COMPUTATIONALLY DECONSTRUCTING MOVIE NARRATIVES: AN INFORMATICS APPROACH

Tanaya Guha[1], *Naveen Kumar*[1], *Shrikanth S. Narayanan*[1], *Stacy L. Smith*[2]

[1]Signal Analysis and Interpretation Lab, University of Southern California, Los Angeles, CA
[2]Media Diversity and Social Change Initiative, University of Southern California, Los Angeles, CA

ABSTRACT

In general, popular films and screenplays follow a well defined storytelling paradigm that comprises three essential segments or acts: exposition (act I), conflict (act II) and resolution (act III). Deconstructing a movie into its narrative units can enrich semantic understanding of movies, and help in movie summarization, navigation and detection of the key events. A multimodal framework for detecting such *three act narrative structure* is developed in this paper. Various low-level features are designed and extracted from video, audio and text channels of a movie so as to capture the pace and excitement of the movie's narrative. Information from the three modalities is combined to compute a continuous dynamic measure of the movie's narrative flow, referred to as the *story intensity* of the movie in this paper. Guided by the knowledge of film grammar, the act boundaries are detected, and compared against annotations collected from human experts. Promising results are demonstrated for nine full-length Hollywood feature films of various genres.

Index Terms— Informatics, multimedia, movie narratives, story intensity, three act structure.

1. INTRODUCTION

With the proliferation of digital media, automated analysis of media content, especially movies, has become critical for abstracting, indexing and navigating through the vast amount of data. Informatics-driven approaches are also enabling new ways to understand media content and personalize interaction experiences.

Significant research effort has been put toward recognizing and extracting semantically meaningful structures, such as shots and scenes in movies. A large number of automated methods have been developed that can detect shot and scene boundaries with reasonable accuracy [1, 2]. A shot is defined as a contiguously recorded sequence of frames and a scene is a higher level structure that is composed of a series of shots [1]. Relatively little effort however has been made to extract higher level structures such as the *narrative structure* of a film.

In general, popular films and screenplays follow a well defined storytelling paradigm that comprises three essential segments or acts: *Act I (exposition)*, *Act II (conflict)* and *Act III (resolution)* [3, 4] (see Fig. 1). Act I introduces the main characters in a movie, and presents an incident (plot point 1) that drives the story; this leads to a series of events in Act II including a key event (plot point 2) that prepares audience for the climax. Act III features the climax and the resolution of the story. According to film theory, Act I ends around the 25-30 minute mark of a film, and Act II ends about 25-30 min before the end of the film [3, 4]. In practice, there is significant variability in the realization of the three act structure.

Detecting the act boundaries can enrich the semantic understanding of movies in several ways. Knowing the locations of the act boundaries makes detecting key events (necessary for summarization and visualization) in a movie much easier since key events, such as the plot points or climax, usually occur in close proximity to the act boundaries. To the film community and experts, the three act structure provides an important basis for comparing different movies and evaluating relative importance of the characters [3, 4]. An automated system that is able to provide an estimate of the act boundaries can assist in further critical analysis of the narrative structure and form.

The importance of the three-act narrative structure has been mentioned in several papers [5, 6, 7], but few have investigated how such structure may be automatically extracted from movie content. To the best of our knowledge, we are aware of only one work that has addressed this problem [8]. This work relies on constituting two separate feature vectors called visual and audio tempo curves [8]. Following a supervised learning paradigm, this method learns the likelihood of locating an act boundary at a high visual tempo and a high audio tempo point. The problem with such a supervised approach is that it requires a large amount of labeled data to account for the large variability inherent in the process of movie making. Creation of such a large database is difficult because accurate manual detection of act boundaries is time consuming (even for the experts), and requires considerable knowledge of film theory and grammar. Hence, a natural choice is to build an unsupervised computational framework.

In this paper, we address the problem of automatically detecting the three act narrative structure in movies in an unsupervised manner. Our goal is to segment a given movie into its three constituent acts by detecting the two act boundaries as shown by dotted lines in Fig. 1. In a case like this where labeled data is challenging to acquire for supervised learning, we believe features play a crit-

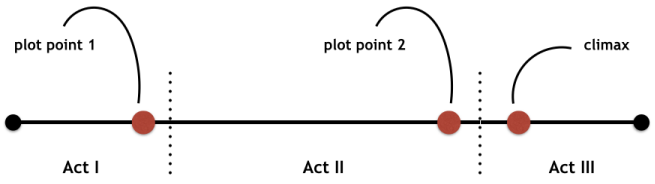

Fig. 1. A schematic diagram of the three act narrative structure in a movie. Our objective is to detect the boundaries between Act I and II (Act Boundary I), and ActII and Act III (Act Boundary II) are the key anchors of a movie's narrative structure.

Many thanks to M. Choueiti and T. Wang from USC Annenberg for the act boundary annotations.

ical role. Our effort is hence focused on carefully designing and extracting features that can capture the cues of transition between acts (or in general, any long-term events or episodes). We extract features from three modalities, video, audio (music) and text (subtitles), and combine them to a continuous dynamic measure of the pace and excitement in a movie, referred to as the *story intensity* in our work. We use this measure of story intensity as a computational index of the (latent) narrative flow, and to mark salient points therein. Guided by film grammar, the act-boundaries are detected from the story intensity curve in an unsupervised manner. Experimental boundary detection results are presented for nine full-length Hollywood movies. For reference, we collected annotations from human experts for these movies. Results of our automated methods are promising when compared against these reference annotations.

2. PROPOSED APPROACH

The objective of this work is to automatically segment a movie into its three acts by detecting the act boundaries. Since we are dealing with overall narrative structure and flow, the usual assumptions held in automated movie content analysis (e.g. in scene detection, frames belonging to the same scene are assumed to have similar illumination, color and audio environment) are not necessarily valid. Hence, it is important to design suitable low-level features that are indicative of the narrative flow. To aid that, we leverage knowledge of techniques that are used in the film-making process to communicate transition between acts.

2.1. Feature design

Visual features: Two visual features that are known to reflect the story intensity of a movie are *shot length* and *motion activity* [8]. Frequent cuts (short shots) and fast motion of actors, objects or camera are often used to create a notion of speed at which the story unfolds, e.g., in the action scenes of a movie.

A) Shot length: Given a movie, we employ a shot-detection method using an open source tool called *ffmpeg* that identifies the key-frame in a movie. This method detects the key frames based on computing pixel-by-pixel differences between consecutive frames. The movie is then partitioned accordingly into N shots. The length of each shot is defined by the number of frames it contains, and is used to construct a feature vector $\mathcal{V}_s = [s_1, s_2, ..., s_N]$ where s_i is the number of frames in the i^{th} shot and N is the total number of shots detected in a movie.

B) Motion activity: To compute a measure of motion activity, optical flow-based motion vectors between each pair of consecutive frames are computed based on the standard Lucas-Kanade algorithm [9]. Flow vectors are computed at a large number of keypoints obtained using a corner detection algorithm [10]. Let us consider a pair of consecutive video frames: F_1 and F_2. Let the number of keypoints detected in F_1 be k. The motion vector associated with the j^{th} keypoint be $[u_j, v_j]$ denoting the movement in horizontal and vertical directions. The motion activity in F_2 is computed as the total motion F_2 has undergone with reference to F_1. Motion activity m_i at shot level is measured as the average motion energy per frame in that shot. A motion activity feature, $\mathcal{V}_m = [m_1, m_2, ..., m_N]$, is computed where m_i is the motion activity for shot i and is computed as follows:

$$m_i = \frac{1}{s_i} \sum_{n=1}^{s_i - 1} \sum_{j=1}^{k} (u_j^2 + v_j^2) \tag{1}$$

The features $\mathcal{V}_s$ and $\mathcal{V}_m$ are mean removed and normalized to have unit standard deviation. They are combined linearly to obtain a single visual feature $\mathcal{V}$ as follows:

$$\mathcal{V}(i) = -w\mathcal{V}_s(i) + w\mathcal{V}_m(i) \tag{2}$$

where, i is the shot index, and $w = 0.5$. We assume equal contribution of the two visual features, $\mathcal{V}_s$ is given a negative value since shot length is assumed to have a negative relationship with movie intensity i.e. smaller shot length indicates higher speed.

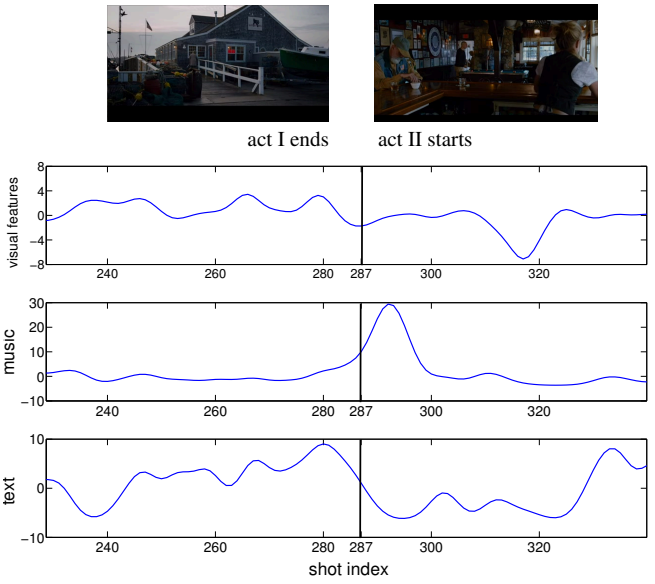

Fig. 2. Behavior of the three modalities around the *act boundary I* (shot # 287) for the movie *Hope Springs*. Here, music and text show strong indications of act transition and for visual features, the cue comes (edge at shot # 284) slightly before the actual act transition.

Music feature: The presence of music plays a very important role in setting different moods in a movie [11]. We believe that its role is particularly significant at act boundaries where it is used as a device to indicate a change in a movie's storyline.

In this work, we extract a feature called *harmonicity* [12] that detects the presence of music in a movie's audio stream. The harmonicity $\mathcal{H}$ of a signal at time t measures the average periodicity in a long term neighborhood using a window. We exploit a standard pitch detection algorithm for this purpose. Pitch is undefined in a non-periodic audio sample, and hence, most of the standard pitch detectors provide this additional information indicating the pitch in non-periodic regions with a value of 0 or -1. We use the *aubiopitch* [13] tool for pitch tracking to obtain pitch values $p_1, p_2, \ldots$ with a time step of $t_s = 0.05s$. The absence of a pitch period is indicated by -1. Using these pitch values, we compute harmonicity values in a window of duration $\mathtt{w} = 3s$, by finding the ratio of the number of frames that contain a pitch period to the total number of frames within that duration as shown in Eqn.(3). This feature is computed with a time-shift of $0.5s$.

$$\mathcal{H}(t) = \frac{\sum_{i=t/t_s}^{(t+\mathtt{w})/t_s} (1 - \delta(p_i + 1))}{\mathtt{w}/t_s} \tag{3}$$

Text feature: We hypothesize that the rate of speech i.e. the *dialog delivery rate* may be relevant to the story, and the pace at which a

movie progresses. In particular, it might be useful in certain genres such as drama or romance, where dialogs are of heavy importance.

We propose to compute the dialog delivery rate, as a feature using the time-aligned subtitles (sub rip text (srt) format). Subtitles in srt format contain subtitle index, subtitle text, and start and end time stamps. Other than dialogs, subtitles also contain narration and song lyrics in movies, marked separately using an italics tag. First, we clean the subtitles to keep only the dialogs. The delivery rate for each subtitle is computed as the number of words uttered per second. To estimate the dialog delivery rate $\mathcal{R}(t)$ at time $\mathtt{t}$, the mean delivery rate of all subtitles within the interval $[\mathtt{t}, \mathtt{t}+\mathtt{w}]$ is computed. The feature is computed with a time shift of $0.5s$.

$$\mathcal{R}(t) = \frac{\text{\# words in subtitle at time } t}{\text{duration in seconds of subtitle at time } t} \tag{4}$$

2.2. Multimodal Fusion

The features extracted from the video channel are computed at shot-level but the audio and text features are computed at different time scales. First, we compute audio and text features at shot-level by averaging the feature values over the length of a given shot. Thus all features are N-dimensional where N is the total number of shots in a movie. Fig. 2 presents the behavior of individual modality around an act boundary using the movie Hope Springs as an example. In order to combine information, we fuse the three modalities into a so-called *story intensity* curve that captures the varying excitement or intensity of the movie narrative as follows:

$$\mathcal{P}(i) = w_v \frac{v_i - \mu_v}{\sigma_v} + w_h \frac{h_i - \mu_h}{\sigma_h} + w_r \frac{r_i - \mu_r}{\sigma_r} \tag{5}$$

where μ_v, μ_h, μ_r and $\sigma_v, \sigma_h, \sigma_r$ are the means and standard deviations of $\mathcal{V}$, $\mathcal{H}$ and $\mathcal{R}$. and i is the shot index. The weights w_v, w_h and w_r control the contribution of each modality towards the definition of story intensity.

Ideally, these weights should be learnt from a set of training data. However, in the absence of enough labeled data we devise a semi-supervised technique to learn the weights. First, we assume equal contribution of each modality and set $w_v = w_h = w_r = 1/3$. Act boundaries are then detected for each of the nine movies using the method described in Section 2.3. Results show that the detected act boundaries for the movie *Cabin in the Woods* are the closest to experts' annotations. Hence, we select this intensity curve as our *story intensity template*, $\mathcal{T}$, assuming that this curve has the 'ideal' structure around the act boundaries. The template $\mathcal{T}$ is subject to cubic spline fitting, and then it is resampled to have its length equal to $\mathcal{P}$. The weights are then learnt by maximizing the correlation between the template and the story-intensity curve.

$$\mathbf{w} = \underset{\mathbf{w}}{argmin}\ (1 - \text{abs}(\text{corr}(\hat{\mathcal{T}}, \hat{\mathcal{P}}))) \tag{6}$$

where $\mathbf{w} = \{w_v, w_h, w_r\}$. Note that $\hat{\mathcal{T}}$ and $\hat{\mathcal{P}}$ denote only those parts of the movie where act boundaries are expected to occur i.e. in between 20 - 30 minute mark in the beginning, and in between 20 - 30 minute mark before the movie ends [3, 4].

2.3. Act boundary detection

We pose the act boundary detection problem as a problem of one-dimensional edge detection in the story intensity curve $\mathcal{P}$. Before edge detection, $\mathcal{P}$ is subject to Gaussian smoothing. The proposed boundary detection steps are as follows:

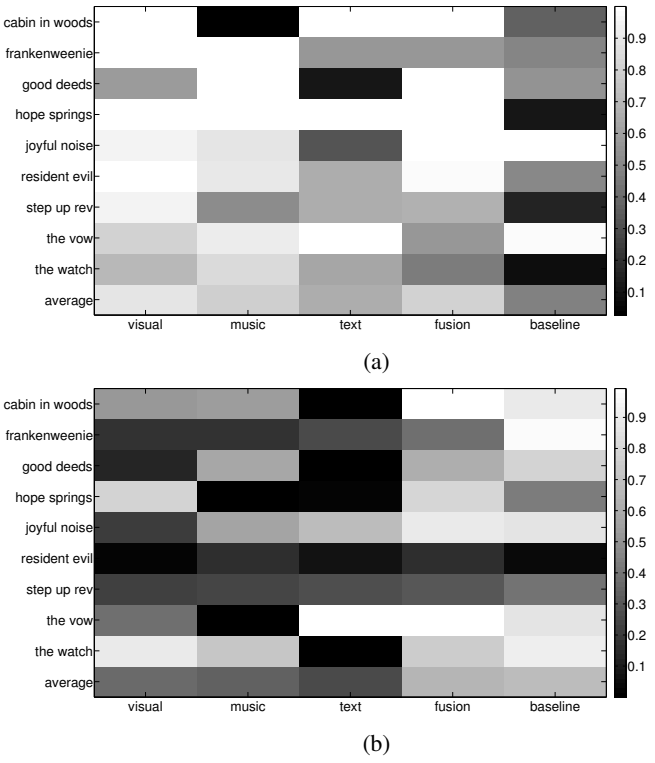

Fig. 3. Detection results for (a) *Act Boundary I* and (b) *Act Boundary II* for the individual modalities as compared to the baseline. Higher brightness indicates higher accuracy.

(i) A Laplacian of Gaussian (LoG) filter is convolved with $\mathcal{P}$ and the zero-crossing points (since they correspond to the edges) are detected in the resulting signal. Note that any standard edge detection method is expected to produce similar end results.

(ii) The zero-crossing points are considered to be the *candidate boundary points*: $\{b_1, b_2, \ldots b_K\}$ where K is the number of zero-crossing points.

(iii) We select an act boundary from the set of all candidate boundary points. A Gaussian probability distribution $\mathcal{N}(\mu, \sigma)$ is used to compute the likelihood of each candidate point to be an act boundary.

$$P(b_j = \text{act boundary}) \propto e^{\frac{-(b_j - \mu)^2}{2\sigma^2}} \tag{7}$$

The candidate boundary point b_j with the maximum P is chosen to be the act boundary. With symmetric pdf, like Gaussian, this is similar to choosing the b_i closest to the 25^{th} minute mark. Nevertheless, any skewed pdf can be used, preferably justified by domain knowledge.

(iv) For *act boundary I*, $\mathcal{N}$ is centered at the 25^{th} minute mark (μ) from the start of the movie with a standard deviation (σ) of 5 minutes. These time intervals are selected based on the knowledge of film grammar [3, 4]. The *act boundary II* is detected in a similar manner with the decaying prior centered at the 25^{th} minute mark from the end of the movie and has the same value of σ.

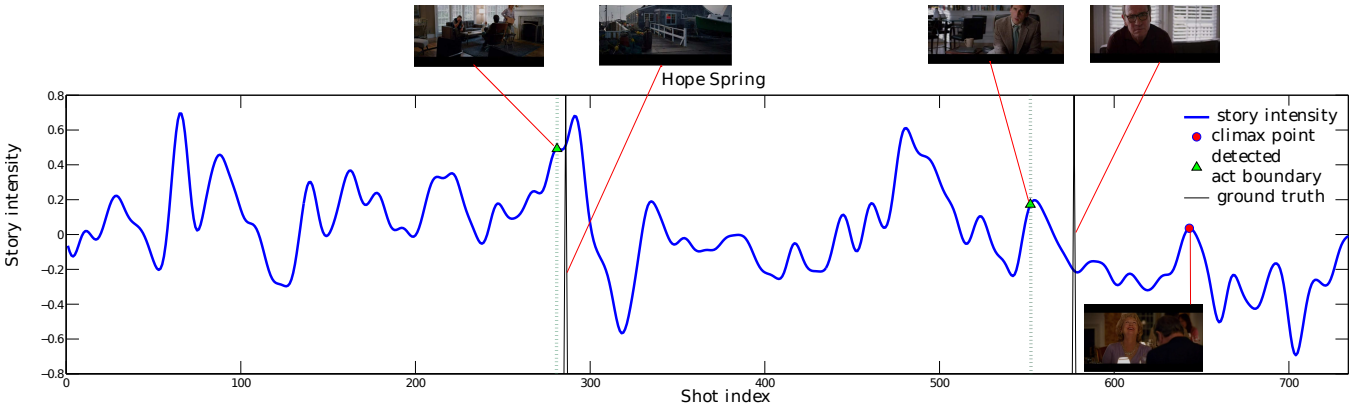

Fig. 4. Example of three act structure detection for the movie Hope Springs where the detected boundaries are shown along with the ground truth obtained from human experts. For act boundary I, error is 76 seconds, and for act boundary II, the error is 205 seconds.

3. EXPERIMENTAL VALIDATION

3.1. Dataset and annotations

We perform act boundary detection experiments on nine full length Hollywood movies released in the year 2012. These movies are: *Cabin in the Woods*, *Frankenweenie*, *Good Deeds*, *Hope Springs*, *Joyful Noise*, *Resident Evil Retribution*, *Step Up Revolution*, *The Vow* and *The Watch*. These movies span various genre including children, drama, action and romance, and were selected by our film expert coauthor SS as the movies that have well-defined three act narrative structure.

Accurate detection of act boundaries requires considerable knowledge of film theory and structure. The reference annotations were obtained from three film experts who watched and marked the act boundaries for each of the nine movies independently. Then, they decided on a final time -stamp for the act boundaries through consensus-seeking discussions. Each movie is annotated with two act boundaries where each act boundary has a single time-stamp precise to the level of seconds.

3.2. Results and discussion

In order to evaluate the performance of the proposed method, we compare a detected act boundary, b_d, with the corresponding reference annotation, b_e, obtained from human experts. A measure of performance i.e. closeness to the human result is obtained in terms of a quantity $\mathcal{Q}$ which is given by $\mathcal{Q} = e^{\frac{-d^2}{2\sigma^2}}$ where $d = |b_e - b_d|$ is measured in seconds; σ has the same value as used in the boundary detection step. The quantity $\mathcal{Q}$ has a maximum value of 1 when b_e coincides with b_d.

Our boundary detection results are validated against the ground truth, and presented in terms of $\mathcal{Q}$ values in Fig. 3 . We also compare the performances of the individual modalities to understand each of their contribution in detecting an act transition. A *baseline* method is used for comparison. The baseline sets the first act boundary at 25^{th} minute mark of the movie, and the second act boundary is at at 25^{th} minute mark from the end of the movie. Results presented in Fig. 3 show that in general, detection of *act boundary I* is more accurate compared to that of *act boundary II*. The visual modality is the most efficient one in detecting *act boundary I*, music features also perform well while text feature seems to be the weakest of all. Clearly, for *act boundary I*, results are better than the baseline in all cases. In fact, to detect *act boundary I*, only one modality, either video or music, seems to be sufficient.

On the other hand, for *act boundary II*, individual performance of each modality is poor. Information from all modalities needs to be fused to improve detection results. Multimodal detection results are better than the baseline for 5 out of the 9 movies, although on average, it performs similar to the baseline. This may be attributed to the fact that movies usually get more complicated towards the end, and hence detection of act boundary II is more difficult. It has also been noted by the human annotators that variability in opinion is much higher in the case of determining *act boundary II* in movies.

Climax detection: Fig. 4 shows the detected and true act boundaries for the movie *Hope Springs*. We have also automatically detected and marked the climax point in the story intensity curve. The climax is detected simply as the highest peak in the third act. This serves as an example of how the three act structure can aid key event detection - a long standing goal in automated movie analysis.

4. CONCLUSION

In this paper, we discussed the importance of the three-act narrative structure in semantic understanding of movies, and developed a multimodal framework to automatically detect the act boundaries. We also demonstrated that the act boundaries serve as a reference to detect key events like the climax scene. Our method relies on computing a continuous dynamic measure of story intensity of a movie. This dynamic measure is computed using a set of low level features extracted from video, music and subtitle text of the movie. The features are designed to capture the transition between acts. To aid this, we leverage knowledge of the techniques used in the film-making process to indicate act transitions.

Computationally detected three act structures are validated against manual annotations collected from domain experts. Highly accurate results are obtained for act boundary I while the accurate detection of act boundary II seems more challenging.

Given the nature of this problem, exploiting multimodal information is critical because filmmakers use different channels to effectively narrate a story to the audience. Designing new features from speech and language can be useful. The proposed story intensity measure can also find applications in genre identification, emotion prediction, and key scene detection in movies.

5. REFERENCES

[1] Jeroen Vendrig and Marcel Worring, "Systematic evaluation of logical story unit segmentation," *IEEE Trans. Multimedia*, vol. 4, no. 4, pp. 492–499, 2002.

[2] Costas Cotsaces, Nikos Nikolaidis, and Ioannis Pitas, "Video shot detection and condensed representation. a review," *IEEE Signal Processing Magazine*, vol. 23, no. 2, pp. 28–37, 2006.

[3] Syd Field, *Screenplay: The foundations of screenwriting*, Random House LLC, 2007.

[4] Robert McKee, *Substance, Structure, Style, and the Principles of Screenwriting*, New York: HarperCollins, 1997.

[5] Ying Li, Wei Ming, and CC Jay Kuo, "Semantic video content abstraction based on multiple cues.," in *Proc. ICME*, 2001.

[6] Zhicheng Zhao and Xiaojuan Ge, "A computable structure model for hollywood film," in *Proc. ICIP*, 2010, pp. 877–880.

[7] Brett Adams, Svetha Venketesh, Hung H Bui, and Chitra Dorai, "A probabilistic framework for extracting narrative act boundaries and semantics in motion pictures," *Multimedia tools and applications*, vol. 27, no. 2, pp. 195–213, 2005.

[8] Brett Adams, Chitra Dorai, and Svetha Venkatesh, "Toward automatic extraction of expressive elements from motion pictures: Tempo," *IEEE Trans. Multimedia*, vol. 4, no. 4, pp. 472–481, 2002.

[9] Bruce D Lucas, Takeo Kanade, et al., "An iterative image registration technique with an application to stereo vision.," in *Proc. IJCAI*, 1981, vol. 81, pp. 674–679.

[10] Jianbo Shi and Carlo Tomasi, "Good features to track," in *Proc. CVPR*, 1994, pp. 593–600.

[11] Kathryn Kalinak, "Settling the score," *Music and the Classical Hollywood Film. Madison*, 1992.

[12] Theodoros Giannakopoulos, Aggelos Pikrakis, and Sergios Theodoridis, "Music tracking in audio streams from movies," in *Proc. IEEE MMSP*, 2008, pp. 950–955.

[13] Paul M Brossier, "The aubio library at mirex 2006," *MIREX 2006*, p. 1, 2006.