
Available from Deakin Research Online:

http://hdl.handle.net/10536/DRO/DU:30044651

Reproduced with the kind permissions of the copyright owner.

Personal use of this material is permitted. However, permission to reprint/republish this material for advertising or promotional purposes or for creating new collective works for resale or redistribution to servers or lists, or to reuse any copyrighted component of this work in other works must be obtained from the IEEE.

Copyright: 2002, IEEE
Narrative Structure Analysis with Education and Training Videos for E-Learning

Quoc Dinh Phung†, Chitra Dorai‡, Svetala Venkatesh†
Department of Computer Science†, Curtin University of Technology
GPO Box U1987, Perth, 6845, W. Australia
{phungquo, svetha}@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 paper deals with the problem of structuralizing education and training videos for high-level semantics extraction and nonlinear media presentation in e-learning applications. Drawing guidance from production knowledge in instructional media, we propose six main narrative structures employed in education and training videos for both motivation and demonstration during learning and practical training. We devise a powerful audiovisual feature set accompanied by a hierarchical decision tree-based classification system to determine and discriminate between these structures. Based on a two-tiered hierarchical model, we demonstrate that we can achieve an accuracy of 84.7% on a comprehensive set of education and training video data.

1. Introduction

As widely acknowledged in the video indexing research community, the most pressing problem in content management is to develop methods to automatically structuralize multimedia data in order to bridge the semantic gap through high-level semantics-based video partitioning, event extraction, and content tagging for automatic media search and browse processes to become more effective. While much research in this field has targeted broadcast videos [5, 9], and motion pictures [2, 10], little attention has been devoted to instructional media, especially education and training videos. This paper addresses the issue of automatically structuralizing these videos for provisioning novel e-learning services over the web.

Partitioning broadcast videos such as news reports into meaningful sections has been studied extensively. In [9], Shearer et al. combine a number of visual and aural low-level processes together with the concept of shot syntax to label different narrative structures used in broadcast news programs such as anchor shots, voice overs, and interview sections. Ichiro et al. [5] propose an automatic indexing scheme for television news video, where shots are indexed based on the image content and keywords. Shots in this work are classified into five different categories: speech/report, anchor, walking, gathering, and computer graphics. Caption text information is then used with the classified shots to build the indices. Film has also been partitioned lately using semantic attributes in the literature. In [2], Adams et al. formulate an algorithmic solution to the computation of movie tempo, and propose segmenting high-level story units based on the ebb and flow of tempo in the film. The work of Wang et al. [10] attempts to extract scenes in film using visual similarity and further improves it with the guidance of cinematic rules.

In this paper, we address the first step towards automatic structuralization of education and training videos, a domain which is gaining notice with recent interest in e-learning media technologies [4]. We present commonly observed rules and conventions in instructional media productions to manipulate the presentation of content in media to match learners' needs. Leveraging production grammar to shape our understanding of the common structural elements employed in instructional media, we propose a hierarchy of narrative structures used. Characteristics of each structural element manifest as a sequence of shots are examined and a set of audiovisual features for capturing the differences between them is proposed. The C4.5 algorithm is used in our hierarchical classification system. Experimental results on six videos are presented to demonstrate the richness and effectiveness of the feature set proposed and the resulting classification of narrative structures.

2. Narrative Structure Hierarchy

As noted by [6], the structure of a video is greatly influenced by the nature of the information in it and the purpose for which the video is used. Education and training video is not an exception and indeed far from being a haphazard placement of content, it is a highly structured
medium aimed at having the maximum impact on the learners. We propose three main narrative structures in these videos based on the study of learning videos: narration sections, conversational/discussion sections, and linkage sections. Narration sections are further categorized into on-screen and voice over as shown in Figure 1.

**Figure 1. Proposed narrative structures in education and training media.**

*On-screen* narration refers to sections in an instructional video with a clear view of a narrator. We observe two finer groupings within this category: *direct* (eye-contact) narration and *assistive* narration. *Direct* narration (DN) is identified as one in which the narrator faces the camera directly and the video frames are spatially dominated by their faces which are shown as close-ups. The *assistive* narration (AN) mode accounts for sections with interrupted presence of the narrator’s face in the frame caused by the movements of the narrator (e.g., shots in a safety video in which the narrator goes to different places of a factory and shows the viewers how accidents can occur). Distinguishing these two categories will be helpful in automatically constructing the storyline followed in the video. For example, DN shots are typically encountered in the beginning with the narrator to introduce the subject and at the end to conclude an event, while the AN shot usually appears to show supplementary material and at the same time, maintain the narrative flow. To some extent, DN shots are analogous to anchor shots and AN to field reporter shots in a news program.

*Voice over* section is captured in the video when the audio track is dominated by the voice of the narrator but without his/her face being shown in the shot. At a finer level, an *un-interrupted* voice over (UV) shot is identified as one with the voice of the narrator consistently delivered across the shot, while an *interrupted* (IV) shot is the one in which the flow of narrative voice is occasionally interrupted (e.g., in a shot showing how a truck can accidentally hit a child, the narrator speaks about the cause, then stops to illustrate the case and then the voice resumes).

The *conversational/discussion* category (DD) captures the interviews or conversation/question and answer segments in a lecture video. The *linkage* section (LF) accounts for the rest of the video and captures connecting shots that cannot be classified into any of above categories. This usually happens to be raw footage and/or shots with superimposed text with no narration, as often used in training video for emphasis.

**3. Feature Extraction and Shot Analysis**

The input to our analysis is a digitally encoded education and training video in MPEG-1 format. As a first step, shot detection on the video stream is carried out. In order to classify and label shots into the proposed narrative categories we analyze both the audio and visual content in the stream.

**3.1. Visual Content Analysis**

Shots have been long recognized as the basic organizational units for video analysis. In this work, we use the Webflix software [1] to detect shot boundaries based on cuts and dissolves.

**Constructing homogeneous shot indices:** In a study of film editing, Brandt [3] notes that the average length for an audience to adjust to a new cut is about 3 seconds. Thus, after merging the cuts and dissolve indices, any shot whose length is less than 3 seconds is combined with its shorter neighbouring shot.

**Face detection and tracking within a shot:** An important characteristic of on-screen narration distinct from others, is the frequent appearance of the narrator. Thus, our primary objective from the visual analysis is to design features that can best detect and examine faces in formal pose appearing in shots. At the simplest level, one can simply exhaustively search for faces in every single frame that makes up the shot. However, this approach is computationally expensive. To avoid this situation, we perform face detection on two different sequences of frames extracted from each shot. The first involves detecting faces on extracted based on the change in visual content of the shot, and the second is on a sequence constructed by selecting frames at fixed positions within the shot.

**Face detection in the keyframe sequence:** The keyframe sequence is computed in the following way. Let \( F = \{ f_1, f_2, \ldots, f_n \} \) be a sequence of frames making up a shot. The sequence of keyframes, \( K = \{ k_1, k_2, \ldots, k_m \} \) (\( m \leq n \)) is inductively computed as follows. Initially, \( k_1 \) is set to \( f_1 \). At the inductive step \( k_{i+1} \) is assigned to the first frame \( f_k \) such that \( \Delta(f_{k_i}, f_k) > T \) for \( l_1 < t \leq n \) where \( \Delta(f_{k_i}, f_k) \) is the histogram difference function and \( l_j \) is the position of \( k_j \) in \( F \). The \( i^{th} \) keyframe essentially represents the visual content of the shot till the next \( (i+1)^{th} \) keyframe. In the case of the shot being static, there will be only one keyframe, which will be the first frame of the shot.

Let \( F(.) \) be the face detector function such that \( F(k) \) has the value of 1 if the frame \( k \) contains a face and 0 otherwise. We define the first metric \( K^\oplus \) for measuring face content in a shot as:

\[
K^\oplus = \frac{\sum_{i=1}^{m} F(k_i)}{m}.
\]
Face detection in a fixed sequence of images: In the keyframe section, face detection is only performed when a significant change in visual content is noticed. Here we perform face detection on a sequence of images selected at fixed positions. Assume that \( \{t_1, t_2, \ldots, t_l\} \) is a series of fixed positions within the shot and \( \{p_1, p_2, \ldots, p_l\} \) are the images at these positions respectively. We define second feature \( T_\Theta \) as:

\[
T_\Theta = \sum_{i=1}^l F(p_i)
\]

where \( F(.) \) is the face detector defined previously. In this work, we have chosen 10 frames equally spaced across the shot.

Features extracted from the sizes of detected faces: In direct narration shots, the narrator is often captured using a close-up, and therefore the area of the detected face is generally larger than that in assistive narration shots. To capture this fact, the areas of detected faces are measured. For either of the keyframe or selected sequence of images, assume \( x \) is the number of images where a face is detected and \( \{u_1, u_2, \ldots, u_x\} \) are the areas of the faces in these frames (if there is more than two faces detected in one frame, their average will be used). The face area feature is computed as \( \sum_{i=1}^x u_i/(x \times A) \) where \( A \) is the area of the whole frame.

3.2. Audio Content Analysis

In order to classify a shot into direct narration, voice over, footage/interview, etc., the information from the faces alone does not suffice. We further examine the audio track as the supplementary source. In this analysis, the audio signal extracted from each shot is labelled as one of the following: shot completely dominated by speech (V), no speech (N) or a mixture of the two (M). To achieve this, a two-pass process is proposed.

Audio labels & voice connectivity feature: First, the audio track is extracted from the video stream and is split into a series of small clips. The nature of education and training videos is to teach and to give instructions, and therefore vocal speech dominates the audio track. This fact leads us to using an unsupervised approach to classify the audio signal using clustering. All audio clips are clustered into two classes and clips belonging to the dominating class will be considered as speech and the others are designated as non-speech.

In the second pass, results from the clustering process are used to classify the audio track. Let \( B \) be the total number of audio clips in a shot and \( N_a \) be number of clips classified as speech, we define the degree of voice activity \( V_a \) as \( \frac{N_a}{B} \).

An audio label \( L^* \) for that shot is then given based on the following rule:

\[
L^* = \begin{cases} 
V & \text{if } V_a \geq T_{high} \\
N & \text{if } V_a \leq T_{low} \\
M & \text{otherwise.}
\end{cases}
\]  

In this work, we choose \( T_{high} = 0.65 \) and \( T_{low} = 0.35 \). These figures are empirically determined.

We further observe that the flow of vocal speech delivered in a narration or voice over section is more consistent than in an interview or a footage. To exploit this fact we propose a simple metric that roughly measures the 'connectivity' of voice (speech) clips. Let \( \{a_1, a_2, \ldots, a_p\} \) be the sequence of audio labels that make up the audio track of shot \( S \) and let \( C(a_i, a_j) \) have a value of 1 if both \( a_i \) and \( a_j \) is labeled as V and 0 otherwise. The voice connectivity feature is then given as:

\[
C_{\text{voice}} = \sum_{i=1}^{p-1} C(a_i, a_{i+1}) 
\]

Essentially, it is the number of contiguous speech-dominant clips normalized by the shot length.

4. Experimental Results

Results are reported from four training videos, House Keeping (16 mins), Safe Guarding (10 mins), Safe Approach (18.5 mins), Think Safe (23.3 mins) and two education videos: Statistics Lesson (27.4 mins) and Geometry Lesson (20.1 mins). This results in a total of 115.2 minutes of video for analysis. The Webflix software has detected a total of 580 shots across these six videos. These shots were hand-labelled as one of DD, DN, AN, UV, IV or LF to form the groundtruth for comparison.

Using the CMU algorithm for face detection [8], we extract face features as proposed in Section 3.1. Each audio track is re-sampled at 44.1kHz with mono channel and divided into a sequence of 1.5s clips with an overlap of 1.0s. Audio features are then extracted using the energy in seven sub-bands of the discrete wavelet transform (DAUB4) with 6 levels of decomposition [7] and 14-order cepstral coefficients using Linear Prediction Coding coefficients. This forms a feature vector of 21 dimensions for each audio clip. The K-means clustering algorithm is then used to form two clusters: speech-dominated and non-speech. Each clip is then assigned one of these two cluster labels. The audio classification scheme proposed in Section 3.2 is used to label the audio track of each shot with either V, N, or M. The voice connectivity metric \( C_{\text{voice}} \) is also computed. This results in a total of six features for each shot, including four from the face detection results and two from audio track analysis.

In the first experiment, we use C4.5 to generate a decision tree for determination of final narrative structure labels. The result from this machine learning process achieves an accuracy as high as 91.6% with the confusion matrix shown in Fig. 2(a). It is to be noted that the size of the tree is also relatively small (43 nodes).

The classification results for DD, DN, AN and UV are high with a low false positive rate of 0.16, 0.09, 0.14 and
0.004 respectively. More importantly, two important sections in narrative structuralization namely, direct and assistive narration sections have been accurately separated from others. Further examination of the videos shows that misclassification between on-screen narration and voice overs is generally due to the following scenarios:

- A voice over shot with the presence of many faces such as people in a party is misclassified with on-screen narration.
- False positives or negatives in face detection will lead to a misclassification between the two.

The new confusion matrix therefore yields an accuracy of 97.6%. The next level DT-rules separate UV, IV with an accuracy of 84.3%. This results in an accuracy about 84.7% for the whole system.

5. Conclusion and Future Work

In this paper, we address the problem of identifying the narrative structure in education and training videos for nonlinear media presentation in e-learning applications. We propose a set of audio visual features to differentiate and classify shots into six distinct narrative structures used in education and training videos. Important structures in these videos have been recognized with a high accuracy. The resulting hierarchical DT-classification system achieves an overall approximate accuracy of 84.7%. The results of this automated narrative structuralization will be used in our continuing work to extract higher level semantics and to discover interesting patterns in instructional videos relating to significant events and story line.

References