INDEXING NARRATIVE STRUCTURE AND SEMANTICS IN MOTION PICTURES WITH A PROBABILISTIC FRAMEWORK

Brett Adams†, Chitra Dorai†, Svetla Venkatesh†, Hung H. Bui†

Department of Computer Science†
Curtin University of Technology
GPO Box U1987, Perth, 6845, W. Australia
{adamsb, svetha, buihh}@cs.curtin.edu.au

IBM T. J. Watson Research Center†
P.O. Box 704, Yorktown Heights
New York 10598, USA
dorai@watson.ibm.com

ABSTRACT
This work constitutes the first attempt to extract an important narrative structure, the 3-Act story telling paradigm, in film. This narrative structure is prevalent in the domain of film as it forms the foundation and framework in which the film can be made to function as an effective tool for story telling, and its extraction is a vital step in automatic content management for film data. A novel act boundary likelihood function for Act 1 is derived using a Bayesian formulation under guidance from film grammar, tested under many configurations and the results are reported for experiments involving 25 full length movies. The formulation is shown to be a useful tool in both the automatic and semi-interactive setting for semantic analysis of film.

1. INTRODUCTION
In this paper we examine the extraction of narrative structure in film, in particular, the 3-Act story telling paradigm. Narrative structure is an integral part of film, forming the foundation and framework within which the film can be made to function as an effective story telling tool. The novelty of this work is that it constitutes the first attempt to formulate a computational framework designed on the basis of production principles, leading to an algorithmic extraction of this construct using a Bayesian formulation. The significance of this work consists of:

i) The extraction of this high-level structure in film, that is at a semantically higher level than previous structuralizing components such as scenes and shot sequences

ii) A correspondingly enriched ability to infer reliably about content, making automatic replies possible to such queries as “Where does this film’s climax occur?” or “How long is the opening sequence, and is it character or action oriented?”, etc.

iii) Comparative analysis among films in these terms, such as 3-Act paradigm confidence, relative location or length of climax, opening sequence, to name a few

iv) It is important to note that although these ideas are be-

Fig. 1. The Aristotelian curve of tension underlying narratives in movies.

ing applied to film initially, other domains exhibit analogous elements that come under the rubric of storytelling, each within the confines of their particular production grammar (e.g., training video, news, documentary) and will benefit from the techniques developed here.

2. BACKGROUND AND MOTIVATION
The Computational Media Aesthetics (CMA) philosophy detailed in [1] advocates scrutinizing a particular domain in order to find the structures that are inherent to it, and are therefore valid and worthwhile targets for extraction. As stated, our chosen domain is film. Film is a subset of the broader domain of story, and a consequence of this, Mehring notes, is “that the largest proportion of our screenplays utilize principles of storytelling that have been around for a very long time”. She goes on to quote Aristotle’s Poetics, which identifies the classical dramatic structure that contains a beginning, a middle and an end, each being a necessary part. “There will always be a beginning, a middle and an end no matter how the parts of the work are arranged and divided. This is the basic structure of storytelling” [2]. See Figure 1.

It is important to note that films issue from screenplays. It is here, in the proto-film that is the script, that the dramatic structure of the work is crafted, ultimately to form the framework within which the derivative work, the movie, lives and breathes.

Identification and extraction of these structures opens the way for automatically generated indexes or semi-interactive tools that allow for query specification in semantic terms fit for general consumption. Plot, climax, opening sequence,
are all terms which are understood to a reasonable degree by
the average user, and are constructs at least an order above
the tedium of shot level. The higher the level of the initial
data structure, the quicker the search space is reduced, and
the more efficient the search.

3. ELEMENTS OF NARRATIVE STRUCTURE
FROM FILM GRAMMAR
Now that we have an understanding of what narrative struc-
ture is, and its usefulness as a high-level construct, the next
question is how is it made manifest? In line with the CMA
philosophy, not only has film grammar alerted us to the ex-
istence of narrative structure, it is our guide in looking for
its signatures.

The following act structure characteristics have been gle-
aned from both explicit and implicit references within the
body of film literature and practice that forms the grammar of
film.
i) Act boundaries occur in known approximate locations, and
may be an instant (generally) or a sequence (span a se-
ries of shots).
ii) Each act contains one climax which is the most intense
or dramatically important in the act, and close to, if not the
last dramatic event in the act.
iii) There is often, but not always, a “breather” after an act
boundary.
iv) Structure is sculpted from the placement of dramatic
events, which in turn leave their signatures in terms of tempo,
sound, colour and so on.
v) As the narrative structure is subject to variations, a com-
putational scheme should intrinsically deal with them.

4. A BAYESIAN FORMULATION FOR
EXTRACTING NARRATIVE STRUCTURE
So we have the key aspects, how are we going to character-
ize them, and synthesize them into a formulation to compute
the boundaries of narrative acts? The most obvious starting
point is the fact that we are dealing with uncertainty and
variations across movies, which recommends a probabilis-
tic framework; we choose a Bayesian formulation in order
to draw the identified factors together.

A nice prior is given theoretically in the form of the ap-
proximate locations of the act boundaries. The prior prob-
ability density \( P(t_a) \), where \( t_a \) is the location of the act
boundary, can be constructed from this theoretical approxi-
mation, or garnered from a simple training exercise against
a number of actual films.

We next utilize another aspect of narrative act structure
from above, namely that each act will have exactly one dra-
matic climax. As stated above, dramatic events, such as an
act climax, leave peaks in the tempo landscape of a film.
Therefore, let \( t_v \) be the time of the single visual tempo peak
corresponding to the visual component of the act climax,
and likewise, let \( t_s \) be the time of the single aural tempo
peak corresponding to the aural component of the act cli-
max. With this assumption, that the location of the act
boundary is solely dependent upon the time of these com-
ponents of the act climax, our formulation hinges upon these
“contributing” tempo peaks. The relationship between \( t_a \),
\( t_v \), and \( t_s \) is summarized in the joint distribution of these
variables:

\[
P(t_a, t_v, t_s) = P(t_v | t_a)P(t_a). \tag{1}
\]

While the act climax can be considered a single dra-
matic event, it can vary with regard to degree of isolation in
time, and as such it is often the case that the visual tempo
peak component of the climax is separated in time (albeit,
only to a small degree in comparison with the length of
the movie) from the aural component. Given this situation,
we will study both the case mentioned above, Equation (1),
which assumes nothing about the degree of dependency be-
tween \( t_v \) and \( t_s \), and the case for which \( t_v \) and \( t_s \) are in-
dependent given \( t_a \), which yields,

\[
P(t_v, t_s | t_a) = P(t_v | t_a)P(t_s | t_a). \tag{2}
\]

Now we consider the evidence for \( t_v \) to be the sequence
of visual tempo peaks \( \epsilon_v \) from the given movie. Likewise
we consider the evidence for \( t_s \) to be the sequence of aural
tempo peaks \( \epsilon_s \). It is to be noted that a more complex for-
mulation could consider an extended body of evidence for
\( t_v \) and \( t_s \). Given that \( t_a \) is dependent only on \( t_v \) and \( t_s \) we have,

\[
P(\epsilon_v, \epsilon_s | t_v, t_s, t_a) = P(\epsilon_v | t_v)P(\epsilon_s | t_s). \tag{3}
\]

And following the simplifying assumption that \( t_v \) is
independent of \( \epsilon_s \), and \( t_s \) is independent of \( \epsilon_v \), Equation (3)
becomes,

\[
P(\epsilon_v, \epsilon_s | t_v, t_s) = P(\epsilon_v | t_v)P(\epsilon_s | t_s). \tag{4}
\]

This assumption is made for mathematical simplicity. It
translates to independence between the visual stream and
the sound track in the movie.

Now we are in a position to fully formulate the problem
of finding \( t_a \). Bayes formula gives us,

\[
P(t_a | \epsilon_v, \epsilon_s) \propto P(\epsilon_v, \epsilon_s | t_a)P(t_a). \tag{5}
\]

This gives us the following, with reference to Equations
(3) and (4),

\[
= \sum_{t_v} \sum_{t_s} P(\epsilon_v | t_v)P(\epsilon_s | t_s)P(t_v, t_s | t_a)P(t_a). \tag{6}
\]

Note that the evidence \( \epsilon_v \) is a sequence of peaks,
\( (\epsilon_v(1), \ldots, \epsilon_v(N)) \), and each of these peaks has an effec-
tive magnitude and time denoted by \( \epsilon_v^{mag}(i) \) and \( \epsilon_v^{time}(i) \)
respectively. We assume that only one of these peaks cor-
responds to the climax visual peak \( t_v \). Thus, the conditional
probability \( P(\epsilon_v | t_v) \) can be simplified to:
\[ P(\epsilon_v|t_v) = \begin{cases} 0 & \text{if } t_v \text{ is not a peak, and} \\ \propto P(e_v^{\text{mag}}(n)) & \text{if } t_v = \epsilon_v^{\text{lim}}(n) \end{cases} \]  

(7)

where \( P(e_v^{\text{mag}}(n)) \) is the probability distribution of the magnitude of the climax visual peak. \( P(\epsilon_s|t_s) \) is defined similarly.

Now we can write Equation (6) as,

\[ P(t_s|\epsilon_v, \epsilon_s, \epsilon_{\text{act}}) \propto P(\epsilon_v, \epsilon_s, \epsilon_{\text{act}}|t_s)P(t_s), \]

(9)

which with the assumption that \( \epsilon_{\text{act}} \) is independent of \( \epsilon_v \) and \( \epsilon_s \) given \( t_s \) yields,

\[ P(t_s|\epsilon_v, \epsilon_s, \epsilon_{\text{act}}) \propto P(\epsilon_{\text{act}}|t_s)P(\epsilon_v, \epsilon_s|t_s)P(t_s). \]

(10)

Following the formulation of Equation (5) through Equation (6), we can rewrite(10) as

\[ P(t_s|\epsilon_v, \epsilon_s, \epsilon_{\text{act}}) \propto P(\epsilon_{\text{act}}|t_s) \sum_{n=1}^{N} \sum_{m=1}^{M} \frac{P(e_v^{\text{mag}}(n))P(e_s^{\text{mag}}(m))}{P(e_v^{\text{mag}}(n))P(e_s^{\text{mag}}(m))}P(\epsilon_v^{\text{lim}}(n), \epsilon_s^{\text{lim}}(m)|t_s)P(t_s). \]

(11)

\( \epsilon_v \) and \( \epsilon_s \) are all computed, with \( e_v^{\text{mag}} \) and \( e_s^{\text{mag}} \) being calculated from the peaks in their respective tempo signals. All distributions on the right-hand side of Equation (11) can be obtained via training. The probability of \( t_s \) being the Act 1 boundary may then be calculated with Equation (11) for all \( t_s \), and the maximum chosen as the winner.

5. IMPLEMENTATION

5.1. Data

The dataset for this experiment includes a number of whole movies (25), approximately half of which have been identified in the film literature (e.g. [3, 4, 5]) as conforming to the 3-Act paradigm, with the remainder verified as such by the authors. The former receive an act 1 boundary ground truth from those sources, whilst the latter have their ground truth manually located by the authors. The selection includes a number of films from different genres, of different ages and length. In addition to the act boundary, the locations of the act 1 climax component peaks (\( t_c \) and \( t_s \)) are also ground truthed similarly. Each movie is processed using shot detection, followed by computations of shot length, motion and raw audio magnitude estimates for visual tempo (using the algorithm from [6]) and aural tempo processing.

5.2. Training

The training is conducted with a Leave One Out policy [7, p. 26], which is computationally expensive, but yields an unbiased result. Thus L-1 movies are used for training, and then the act boundary of the Lth movie is determined. The inputs to the training phase consist of:

i) The act boundary location (\( t_a \)), as a percentage of the total movie time. This is the source for \( P(t_a) \)

ii) The act climax, both the visual component (\( t_v \)), and the aural component (\( t_a \)), both as a percentage of total movie time before the act boundary (ie. \( \frac{\text{ts} - t_a}{\text{ts}} \times 100 \)). This is the source for \( P(t_v|t_a) \) and \( P(t_a|t_a) \).

iii) The effective magnitude for the peaks at both \( t_v \) and \( t_a \), where effective magnitude is defined as: Avg. tempo in window \( w_1 \) - Avg. tempo in window \( w_2 \). This is the source for \( P(\epsilon_{\text{act}}|t_a) \).

Using Matlab, \( P(\epsilon_{\text{act}}|t_a) \), \( P(\epsilon_v|t_v) \), \( P(\epsilon_s|t_s) \), \( P(t_v|t_a) \), \( P(t_a|t_a) \) are computed with normal fit, both at \( \mu \) and \( \sigma \), and at the higher (broader) \( \sigma \) of a 95% confidence interval. \( P(\epsilon_{\text{act}}|t_a) \), \( P(\epsilon_v|t_v) \) and \( P(\epsilon_s|t_s) \) are also computed with a triangular pdf \((f(x) = 2x/h^2) \), to test the assumption of dramatic intensity being proportional to tempo (i.e., the more important the dramatic event, the bigger the tempo peak; the act climax is the most important event in the act).

5.3. Calculating \( P(t_a|\epsilon) \)

Visual tempo is obtained for each movie according to the algorithm of [6], and an aural tempo measure is obtained using averaged raw audio magnitude. From these signals the local maxima are obtained which provides the evidence sequences, \( \epsilon_v \) and \( \epsilon_s \) respectively. These are recorded as a sequence of frame location, and effective magnitude as defined above. Tempo average difference is calculated as above for every shot.

Every shot is then considered, and \( P(t_a|\epsilon) \) is calculated as per Equations in Section 4. The pdfs used are those that were trained from the set of all movies excluding the movie being tested, as mentioned above.

Figure 2 shows a plot of sound and visual tempo for the movie “Starwars: Episode 1”, and below, a plot of \( P(t_a|\epsilon) \).

5.4. Results

In order to ascertain the performance of \( P(t_a|\epsilon) \), the shot corresponding to the maximum a posteriori (MAP) is considered the act boundary, the location of the computed act boundary is compared with the ground truth location of the actual act boundary, and an error returned as the difference between the two in frames. A number of different experiments were performed, each with a different configuration for \( P(t_a|\epsilon) \). The best result across the 25 movies was 4.1 minutes, and was obtained with a wide prior \( P(t_a) \), triangular \( \epsilon_v \), \( \epsilon_s \), and the \( \epsilon_{\text{act}} \) information. For an average length
movie this equates to about 3.5% error in act boundary location. A “best-of-2” analysis, taking into account the second best local maxima, reports an overall avg. error of 2.9 mins, and is more appropriate to a semi-interactive setting.

It is interesting to note that there are a handful of movies that do not perform well with any configuration for $P(t_a | e)$. Upon examination we find that there are examples of all three possible reasons for this failure (and some movies contain more than one):

1. Strange dramatic structure: These can be considered as “dramatic structure outliers”, where the film includes a single very rare narrative structure, or a number of them in combination. E.g. OrdinaryPeople has a very early act 1 boundary (ground truth by Vogler [4, p. 128]).

2. Elusive structure: Where the mapping between dramatic event and tempo breaks down, $P(t_a | e)$ suffers accordingly. This happens particularly where the dramatic event concerned is subtle (character-based, emotional, intellectual) and the filmmaker doesn’t reinforce it with any of the usual tempo signatures such as motion, shot length, etc. Starwars4 has a purely emotional climax and no breather after the boundary (a case of problem 1) which together conspire to defeat detection of the boundary (ground truth by Vogler [4]).

3. Problems with our process: This includes problems of inaccuracies in the foundational data, such as shot indices or motion estimates, upon which $P(t_a | e)$ is ultimately built. In SilenceOfTheLambs, the act climax occurs in a very dark environment and the shot index contains false negatives (ground truth by Thompson [5]).

6. CONCLUSIONS

In summary, the optimal performance of the resulting formulation is 4.1 minutes error across 25 movies for MAP. This performance increased, as expected, in the best-of-x style analysis, to 2.9 minutes (e.g. best of 2, at 50% of MAP). $P(t_a | e)$ would thus prove useful in both a fully automatic and semi-interactive setting.

While the immediate domain of this work is film, we also highlight the fact that many other domains exhibit analogues to the structural components extracted here. News, documentary, training videos, sitcoms, and computer games, all have identifiable grammars, and structural conventions that are adhered to and may be exploited for the purpose of automated understanding.

7. REFERENCES