RC22776 (W0304-105) April 11, 2003
Computer Science

IBM Research Report

On the Extraction of Thematic and Dramatic Functions of Content in Educational Videos

Dinh Quoc Phung, Svetha Venkatesh
School of Computing
Curtin University of Technology
GPO Box U1987, Perth, 6845, Western Australia

Chitra Dorai
IBM Research Division
Thomas J. Watson Research Center
P.O. Box 218
Yorktown Heights, NY 10598

Research Division
Almaden - Austin - Beijing - Delhi - Haifa - India - T. J. Watson - Tokyo - Zurich

ON THE EXTRACTION OF THEMATIC AND DRAMATIC FUNCTIONS OF CONTENT IN EDUCATIONAL VIDEOS

Dinh Quoc Phung, Svetha Venkatesh

School of Computing
Curtin University of Technology
GPO Box U1987, Perth, 6845, Western Australia
{phungquo,svetha}@computing.edu.au

Chitra Dorai

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

ABSTRACT

In this paper, we propose novel computational models for the extraction of high level expressive constructs related to, namely *thematic* and *dramatic* functions of the content shown in educational and training videos. Drawing on the existing knowledge of film theory, and media production rules and conventions used by the filmmakers, we hypothesize key aesthetic elements contributing to convey these functions of the content. Computational models to extract them are then formulated and their performance evaluated on a set of ten educational and training videos is presented.

1. INTRODUCTION

Along with an ever increasing volume of multimedia data nowadays, comes an urgent demand for content management technology. Effective management techniques will no doubt play a key role in the design of future intelligent multimedia systems. One of the pressing problems in media content management is the challenge to *bridge the semantic gap* between low-level, easily computable media features and semantic constructs that are capable of handling users' queries at higher levels of understanding. Our ongoing work attempts to design algorithms to facilitate semantic analysis and automatic structuralization of educational videos. In [1], we proposed a hierarchy of narrative structures used in educational videos along with a tiered-classification system to determine them. In [2], the *content density* function was introduced and evaluated in the context of topic boundary determination in instructional media. This paper presents our continuing investigation of this domain with two new high level expressive constructs, namely the *thematic* and *dramatic* functions of the educational content. Drawing on the extensive body of film theory, and media production rules and conventions used by filmmakers, we hypothesize key aesthetic elements that influence the 'perception' of these expressive constructs in educational video. Computational models for these functions are then formulated, and an evaluation of the performance of these functions is presented on several industrial safety training videos.

Related work is identified extensively w.r.t multimedia content management, but few dealing directly with education and training videos. Studying meaning of signs – or semiotic – and relating them to express high level semantic content of commercial videos is found in [3]. Low-level features are used to map a piece of commercial segment into more expressive description including: *practical, playful, utopic* and *critical*. Also recently found in the literature is the study of films for mining semantic descriptions of the content. In [4], for instance, Adams et al. formulate an algorithmic solution for the computation of movie *tempo*, a high-level semantic construct, and propose segmenting a movie into meaningful story units based on the ebb and flow of drama in the film as underlined by its tempo. Mining education videos for topical events is studied in [5]. In this work, visual events in the image sequence of foils are detected and then incorporated with audio information in a probabilistic framework for topic detection.

2. THEMATIC AND DRAMATIC FUNCTIONS OF EDUCATIONAL CONTENT

Educational video is defined as "a motion picture designed to teach" [6, p.9]. A well-crafted video segment that motivates learners to actions or enables them to retain a message that can last in their memory is not simple to produce. Numerous aesthetic choices must be made by the director in video production and the issue is not only *what* is to be shown but also *how* it is to be presented to achieve the maximum impact on learners. Addressing film techniques in general, Foss [7] approaches the *what* and *how* in films from two distinct views of what he calls, *Plane of Events* and *Plane of Discourse*. Elements that "can or could be perceived by the characters in the film or program" belong to the *Plane of Events*, whilst the *Plane of Discourse* contains factors that "are imperceptible to the characters in the film". Hence, in the context of edu-

	What we see and hear	*How* we see and hear
General films	Physical appearance, acting, setting, props, costume and make-up, time, weather, movement, physical relations, real sound, dialog, etc ..	Format, shot composition, size of shot, camera angle, color and monochrome, type of film and exposure, sound effects, etc ...
Educational videos	Appearance of narrator(s)/presenter(s), motion in scenes, music, speech, voice over and use of superimposed text captions.	Use of color (hue, saturation, lightness values), camera pan, tilt and zoom, sound effects (e.g., expressive silence)

Table 1. Media elements for expression [7].

cational videos, contents and materials presented are captured in Foss's *Plane of Events* and different techniques employed to bring them to the viewers belong to the *Plane of Discourse*.

What media elements are then available for the filmmakers to manipulate these *planes*? Table 1 shows a partial list of suggested factors [7]. Obviously, this list comes from the viewpoint of a filmmaker, and not all of them are applicable or amenable to our computing framework. The element *weather*, for example, is still far from being automatically extractable by computers. Limiting our analysis to those computable elements and educational domain, we propose a list of elements for further study as shown in Table 1.

Given the set of available aesthetic choices to express the *what* and *how* in motion pictures, Foss [7] comes to define six distinct functions of narrative expression, including: realistic, dramatic, thematic, lyrical, comic, and extraneous. In this study, we are mainly interested in the thematic and dramatic functions.

Thematic Function

A narrative function is said to be thematic when it acts "as a *comment* on or *interpretation* of what happens on the Plane of Events" [7]. In the case of educational videos, we interpret the thematic function as being *instructional* and *informative*. It reflects portions of the video where the filmmaker decides to 'step in' and interfere in the subject being shown. Voice of authority over a raw footage in documentary videos, for example, would help the video makers to clarify the visual content, and perhaps to make an appeal for his/her subjective point of view. Superimposed text in training videos would draw trainees to the specifics and emphasize key messages. Extraction of such a function that can disclose the degree of the videomakers' involvement, or mediation level in a sense, would be useful for content management, especially in educational domain. This will immediately allow us to segment sections with high-level of informative/instructional contents, and facilitate queries such as "finding me segments with specific instructions".

Dramatic Function

If the thematic function is designed to capture the degree of the videomakers' mediation, the dramatic function reflects the 'interesting' or the dramatic nature of presentation in the mediating process. Foss [7] sees dramatic function as that which "influences human relations as well as people's wishes, opinions, and choices", and the lyrical function as that which "create[s] a particular atmosphere or feeling". In the domain of educational videos, we shall combine these two and call them the *dramatic function*. So what do we expect this function to tell us ? Consider the example about the dramatization of ladder safety in the training video, "Maintenance". Here, the filmmaker chooses to not only talk about safety rules, but also decides to *dramatize* what is said by showing a sequences of images. Narration is stopped, ambient music is played in the background and the falling actions shown in the scenes. The dramatic function should reach a climax for this sequence.

Based on these definitions and examination of several educational videos, we hypothesize that the key elements that are manipulated to influence thematic and dramatic functions of the content are: Use of color, degree of motion shown, use of superimposed text, soundtrack (e.g., augmented with music, voice over narration) and the appearance of the narrator(s). Table 2 shows these elements and their occurrence styles and impact on these two categories of functions.

3. RELATING MEDIA ELEMENTS TO THEMATIC AND DRAMATIC FUNCTIONS

Given these media elements and the hypothesized impact (Table 2), how do we justify which element should be accounted for in determining the functions of the content automatically? For example, should Music be counted as a variable in computing a measure of thematic function for that shot? While we do not have groundtruth from the literature to justify what should be candidates to highlight 'good' thematic/dramatic functions, we resort to what we call an *extreme-case* methodology. Given an educational video, we watch and manually label segments, which, in our opinion and according to the definitions, are deemed to be *extreme* thematic or dramatic sections. A section with scrolling text and voice overs, for example, is a form of extreme thematic. By extreme, we mean in a way "undoubtedly". A fast editing segment augmented by music, contains no narration, as observed in the beginning of some educational films is an example of extreme dramatic.

Ten videos are used in this experiment. Shot indexes for each video are first generated by the Webflix software and then shot detection errors are corrected manually. Extreme segments are manually identified at the shot level, i.e., each segment consists of a sequence of shots. Let $\{\mathcal{T}\}$ be the set of all segments labeled as extreme thematic and $\{\mathcal{D}\}$ be the set of those labeled as extreme dramatic. From the visual channel of the video, we extract the following features (where n is a shot index):

+ *Color*. Hue, Lightness, Saturation, Warmth/Cold (detailed in [8]).
+ *Motion*. Average shot motion, (**Mo**$[n]$) is estimated as the average of camera tilt and pan [1].
+ *Superimposed Text*. We use `TextContentRatio`, (**Tx**$[n]$), which is measured as the ratio of number of frames with text captions detected to the total number of frames in the shot [1].
+ *Narrator Presence*. We use `FaceContentRatio` (**Fa**$[n]$), which is measured as the ratio of number of frames with face(s) detected to the total number of frames in the shot [1].

Further, from the audio channel we extract `MusicRatio` (**Mu**$[n]$), `SpeechRatio` (**Sp**[n]), `SilenceRatio`, and `NonliteralRatio` as detailed in our previous work [1]. `MusicRatio`, for example, is measured as the ratio of number of audio clips classified as music to the total number of clips making up the sound track of the shot.

Figure 1 shows a part of our analysis. The bargraphs are plotted with average values of these features computed over $\{\mathcal{T}\}$ and $\{\mathcal{D}\}$ segments under study. The linegraphs show some examples of how these features vary across dramatic and thematic segments. Regarding audio information (Fig. 1a), it appears clearly that the `MusicRatio` dominates in dramatic segments, while the `SpeechRatio` dominates in thematic segments. Although there are significant differences in `SilenceRatio` and `NonliteralRatio` between thematic and dramatic, we do not consider them as key influencing factors as their values all fall short under 30%. In Figure 1b, we observe the dominance of both `FaceContentRatio` and `TextContentRatio` for thematic segments. In studying the contribution of color, Fig. 1c shows no significant difference of the Hue and Saturation values between thematic and dramatic sections (note that the vertical axis here denotes the average values of Hue and Saturation, not the ratios), and does not strongly support our hypothesized impact of color as indicated in

	Color	Motion	Music	Voice Over	Superimp. Text	Narrator Presence
THEMATIC	*high Hue*	*low*	*rarely used*	*high*	*high*	*high*
	simple background color is typically used to avoid distraction from text captions and presenters.	less movement is involved.	mainly narrating voice or silence (e.g., with scrolling text)	voice of authority to narrate over the content	introduce topics, remarks, emphasize, keypoints, reminder messages, etc.	appears to directly speak to the viewers.
DRAMATIC	*low Hue*	*high*	*high*	*low*	*low*	*low*
	usually diversity in color.	lots of motion due to actions in the scenes.	sometimes music is added to create the mood.	usually narration is stopped.	possible appearance of scene text *but not* superimposed text	narrators do not show up.

Table 2. Media elements and their utility in conveying thematic and dramatic functions of the content in educational video.

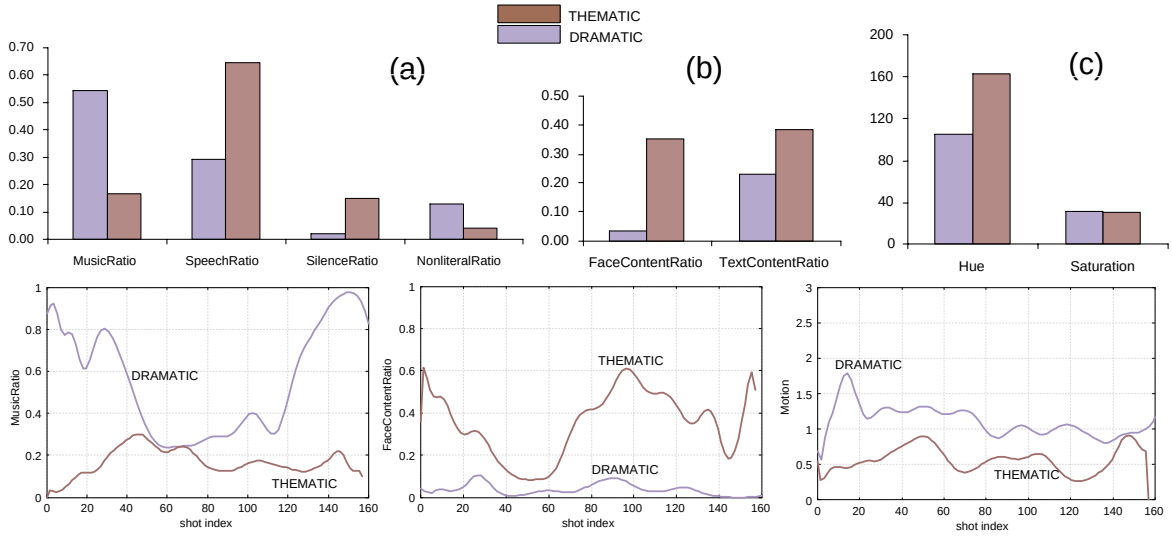


Fig. 1. Establishing relationships between proposed media elements and their impact on thematic and dramatic nature of the educational content.

Table 2. We, therefore, decide to omit them from the list of primary contributing factors. We also observe the dominance of motion in dramatic sections, but not in thematic sections (Fig. 1).

Our results, from analyzing the complete data set show that thematic segments are influenced primarily by: `SpeechRatio`, `TextContentRatio` and `FaceContentRatio`. Dramatic segments are influenced primarily by: `MusicRatio` and `Motion`. Based on these results, we formulate a function indicating the thematic function of the content as:

$$\mathbf{Th}[n] = \alpha_{\mathrm{Th}}\mathbb{N}\{\mathbf{Sp}[n]\} + \beta_{\mathrm{Th}}\mathbb{N}\{\mathbf{Tx}[n]\} + \gamma_{\mathrm{Th}}\mathbb{N}\{\mathbf{Fa}[n]\} \quad (1)$$

and likewise for the dramatic function:

$$\mathbf{Dr}[n] = \alpha_{\mathrm{Dr}}\mathbb{N}\{\mathbf{Mu}[n]\} + \beta_{\mathrm{Dr}}\mathbb{N}\{\mathbf{Mo}[n]\} \quad (2)$$

where n is the shot index, $\mathbb{N}\{.\}$ is the normalization operator defined as: $\mathbb{N}\{x\} = \frac{x-\mu_X}{\sigma_X}$, and $\alpha_{\mathrm{Th}}, \beta_{\mathrm{Th}}, \gamma_{\mathrm{Th}}$ and $\alpha_{\mathrm{Dr}}, \beta_{\mathrm{Dr}}$ are weighting factors, which are set to 1 in this work. In essence, it means each chosen media element equally contributes to the perception of these expressive functions. These functions are then smoothed with a Gaussian filter.

4. EXPERIMENTAL RESULTS

In this experiment, we aim to study whether hand-labeled *extreme* segments do correspond to the *extrema* (ie. maximum) in these indicator measures for thematic and dramatic nature of the content. For a video, **V**, to evaluate extreme *thematic* segments, we first compute its corresponding measure, and detect all peaks of this measure. A peak p whose thematic value ($\mathbf{Th}[p]$) exceeds a threshold, **TH** is deemed to be an extreme peak and denoted as $p^{\mathbf{ext}}$. Let $\mu_{\mathbf{Th}}$ be the mean of thematic values for the video **V** computed from $\mathbf{Th}[.]$. In this study, we set the threshold to be $\mathbf{TH} = \mu_{\mathbf{Th}} + 20\% \times |\mu_{\mathbf{Th}}|$ (i.e., 20% more than average).

A true positive is recorded if $p^{\mathbf{ext}}$ belongs to one of groundtruthed segments, and a false positive is counted otherwise. A miss is

found when a groundtruthed segment does not correspond to any of the peaks that exceed **TH**. We evaluate the *dramatic* content segments in a similar way. The results for this study are shown in the table below.

V	THEMATIC					DRAMATIC				
	*	**	TP	FP	Ms	*	**	TP	FP	Ms
1.	10	6	5	1	0	8	2	1	1	0
2.	6	5	5	0	0	8	2	1	1	0
3.	7	6	4	2	1	7	1	1	0	0
4.	8	7	5	2	0	6	2	2	0	0
5.	12	8	4	4	0	10	2	0	2	0
6.	13	8	7	1	0	12	4	4	0	1
7.	18	10	8	2	0	21	3	3	0	1
8.	7	5	4	1	0	7	5	4	1	0
9.	8	7	5	2	1	9	2	2	0	0
10.	14	11	8	3	0	12	2	2	0	2
all	103	73	**55**	**18**	**2**	100	25	**20**	**5**	**4**
*** : Number of detected maxima** **** : Number of detected maxima exceeding TH**										

Fig. 2a

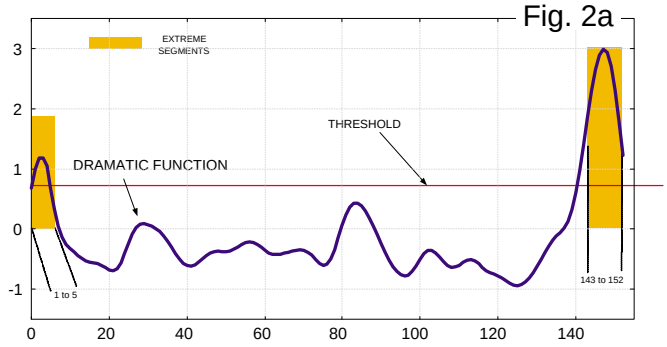
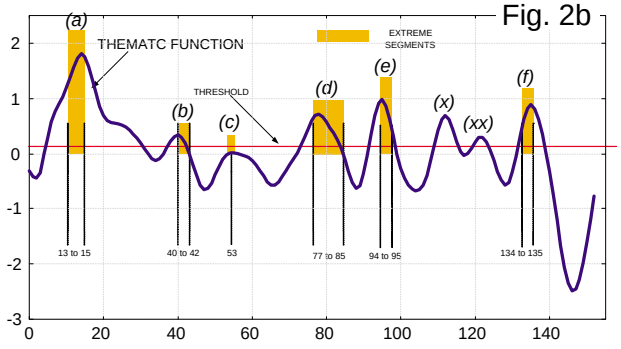


Fig. 2b



The recall and precision recorded for thematic segments are 96.5% and 75.3% respectively, and 83.3%, 80% respectively for dramatic segments, where recall is defined as $\frac{TP}{TP+Miss}$ and precision as $\frac{TP}{TP+FP}$. An analysis of these results reveals that:

- False positives generally correspond to very weak peaks when compared to peaks identified as true positive.
- Misses are mostly attributed to errors in text/face detection results.

Figure 2 shows an example of dramatic and thematic functions computed for video number 9 ("Electronics-Safety"). Segments groundtruthed as extreme are also shown. In Fig.2a, we observe two strong peaks for dramatic content at the beginning and the end of the video, corresponding to extreme dramatic sections as indicated in the groundtruth. From the video, it shows that the strongest peak at the end corresponds to the section where the video makers summarize the video and 'dramatize' what was said. Music is played, narration is stopped, and a sequence of dissolves with lots of actions in the scenes are shown.

In Fig.2b, we observe strong peaks at (a),(d),(e) and (f). These peaks all correspond to manually labeled extreme sections. The segment for shot from 40 to 42 (peak b) also exceeds the threshold, though it is a weak peak. However, there is one shot (53, peak c) that does not exceed the threshold and we also observe two cases of false positive (peaks x and xx).

5. CONCLUSION

We have defined and studied the thematic and dramatic functions of the content in educational videos. We first hypothesize key media elements, followed by an experiment on ten videos to point out primary contributing factors to these constructs. We then define two novel measures to indicate the thematic and dramatic nature of the content portrayed. Finally, we describe an experiment to evaluate the performance of these functions. The results have shown the validity and the usefulness of these measures. Future work includes the use of more advanced classifiers in labeling the two functions of educational content and exploiting them for hierarchical topic segmentation.

6. REFERENCES

[1] Dinh Quoc Phung, Chitra Dorai, and Svetha Venkatesh, "Narrative structure analysis and topic segmentation in educational films," *Submitted to Signal Processing: Image Communication, Special Issue on Multimedia Adaption*, 2002.

[2] Dinh Quoc Phung, Chitra Dorai, and Svetha Venkatesh, "High level segmentation of instructional videos based on content density," in *ACM International Conference on Multimedia*, Juan Les Pins, France, 1-6 December 2002, pp. 296–298.

[3] Carlo Colombo, Alberto Del Bimbo, and Pietro Pala, "Retrieval of commercials by semantic content: the semiotic perspective," *Multimedia Tools and Applications*, 2001.

[4] Brett Adams, Chitra Dorai, and Svetha Venkatesh, "Novel approach to determining movie tempo and dramatic story sections in motion pictures," in *2000 International Conference on Image Processing,*, Vancouver, Canada, September 2000, vol. II, pp. 283–286.

[5] Tanveer Seyeda-Mahmood and S. Srinivasan, "Detecting topical events in digital video," in *ACM Multimedia*, 2000, pp. 85–94.

[6] Lewis Herman, *Educational Films: Writing, Directing, and Producing for Classroom, Television, and Industry*, Crown Publishers, INC., New York, 1965.

[7] Bob Foss, *The Filmmaking: Narrative and Structural Techniques*, Silman-James Press, Los Angeles, 1992.

[8] Ba Tu Truong, Chitra Dorai, and Svetha Venkatesh, "Automatic scene extraction in motion pictures," *IEEE Trans. in CSVT, Special issue on multimedia content description*, 2002.