing movie narratives: An informatics approach, **T Guha**, N Kumar, S S Narayanan and S Smith, ICASSP 2015
3) ffmpeg.org
4) aubio.org
5) http://download.cnet.com/Free-MP4-to-WAV-Converter/3000-2140_4-76169127.html
6) https://github.com/vanstorm9/Emotion-Recognition-DOF
7) www.inf.fu-berlin.de/lehre/SS06/.../origReport_von_Carlo_**Tomasi**.pdf

Thank you.