Detection of Documentary Scene Changes by Audio-Visual Fusion

Atulya Velivelli[1], Chong-Wah Ngo[2], and Thomas S. Huang[1]

[1] Beckman Institute for Advanced Science and Technology, Urbana
{velivell,huang}@ifp.uiuc.edu
[2] City University of Hong Kong, Hong Kong
cwngo@cs.cityu.edu.hk

Abstract. The concept of a documentary scene was inferred from the audio-visual characteristics of certain documentary videos. It was observed that the amount of information from the visual component alone was not enough to convey a semantic context to most portions of these videos, but a joint observation of the visual component and the audio component conveyed a better semantic context. From the observations that we made on the video data, we generated an audio score and a visual score. We later generated a weighted audio-visual score within an interval and adaptively expanded or shrunk this interval until we found a local maximum score value. The video ultimately will be divided into a set of intervals that correspond to the documentary scenes in the video. After we obtained a set of documentary scenes, we made a check for any redundant detections.

1 Introduction

A rapid increase in digital video data over the past few years has given rise to importance for video data indexing. The first step towards alleviating this problem is organizing a video into semantically tractable units called *scenes*. A scene is defined as a collection of shots that occur at the same location or that are temporally unified. A *shot* is defined as an unbroken sequence of frames taken from one camera. In [1, 2], scene change is detected by extracting visual features such as chromatic edit detection feature and color feature from the video (image sequence). Recently there has been interest in using both audio and visual information for scene change detection. In [3], different audio classes are detected sequentially, and this information is later combined with the probability value for a visual cut detection. In [4], scene breaks are detected as audio-visual breaks, where each break is based on dissimilarity index values calculated for audio, color and motion features. Even in this case there is no specific audio-visual fusion strategy.

In [5], they first obtain a set of audio scenes and a set of video (visual) scenes by using different criteria for audio scene break and video (visual) scene break. These breaks are then merged to obtain audio-visual scene breaks.

E. M. Bakker et al. (Eds.): CIVR 2003, LNCS 2728, pp. 227–238, 2003.

We introduce the concept of a documentary scene and use an audio-visual score value (which is a weighted combination of an audio score and a video score) to divide the video into a set of documentary scenes. The audio score and the visual score are generated by procedures evolved out of the observations that we make on the video data. The video data that we used for making observations and for experimenting is from the NIST Special Database 26.

The rest of paper is organized as follows. Section 2 introduces the concept of documentary scene and then outlines the major steps of our approach. Section 3 describes the generation of visual label patterns while Section 4 presents a maximum-likelihood based method to generate audio class labels. Section 5 describes our audio-visual fusion strategy. Section 6 proposes an adaptive scheme and redundancy check to detect documentary scene changes. Finally, Section 7 presents the experimental results and Section 8 concludes this paper.

2 Overview of Our Approach

In this section we define the concept of a documentary scene, which was inferred by analyzing the video data. Critical observations made on the data will also be stated. The proposed approach in this paper is then presented based on these observations.

2.1 Concept of a Documentary Scene

The NIST videos show contiguous shots with little visual similarity while the topic or the semantic context described by the audio remains same. Hence we conclude that a documentary video can be modelled as a union of several semantically tractable topic level units known as *documentary scenes*, where there is a semantic correlation between the audio component and the visual component of videos. The common observations from documentary videos include:

1. Within a documentary scene, similar video frames occur either contiguously or with a temporal separation of some dissimilar frames; hence, the total number of similar frames within an interval is a measure for a documentary scene.
2. In most cases, the audio class label at the beginning of a documentary scene is same as the audio class label at the end of a documentary scene.
3. In some cases the audio component has less information to contribute, while visually there is a strong semantic correlation in the background image of the video frames. For example as shown in Figure 1, the semantic context is the discussion on a new project. The selected frames in Figure 1 show the machine drawings associated with the project in background.
4. In most cases the visual pattern has a counterpart audio pattern. An audio-visual sequence shown below explains this observation.

$$\begin{array}{rl} \text{audio class :} & \text{speech} \leftarrow \text{speech+siren} \leftarrow \quad \text{speech} \\ \text{visual sequence :} & \text{aircraft} \leftarrow \quad \text{hanger fire} \quad \leftarrow \text{officer speaking} \end{array}$$

Fig. 1. A few frames with a similar visual background

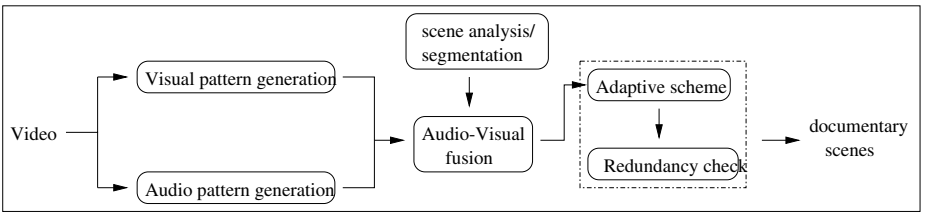

Fig. 2. Proposed approach for detecting documentary scenes

2.2 Proposed Approach

Based on the above observations, we propose a scheme, as illustrated in Figure 2, for detecting documentary scenes. Given a video, the proposed approach first generates a visual pattern (observation 1) and an audio pattern (observations 2, 4) respectively based on similarity measures. Scene analysis (observation 3) will be conducted for foreground and background segmentation[1]. The collected data from visual, audio and scene analysis will be integrated for audio-visual fusion. The scene change detector is composed of two main steps: adaptive scheme and redundancy check.

3 Generating Visual Label

In this section, we present the method for generating a visual label pattern. We first perform shot boundary detection and then select the keyframes. Using the auto-correlogram method described in [6], we label the keyframes extracted from the video.

3.1 Selection and Labelling of Keyframes

To label the video keyframes, we first detect the shot boundaries using the technique in [7]. Each shot is composed of several temporal units. We extract a video frame at the middle of a temporal unit (30 frames) as a keyframe. The video frame rate is 30 frames/s. The similarity measure for labelling the

[1] In this paper, we will not present the details of scene analysis, interested readers can refer the details of our approach in [9, 10].

keyframes is based on the color auto-correlogram [6] of keyframes. A critical part of this algorithm is the appropriate selection of similarity threshold η'. In our approach, we employ Neyman-Pearson hypothesis to determine η', which will be elaborated later.
To label a keyframe $\mathcal{I}^{(i)}$, we find $\beta_i = \text{argmin}_{1 \leq j < i} \left| \mathcal{I}^{(i)} - \mathcal{I}^{(j)} \right|$. If $\left| \mathcal{I}^{(i)} - \mathcal{I}^{(\beta_i)} \right| < \eta'$, we label it with index β_i ($\mathcal{I}^{(i)}$ is considered similar to $\mathcal{I}^{(\beta_i)}$), else with i ($\mathcal{I}^{(i)}$ is considered dissimilar to preceding frames). Where $\left| \mathcal{I}^1 - \mathcal{I}^2 \right|$ denotes the auto-correlogram distance measure between $\mathcal{I}^1$ and $\mathcal{I}^2$. The example below shows the original keyframe index and the resulting keyframe labels after applying the algorithm:

$$\begin{aligned} Keyframe\ index &\leftarrow 1\ 2\ 3\ 4\ 5\ 6\ 7\ 8\ 9\ 10\ 11 \\ Keyframe\ label &\leftarrow 1\ 2\ 2\ 4\ 5\ 6\ 2\ 4\ 9\ \ 9\ \ 5 \end{aligned}$$

The above labelling pattern indicates that the keyframe with index 2 is similar to the keyframe with index 3; hence the keyframe with index 3 also gets a label 2, while the keyframe with index 4 is not similar to any of the 3 preceding keyframes and hence gets a label 4.

3.2 Threshold Selection

We use the Neyman-Pearson hypothesis testing, which is explained in detail in [8], to determine the value of threshold η'. We first calculate the distances between all combinations of the N images and we assume that the distances can be modelled as two Gaussian distributions, corresponding to the similar and the dissimilar images . The value of threshold η' is

$$\eta' = \sigma\ \phi^{-1}(\alpha) + \mu \tag{1}$$

where, σ^2, μ are the variance and mean of the Gaussian distribution with a larger variance. We assume that the larger variance Gaussian corresponds to the dissimilar images.

$$\phi(x) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{x} e^{\frac{-t^2}{2}}\, dt \tag{2}$$

α is the false alarm tolerance and $\phi^{-1}(x)$ denotes the inverse of $\phi(x)$.
In our case we have decided a false alarm tolerance of 5%, as it is a good trade off between false alarm and the detection over a wide range of data; hence, $\alpha = 0.05$. Substituting $\alpha = 0.05$ in Equation (1), we obtain a corresponding value for η'.

4 Generating Audio Label

In this section we describe the generation of an audio class label pattern. We select 3 videos from the total of 8 in the NIST database for training purposes. The audio training data is annotated into 6 audio classes.

4.1 Audio Data Models

We manually segment the training data containing 3 different audio files that are extracted as the audio stream of the video files, and then label each segment as belonging to one of 6 different classes: speech, speech + music, music, speech + noise, noise, and silence. We employ a Gaussian mixture model [11] for audio class identification. For each of these six audio classes, we first calculate a 30-dimensional audio feature vector. The audio feature vector consists of 13 mel frequency cepstrum coefficients. Another 13 components correspond to the first derivative of the original 13 coefficients. Then we have the short time energy, the short time zero crossing rate, and their first derivatives. We extract this feature vector over all the audio files meant for training. Then we estimate a Gaussian mixture model M_i for the audio class i which is defined by the parameter set $P_i(\mu_i, \Sigma_i, w_i)$, consisting of the mean vectors μ_i, covariance matrices Σ_i, and mixture weight vectors w_i.

4.2 Class Label Identification

The goal of audio class label identification is to find the model M_i that best explains each frame of the test data represented by a sequence of N audio frames $\{f_n\}_{n=1,..,N}$. The audio frame rate is 100 frames/s. Maximum-likelihood criterion is used for classification. The class label c_n for frame f_n is $c_n = \arg\max_i \log\ p_i(f_n|\mu_i, \Sigma_i, w_i)$, where p_i is the probability of a frame belonging to the Gaussian mixture model $P_i(\mu_i, \Sigma_i, w_i)$.

Similar to the visual data labelling explained in the previous section, we also generate a counterpart audio class label pattern for N frames $\{f_n\}_{n=1,..,N}$, as $\{c_n\}_{n=1,..,N}$.

5 Audio-Visual Fusion

In this section, we first define some terms that will be used for audio-visual fusion with respect to the interval $[s \quad t]$, where s and t denote the indices of video keyframes. In the case of audio definitions we map the index to corresponding audio frame. Appropriate weights are then statistically selected to linearly combine audio and visual information for detecting documentary scene changes.

Visual Similarity Count: We denote the set of similar keyframes as $VS = \{V_i | V_i = V_j, j \neq i\}$, the normalized visual similarity count $vsc(s,t)$ is

$$vsc(s,t) = \frac{|VS|}{(t-s+1)} \tag{3}$$

where $|.|$ denotes the size of a set.

Audio Similarity Count: AC_m is the number of audio frames $\in$ class m, the normalized audio similarity count $asc(s,t)$ is

$$asc(s,t) = \frac{\max_{1 \le m \le K} AC_m}{Number\ of\ audio\ frames\ in\ this\ interval} \tag{4}$$

where K is the total number of audio classes.

Audio Delta Function: This function returns 1 if the audio class of the frames in the begining of the interval is same as those in the end. S_i is the log-likelihood of the audio class of the initial r audio frames, denoted as $A_s = \{a_n\}_{n=s,..,s+r-1}$ belonging to the i^{th} audio class model. We have

$$S_i = \log\, p_i(A_s|\mu_i, \Sigma_i, w_i) = \sum_{n=s}^{s+r-1} \log\, p_i(a_n|\mu_i, \Sigma_i, w_i). \tag{5}$$

Furthermore, let T_i denotes the log-likelihood of the audio class of the ending r frames, denoted as $A_t = \{a_n\}_{n=t-r+1,..,t}$ belonging to the i^{th} audio class model. We have

$$T_i = \log\, p_i(A_t|\mu_i, \Sigma_i, w_i) = \sum_{n=t-r+1}^{t} \log\, p_i(a_n|\mu_i, \Sigma_i, w_i). \tag{6}$$

Finally, the audio delta function $\delta_a(s,t)$ is defined as

$$\delta_a(s,t) = \delta(L_s - L_t). \tag{7}$$

where $L_s = \arg\max_i S_i$ is the label corresponding to the initial r frames, and $L_t = \arg\max_i T_i$ is the label corresponding to the ending r frames.

Visual Delta Function: This function $\delta_v(s,t)$ returns 1 if the video frames between the keyframes within the index interval $[s \quad t]$ have a similar background. Figure 3 shows the detailed algorithm of this function. Each common background scene is represented as $back_i$, and x_i is used to represent the start of $back_i$, while y_i represents its end. The details for finding the common background scenes can be found in [9].

5.1 Audio-Visual Score

The audio score $S_a(s,t)$ is the sum of the audio similarity count, and the audio delta function :

$$S_a(s,t) = asc(s,t) + \delta_a(s,t). \tag{8}$$

The visual score $S_v(s,t)$ also is the sum of the visual similarity count, and the visual delta function :

$$S_v(s,t) = vsc(s,t) + \delta_v(s,t). \tag{9}$$

Input: $\{back_i\}_{i=1,\ldots,N}$ the set of common background scenes, interval $[s \quad t]$, tolerance value ϵ
Output: $\delta_v(s,t)$
1. **for** $i = 1, \ldots, N$ **do:**
  1.1 **Initialize**: $\delta_v(s,t) \leftarrow 0$
  1.2 **if** $|x_i - s| < \epsilon$ and $|y_i - t| < \epsilon$
  1.3 **then** $\delta_v(s,t) \leftarrow 1$

Fig. 3. Algorithm for visual delta function

The audio-visual score [12], $S_{av}(s,t)$ is a weighted combination of the audio and visual score.

$$S_{av}(s,t) = w_a \; S_a(s,t) + w_v \; S_v(s,t). \tag{10}$$

where w_a is the audio weight, w_v is the visual weight, and $w_a + w_v = 1$.

5.2 Selection of Mixture Weights

Since w_a and w_v can affect the result of audio-visual score, we apply a statistical method to approximate their optimal values. We learn the optimal value of the mixture weight ω denoted as ω^{opt} from the training data by minimizing a cost function $C(\omega)$ which is the smoothed recognition error rate [13]:

Input: $L + 1$, Δ, *entries*.
- $L + 1$: The initial size of the interval $[s \; t]$, which will be either expanded or shrunk
- Δ: The step size by which the interval is each time expanded or shrunk as shown in Figure 5
- *entries*: The index of the last keyframe in the video as shown in Figure 5

Output: $\{documentary_i\}_{i=1,\ldots,N}$ a set of documentary scenes

1. Initializing the begining and end of the interval and its index:
$s \leftarrow 1$, $t \leftarrow L + 1$ and $i \leftarrow 1$

2. while $t <$ *entries*, perform the below steps:
2.1 **if** $S_{av}(s,t) > \{S_{av}(s, t + \Delta)$ and $S_{av}(s, t - \Delta)\}$ (indicates a local maximum)
2.2 $documentary_i \leftarrow [s \quad t]$ (indicates the detection of documentary scene)
2.3 $i \leftarrow i + 1$ (counting each detected documentary scene)
2.4 we re-initialize the interval: $s \leftarrow t + 1, t \leftarrow s + L - 1$
2.5 **else if** $S_{av}(s, t + \Delta) > S_{av}(s, t - \Delta)$
2.6 we expand the interval by Δ: $t \leftarrow t + \Delta$
2.7 **else** we shrink the interval by Δ: $t \leftarrow t - \Delta$

Fig. 4. Algorithm for adaptive scheme

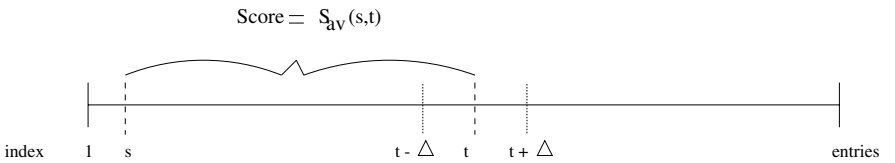

Fig. 5. The audio-visual score $S_{av}(s,t)$ evaluated over the interval $[s\ t]$ is compared with the score evaluated over $[s, t+\Delta]$ and $[s, t-\Delta]$

$$C(\omega) = \frac{1}{R}\sum_{i=1}^{R} \frac{1}{1+e^{\{-\xi\ S^i_{av}(\omega)\}}}\ , \xi > 0 \tag{11}$$

where

$$S^i_{av}(\omega) = \omega\ S^i_a + (1-\omega)\ S^i_v. \tag{12}$$

S^i_a is the audio score for the i^{th} training clip, and S^i_v is the visual score for the i^{th} training clip. This cost function is evaluated by selecting R clips from the training videos such that they do not correspond to a documentary scene. Minimizing the cost function $C(\omega)$ with respect to ω, while substituting a positive value for ξ results in $w^{opt} = 0.7$. Hence, the optimal value of audio weight $\omega_a = 0.7$ and the optimal value of the visual weight $\omega_v = 0.3$.

6 Detecting Documentary Scene Change

6.1 Adaptive Scheme

In the adaptive scheme, the audio-visual score within an interval is first evaluated. This interval will be adaptively expanded or shrunk until a local maximum is found. The detailed algorithm can be found in Figure 4.

6.2 Redundancy Check

The visual delta function and the audio delta function tend to be the cause of some redundant detections. In fact, certain redundant detections can be eliminated by the careful investigation of neighbouring documentary scenes. To cope with this problem, we merge neighbouring documentary scenes on the basis of a new score. The audio-visual merging score $S^M_{av}(s,t)$ is a weighted combination of the audio similarity count and the visual similarity count:

$$S^M_{av}(s,t) = \omega^{opt}\ asc(s,t) + (1-\omega^{opt})\ vsc(s,t) \tag{13}$$

The details of algorithm for redundancy check can be found in Figure 6.

Table 1. Results at the end of the adaptive scheme stage for Δ=3

video	$\Delta = 3$			
	Duration	Human Detections	Machine Detections	Number of hits
1	489 s	29	27	21
2	382 s	13	22	6
3	523 s	11	14	8
4	743 s	22	40	9
5	493 s	14	28	5

Table 2. Results at the end of check for redundancy stage for Δ=3

video	$\Delta = 3$			
	Duration	Human Detections	Machine Detections	Number of hits
1	489 s	29	27	21
2	382 s	13	12	10
3	523 s	11	14	8
4	743 s	22	22	17
5	493 s	14	16	9

7 Experimental Results

The experimental data used is from NIST Special Database 26. Out of the eight videos, three were used for training while the remaining five were used for testing. We use Figure 7 to depict a detected documentary scene. In Figure 7, the top row, from left to right, shows the award, a performer, and an instrument. The bottom row shows President Clinton, the audience, and the award being handed over with an industrial lab in the background. This documentary scene is a description of the Malcolm Baldrige Quality Award. Although there are many visual changes throughout this documentary scene, the underlying semantic context remains the same.

Input: Result of the adaptive scheme $\{documentary_i\}_{i=1,..,N}$
Output: The merged output set $\{documentary_i\}_{i=1,..,N}$

1. **while** $i < N$, **do**:
   1.1 **if** $S^M_{av}(s_i, t_{i+1}) > \{S_{av}(s_i, t_i)$ and $S_{av}(s_{i+1}, t_{i+1})\}$
   1.2 **merge** $[s_i\ t_i]$ and $[s_{i+1}\ t_{i+1}] : documentary_i \leftarrow [s_i\ t_{i+1}]$
   1.3 **adjust** the index and length: $i \leftarrow i-1$, $N \leftarrow N-1$
   1.4 **else** go to the next documentary scene: $i \leftarrow i+1$

Fig. 6. Algorithm for redundancy check

Fig. 7. A documentary scene from test video 2

We adopt recall and precision for the performance evaluation:

$$recall = \frac{Number\ of\ hits}{Human\ Detections} \tag{14}$$

$$precision = \frac{Number\ of\ hits}{Machine\ Detections} \tag{15}$$

where, hit is defined as the detection by both human subjects and machine. Tables 1 and 2 show the experimental results, by the adaptive scheme and redundancy check, for Δ=3. As indicated in Tables 1 and 2, number of hits in videos 2, 4 and 5 are increased after the redundancy check. The number of hits remain the same for videos 1 and 3 before and after applying the redundancy check. The recall and precision of the proposed approach on the five tested videos can be found in Table 3.

We further investigate the effectiveness of the approach by varying the parameter Δ. The recall-precision values for $\Delta = 3$ are constantly better than the values for $\Delta = 4$ or 5, as indicated in Table 3.

Table 3. A comparison of recall - precision values for $\Delta = 3, 4, 5$ after making a check for redundancy

video	$\Delta = 3$		$\Delta = 4$		$\Delta = 5$	
	recall	precision	recall	precision	recall	precision
1	.72	.78	.59	.71	.45	.57
2	.77	.83	.54	.64	.62	.80
3	.73	.57	.55	.40	.45	.36
4	.77	.77	.45	.56	.41	.45
5	.64	.56	.50	.54	.43	.46

8 Conclusion

We have presented a scheme out of certain observations that we made on the audio-visual characteristics of documentary videos. This scheme is basically a two stage process. In the first stage, we find a set of documentary scenes by a weighted fusion of the audio score and the video score. In the second stage, we make a check for any redundant detections, and, if any, we merge those documentary scenes. It is observed through experiments that in the cases where the end of the adaptive scheme itself gives optimal number of hits, even after making a check for redundancy they remain unchanged. However, in the cases where there is actually a redundancy in detection, merging neighbouring documentary scenes actually increases the number of hits. This scheme has been successful in detecting documentary scene changes, with each of them having a common underlying semantic context. Future work would focus on how to identify this semantic context probably by classifying them into few learnt categories.

References

[1] A. Hampapur, R. Jain , and T. Weymouth, "Digital video segmentation," in *Proceedings of ACM Multimedia ,* San Francisco CA, October 1994, pp. 357-364. 227

[2] J. R. Kender and B. L. Yeo, "Video scene segmentation via continuous video coherence," in *CVPR ,* Santa Barbara CA, June 1998. 227

[3] C. Saraceno and R. Leonardi, "Audio as support to scene change detection and characterization of video sequences," in *Proceedings of ICASSP,* vol 4, 1997, pp. 2597-2600. 227

[4] J. Huang, Z. Liu, and Y. Wang, "Integration of audio and visual information for content-based video segmentation," in *ICIP*, Chicago, 1998. 227

[5] H. Sundaram and S.-F. Chang, "Video scene segmentation using audio and video features," in *ICME,* New York, July 28-Aug 2, 2000. 227

[6] J. Huang, "Color-spatial image indexing and applications," Ph.D. dissertation, Cornell University, 1998. 229, 230

[7] 229
C. W. Ngo, T. C. Pong, and R. T. Chin, "Video partitioning by temporal slice coherency," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 11, Aug 2001, pp. 941-953.

[8] H. V. Poor, *An Introduction to Signal Detection and Estimation.* New York: Springer, 2nd ed., 1994. 230

[9] 229, 232
C. W. Ngo, T. C. Pong and H. J. Zhang, "Motion-based video representation for scene change detection," *International Journal of Computer Vision*, vol. 50, No. 2, Nov, 2002.

[10] 229
C. W. Ngo, "Motion Analysis and Segmentation through Spatio-temporal Slices Processing," *IEEE Trans. on Image Processing*, Feb, 2003.

[11] D. A. Reynolds and R. C. Rose, "Robust text-independent speaker identification using Gaussian mixture speaker models," *IEEE Transactions on Speech and Audio Processing,* vol. 3, no. 1, pp. 72-83, Jan 1995. 231

[12] B. Maison, C. Neti, and A. Senior, "Audio-visual speaker recognition for video broadcast news: some fusion techniques," in *IEEE Multimedia Signal Processing Conference (MMSP99),* Denmark, Sept 1999. 233

[13] L. Rabiner and B. H Juang, *Fundamentals of speech recognition.* New Jersey: Prentice Hall International Inc, 1993. 233