Movie Analysis with Emphasis to Dialogue and Action Scene Detection

Emmanouil Benetos, Spyridon Siatras, Constantine Kotropoulos, Nikos Nikolaidis, and Ioannis Pitas

Department of Informatics, Aristotle Univ. of Thessaloniki, Box 451, Thessaloniki 541 24, Greece
{empeneto, siatras, costas, nikolaid, pitas}@aiia.csd.auth.gr

2.1 Introduction

Movies constitute a large portion of the entertainment industry, as over 9.000 hours of video are released every year [4]. As the bandwidth available to users increases, online movie stores – the equivalent of popular digital music stores – are emerging, providing the users an opportunity to build large personal movie repositories. The convenience of digital movie repositories will be in doubt, unless multimedia data management is employed for organizing, navigating, browsing, searching, and viewing multimedia content. Semantic content-based video indexing and annotation offer a promising solution for the efficient digital movie management.

Semantic video indexing aims at extracting, characterizing, and organizing video content by analyzing the visual, aural, and textual information sources of video. The need for content-based audiovisual analysis has been realized by the MPEG committee, leading to the creation of the MPEG-7 standard [1]. The current approaches for automatic movie analysis and annotation mostly focus on the visual information, while the audio information receives little or no attention. However, the integration of the audio information with the visual one can improve semantic movie content analysis.

The predominant approach to semantic movie analysis is to initially extract some low-level audiovisual features (such as color and texture from the video or energy and pitch from the audio), derive some mid-level entities (such as video shots, keyframes, appearance of faces and audio classes), and finally understand video semantic content by analyzing and combining these entities. A hierarchical video indexing structure is displayed in Figure 2.1.

Movie analysis aims at obtaining a structured organization of the movie content and understanding its embedded semantics like humans do. It has been handled in different ways, depending on the analysis level and the assumptions on the film syntax described in Section 2.2. Most movie analysis efforts concentrate on movie scene or shot detection, while other works focus on the separation of dialogue and non-dialogue scenes. Several efforts have been made for dialogue scene detection,
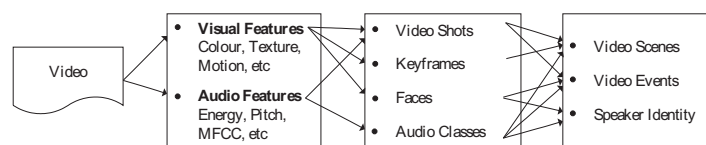
Fig. 2.1. Generic video indexing structure, where arrows between nodes indicate a causal relationship (adapted from [29]).

some efforts have concentrated to action scene detection, and limited work has also been performed for movie genre categorization.

In this chapter, we put emphasis on the detection of dialogue and action scenes in a video sequence using visual and audio cues. Dialogue and action scenes can be interpreted as high-level semantic features that are appropriate for inclusion in more sophisticated organization, browsing, and retrieval techniques applied to movies and television programs. Their successful detection provides significant semantic information for the video sequence and it is especially useful for managing certain classes of video content. For instance, a dialogue detection system can enrich a generic video retrieval/browsing system enabling the detection and retrieval of scenes where a dialogue is taking place. In conjunction with face or speaker identification methods, it can also be able to identify the scenes where two (or more) particular persons are conversing. Furthermore, a quantitative comparison between the duration of dialogue scenes and the duration of non-dialogue scenes in a movie can be used for movie genre classification. As far as action scene detection is concerned, it can be applied to a film summarization system, where users can quickly and easily browse the content of a film. Dialogue and action scenes follow specific patterns, concerning their constituent shots, that makes their detection in a video sequence feasible.

The main aim of this chapter is to review the research related to dialogue and action scene detection in order to assess qualitatively and quantitatively the various methods. These methods can be broadly classified as video-only, audio-only, or audiovisual ones. A second classification distinguishes them to deterministic methods and probabilistic ones.

The remainder of the chapter is organized as follows. In Section 2.2, the basic principles of film structure and video editing rules for constructing dialogue and action scenes are discussed. The most commonly employed figures of merit are described in Section 2.3, along with the datasets utilized in movie analysis literature. Sections 2.4 and 2.5 review the basic principles and state-of-the-art algorithms for visual-only, audio-only, and audiovisual dialogue-action scene detection, respectively. Conclusions are drawn in Section 2.6.

2.2 Film Syntax Basics

A movie or television program can be divided into *shots* and *scenes*. A shot is defined as a single continuous camera recording, whereas a scene consists of a concatenation

of shots, which are temporally and spatially cohesive *in the real world*, however not necessarily cohesive in the projection of the real world on film [3, 12]. Rasheed gives a similar definition, stating that similar shots of a movie must be combined in order to form a scene or a *story unit* [42]. The notion of *computable scenes* (c-scenes) is proposed to characterize scenes that can be reliably computed using only low-level features [46]. They are derived by fusing information from audio and visual boundary detectors. Another term that has been proposed is the *logical story unit* (LSU) which is a high-level temporal movie segment characterized by a single event (dialog, action scene). The LSU segmentation is based on the investigation of visual information and its temporal variations in a video sequence. A movie can be modeled as a sequence of states and events organized in space and time by creating a *state graph* representing the film story [47].
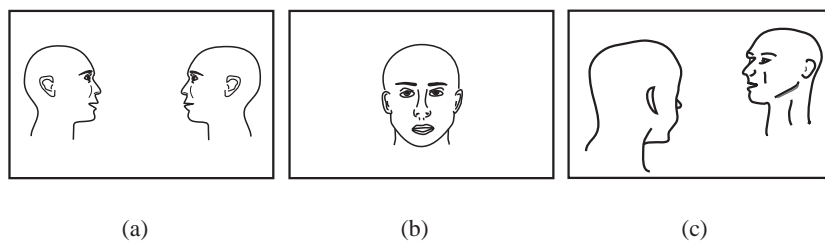

(a)   (b)   (c)

Fig. 2.2. Sample frames from shots usually employed in dialogue scenes. (a) Side view of two persons. (b) Frontal view of one person. (c) Over-the-shoulder shot (a shot of one person taken from over the shoulder of another person).

As far as dialogues are concerned, a *dialogue scene* can be defined as a set of consecutive shots, which contain conversations of people [3, 22]. In Figure 2.2, frames from shots broadly employed in dialogue scenes are depicted. In such a scene, the persons who participate in the dialogue will be present either one at a time (Figure 2(b)) or all in the same image frame, in frontal or side view (Figures 2(a), 2(c)). In general, a dialogue scene includes a significantly repetitious structure of shots that depict the dialogue participants. However, a dialogue scene might include shots that do not contain any conversation or do not even depict a dialogue participant. For example, shots of other persons or objects might be inserted in the dialogue scene. In addition, the shot of the speaker may depict the rear view of his head. Evidently, these shots add to the complexity of the dialogue detection problem. According to Chen, the elements of a dialogue scene are: the people, the conversation, and the location where the dialogue takes place [9]. The basic shots in a 2-person dialogue scene are:

- Type A shot: Shot of actor A's face.
- Type B shot: Shot of actor B's face.
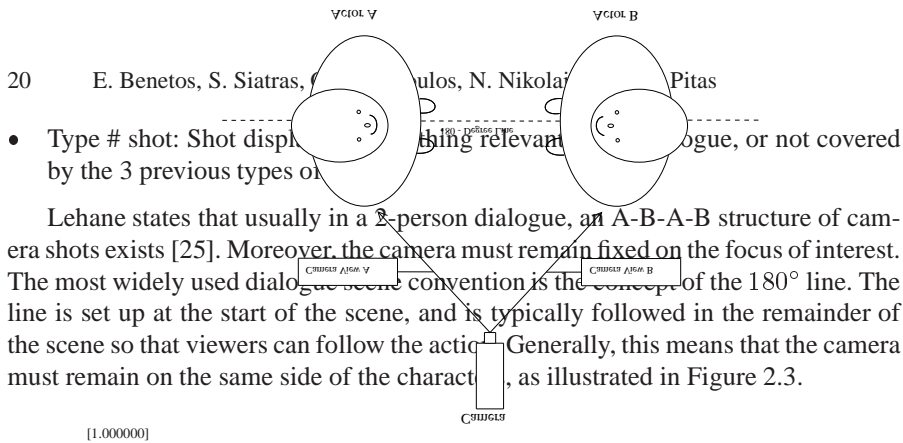- Type C shot: Shot with both faces visible.

- Type # shot: Shot disp... ...ing relevant ...ogue, or not covered by the 3 previous types o...

Lehane states that usually in a 2-person dialogue, an A-B-A-B structure of camera shots exists [25]. Moreover, the camera must remain fixed on the focus of interest. The most widely used dialogue scene convention is the concept of the 180° line. The line is set up at the start of the scene, and is typically followed in the remainder of the scene so that viewers can follow the action. Generally, this means that the camera must remain on the same side of the characters, as illustrated in Figure 2.3.

[1.000000]

Fig. 2.3. The concept of the 180° line (adapted from [25]).

Concerning film syntax for action sequences, Lehane mentions that it is a general concept meant to keep the audience's attention at all times [24]. The objective of the director is to excite the viewer by a rapid succession of shots, strong movement within shots, and variation in the length of shots. Pans, tilts, and zooms are used to follow characters moving within shots. According to Chen, the rules governing the actor arrangement and camera placement in simple action scenes are the same to those for producing simple dialogue scenes, even though in action scenes actors move rapidly and cameras follow the actors [9].

A 2-person dialogue scene, from the audio point of view, can be defined as a proper alternation between two speakers [22]. Dialogues in an audio framework can be detected by using the cross-correlation function between the speaker indicator functions or their respective cross-power spectra. A set of recognizable dialogue acts, according to semantic content, based on audio analysis, is proposed in [23]: (i) Statements (ii) Questions (iii) Backchannels (iv) Incomplete utterance (v) Agreements (vi) Appreciations.

In contrast with dialogue scenes, the audio channel in an action scene usually consists of less speech and more environmental sounds or music [10]. The soundtrack of an action scene is chosen in a way to create tension and suspense to the

viewers. It is much different than the soundtrack of a dialogue scene, where, if music accompanies the dialogue, it is discrete and unobtrusive. Hence, action scenes exhibit a higher audio energy due to tense music, explosions, people fights, etc. A more detailed description of the basic principles of film syntax can be found in [7, 8].

2.3 Figures of merit and movie datasets

The most commonly used figures of merit in dialogue and action scene detection experiments are *recall* (R), *precision* (P), and F_1 *measure*, defined as

$$P = \frac{hits}{hits + false\ alarms}, \quad R = \frac{hits}{hits + misses}, \quad F_1 = \frac{2R \cdot P}{R + P}. \quad (2.1)$$

Hits are defined as correctly detected dialogue or action scenes. False alarms should not have been detected as dialogue/action scenes, but are nevertheless detected as such. Misses are defined as scenes that should have been identified as dialogue/action scenes, but were not. Other performance metrics used for the evaluation of dialogue/action scene detection algorithms are the *hit rate*, the *miss rate*, and the *false hit rate* [36]. The authors employing these figures of merit, argue that scene determination is equivalent to eliminating the shot boundaries which do not correspond to scene boundaries. The hit rate is the ratio of correctly eliminated shot boundaries plus the correctly detected scene boundaries over the number of all shot boundaries. The miss rate is the ratio of missed scene boundaries to the number of all shot boundaries. The false hit rate determines the ratio of falsely detected scene boundaries to the number of all shot boundaries. Finally, Alatan et al [2, 3, 4] employ the *shot accuracy* measure, which is defined as the ratio of correct shot assignments to the total number of shots.

The movies and TV shows used for dialogue and action scene detection are listed in Table 2.1. It should be noted that there is no common database used for dialogue and action scene detection experiments.

2.4 Visual-only and Audio-only Dialogue and Action Scene Detection

In this section, a review of the recent advances in dialogue and action scene detection techniques, using only the visual information or the aural one, will be undertaken. The features extracted from the video and audio are described, selected algorithms are examined, and their results are presented and discussed.

The proposed approaches for dialogue and action scene detection can be classified into two main categories: *deterministic* and *probabilistic* ones. Deterministic techniques exploit the repetitive structure exhibited by visually similar shots that are temporally close to each other [5, 9, 22, 24, 25, 36, 46], whereas probabilistic techniques use Hidden Markov Models (HMMs). To assign semantically meaningful scenes to model states. The video content is segmented into dialogue or action scenes using the state transitions of the HMM [19, 49].

Table 2.1. Movies and TV shows used in scene analysis and dialogue detection experiments.

Movie	Reference	Movie	Reference
MPEG-7 Data Set (CDs 20-22)[1]	[4][2][3]	Braveheart	[29]
Crouching Tiger, Hidden Dragon	[9][10]	When Harry Met Sally	[29]
Gladiator	[9][10]	Forrest Gump	[36]
Patch Adams	[9]	Groundhog Day	[36]
Analyze That	[22]	A Beautiful Mind	[42]
Cold Mountain	[22]	Goldeneye	[42]
Jackie Brown	[22]	Gone in 60 Seconds	[42]
Fellowship of the Ring	[22]	Terminator II	[42]
Platoon	[22]	Top Gun	[42]
Secret Window	[22]	Four Weddings and a Funeral	[46]
Dumb and Dumberer	[24][25][26]	Pulp Fiction	[46]
Kill Bill vol. 1	[24][25]	Sense and Sensibility	[46]
Reservoir Dogs	[24][25][26]	CNN Headline News	[53]
Snatch	[24][26]	Dr. No	[53]
American Beauty	[25][26]	Jurassic Park III	[53]
High Fidelity	[25][26]	Larry King Live	[53]
Shaft	[25]	Mission Impossible II	[53]
Life of Brian	[26]	Scream	[53]
Legends of the Fall	[28][29]	The Others	[53]

[1] The MPEG-7 Data Set CDs 20, 21, and 22 contain a Spanish TV movie, a Spanish TV sitcom, and a Portuguese TV sitcom, respectively.

2.4.1 Deterministic Approaches

The deterministic approaches to visual-only or audio-only dialogue and action scene detection are based on the extraction of low-level features such as color, motion, texture, silence ratio, and audio energy. Shots which exhibit similar attributes and are temporally close to one another are clustered together. The presence of a dialogue scene is revealed by a repetitious structure of similar shots or a repetitive change of speakers. However, errors emerge in methods where only low-level information is used. A scene simply exhibiting a repetitive shot structure could be classified as a dialogue scene. Furthermore, errors might appear when a speaker dominates the dialogue and the other participants are less frequently shown. Hence, most recent methods include post-processing steps in order to eliminate the errors and improve their performance. For action scene detection, dialogue detection is extended by employing the average shot length and measuring motion activity.

In [46], dialogues are detected by exploiting the local topology of an image sequence and employing statistical tests. A topological framework examining the local metric relationships between images is introduced. The analysis assumes that each shot in the video is represented by a single keyframe. The topological graph $T_G = \{V, E\}$ of a sequence of K images is a fully connected graph with vertices being the video sequence images and edges specifying the metric relationship between

the images. Let T_{MAT} be the $K \times K$ adjacency matrix of T_G. An ideal dialogue is a structure, where every 2nd keyframe is alike, while adjacent keyframes differ. In such a case, T_{MAT} contains ones in the 1st off-diagonal elements, zeros in the 2nd off-diagonal elements, ones in the 3rd off-diagonal elements, and so forth. The following periodic analysis transform $\Delta(n)$ is proposed to identify the aforementioned structure in a sequence of N shot keyframes. If o_i, $i \in \{0, N-1\}$, is a time-ordered sequence of keyframes, then

$$\Delta(n) = 1 - \frac{1}{N} \sum_{i=0}^{N-1} d(o_i, o_{\mathrm{mod}(i+n,N)}), \qquad (2.2)$$

where $d()$ is a color histogram-based distance function. The system detects dialogues by determining whether $\Delta(2) > \Delta(1)$ and $\Delta(2) > \Delta(3)$ are statistically significant decisions. The dialogue detection algorithm is applied using a sliding window in the entire video sequence. Experiments performed in three movies (cf. Table 2.1) have produced a recall rate between 80% and 91% at a precision rate fluctuating between 84% and 100%. However, the system under discussion is operating at its full potential only when the dialogue exhibits a periodic structure.

In [5], *shot interactivity* is introduced, expressing how actively shots in a particular time segment relate to one another. The algorithm is based on the observation of the repetitive appearances of similar shots. Similar shots are determined with respect to the characteristics of the included frames, such as the color histogram and the luminance layout of mosaic picture [6]. Dialogue scenes are identified by clustering groups of neighboring shots whose shot interactivity exceeds a threshold. Two parameters, *dialogue density* δ, which expresses the sum of shot durations, and *dialogue velocity* v, which expresses how frequently the speakers change, are defined:

$$\delta_{\alpha b} = \frac{\sum_{i=\alpha}^{b} \rho_{\alpha b,i} \lambda_i}{\sum_{i=\alpha}^{b} \lambda_i} \qquad v_{\alpha b} = \frac{\sum_{i=\alpha}^{b} \rho_{\alpha b,i}}{\sum_{i=\alpha}^{b} \lambda_i} \qquad (2.3)$$

where λ_i is the duration of shot i, and $\rho_{\alpha b,i}$ is a binary variable which admits the value 1, when shot i contains a dialogue in the shot range $[\alpha, b]$. The shot interactivity from shot α to shot b, is the product of $\delta_{\alpha b}$ and $v_{\alpha b}$, which increases either with the increase of the length of shots which include a dialogue or when frequent transitions between the speakers occur. Experiments were conducted in 4 news shows and 3 variety shows. On average, the recall rate for news programs was 86% and the corresponding precision was 94%. For variety shows, both rates were found to be 100%.

In [25], a dialogue detection system is described, that employs low and mid-level visual features. The system is depicted in Figure 2.4. The first level of the system involves the processing of low-level visual data, determining the shot boundaries and the motion present within each shot of a video sequence. Histogram-based shot boundary detection is applied in order to extract keyframes, whereas the motion extraction block employs the motion vectors exported from the MPEG-1 bitstream. In

the second level of the system, visually similar shots that are temporally close are clustered together. The clustering method is based on the difference of the average color histogram between the shot keyframes [51]. At the same level, camera motion analysis is performed determining if significant motion is present in a shot. In the third level of the system, dialogue detection is performed. First, potential dialogue sequences (PDS) are identified solely from the camera motion analysis output. Hence, when a number of consecutive static shots is encountered, a PDS is declared. When non-static shots begin to dominate over the static ones, the PDS ends. After having identified all PDS, a further processing step is applied in order to verify whether these scenes are indeed dialogue scenes or not. This process involves the calculation of the so-called *cluster to shot ratio* ($C : S$) in the PDS, which determines the percentage of visually unrelated shots in the PDS. $C : S$ ratio is simply the number of clusters that have shots within the PDS to the total number of shots in the PDS. The authors argue that a low $C : S$ ratio is consistent with a dialogue scene, since it reveals a repetitive structure of similar shots. Five movies with a total of 171 manually marked-up dialogues were used to evaluate system performance (cf. Table 2.1). Scenes marked as a dialogue by the authors were sequences of five or more shots containing at least two people conversing, where the main focus of the sequence is conversation. For instance, two people conversing in the middle of a car chase would not apply to that rule, as the main focus is considered to be in the chase. The average recall and precision rates were 86% and 77.8%, respectively. However, as the authors state, an improvement in the exact start and end points of the dialogues is necessary.

The same authors have extended the work in [25] by proposing a similar configuration for detecting action sequences in movies, where the final level of the system differs [24]. The detection of action sequences is performed by using a state machine that was created to search for sequences that match the structure of action scenes. In particular, the state machine looks for sequences in which temporally short shots with high motion activity are dominant. These potential action sequences (PAS) are either accepted or rejected as being true action sequences based on the clustering input. The authors consider that an action scene should lead to a quite high $C : S$ ratio. For this reason, they apply to the $C : S$ ratio an empirically chosen threshold. Experiments were performed on 4 movies (cf. Table 2.1), and the reported recall and precision rates exceeded 80% and 40%, respectively.

Chen and Özsu proposed a rule based model to extract simple dialogue and action scenes instead of clustering shots into scenes using image features [9]. The rules utilized the four types of shots defined in Section 2.2, which determine whether participants' faces in a 2-person dialogue scene are visible in the shot or not and define what type of shot may follow an A, B, or C-type shot. Based on these rules, a finite state machine (FSM) was developed, being able to extract simple (2-person) dialogue or one-on-one fighting scenes. More specifically, a small number of consecutive shots, used to establish a dialogue scene, was characterized as *elementary dialogue scene*. The authors empirically identified 18 different types of elementary dialogue scenes.
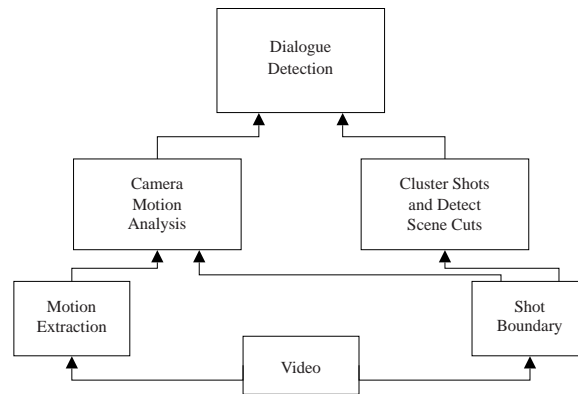
Fig. 2.4. Dialogue detection system proposed by Lehane [25].

The concept of a *video shot string* (VSS) is introduced, in order to represent the temporal occurrence of the different shot types in a video sequence. A VSS is a set of video shots whose types belong to one of the four video shot types defined in Section 2.2. A VSS of a dialogue scene (VSSDS) is defined as a VSS whose prefix is the one for the elementary dialogue scenes expanded by appending some of the three types of shots that include the faces of the dialogue participants. An elementary dialogue scene ending with a shot A can be expanded by appending either shot B or C. An elementary dialogue scene with no additional shots appended to it is classified as a VSSDS as well. In order to extract VSSDS, the VSS is input to a deterministic FSM. A dialogue scene is extracted when a path corresponding to a VSSDS is encountered. The differentiation between dialogue and action scenes was based on the average shot length in a scene, considering that the average shot length in action scenes is smaller than that in dialogue scenes. Experiments were conducted in 3 movies for the dialogue detection system and 2 movies for the action detection (cf. Table 2.1). The movies were first segmented into shots and the actor appearances were manually marked and used as input to the FSM. For the three movies, the dialogue scene detection algorithm exhibited a recall rate equal to 96.6%, 90.51%, and 97.28% at precision rate of 89.47%, 80.52%, and 91.79%, respectively. Correspondingly, the action scene detection algorithm had a recall rate equal to 84% and 81.6%, at a precision rate of 84%, 76.56%, respectively.

In [36], a technique for clustering shots into settings or dialogues is described. The dialogue scenes are considered to have alternating shots of the participants, with only one character displayed at any given time, in frontal view. A face detector [44] and a face classification method are also employed. Faces in neighboring frames which exhibit similarity in position and size are assigned to groups called *face-based classes*. In a second step, face-based classes with similar faces within the same shot are merged by the eigenfaces [35], in order to obtain the largest possible face-based classes. A sequence of at least three consecutive shots is identified as a dialogue when the following conditions apply. At least one face-based class should be present in

each shot, being no more than 1 s apart from its neighbor. Additionally, the eigenface merged face-based classes should alternate within the shot sequence. Experiments performed in two movies for the determination of dialogue scene boundaries yielded hit rates equal to 80% and 86%, miss rates equal to 7% and 4%, and false hit rates 13% and 10%, respectively.

In [22], dialogue detection using audio-only information is presented. Each speaker is characterized by an indicator function. It is demonstrated that a dialogue scene should have a high correlation between pairs of indicator functions. The features utilized are the cross-correlation of indicator functions, and their respective cross-power spectra. Experiments were performed in 6 movies, exhibiting a precision rate of 100% at a recall rate of 85.7%, yielding an F_1 measure of 0.922.

2.4.2 Probabilistic Approaches

In addition to the deterministic approaches, probabilistic ones using HMMs have been proposed and implemented for the efficient characterization of dialogue scenes [40, 41]. The design of an HMM consists in defining its states, specifying its topology, and determining the parameters at each state. Then, the HMM parameters are computed using the Baum-Welch algorithm and the best state sequence for a given input is determined using the Viterbi algorithm.

HMMs were used by Ferman and Tekalp for extracting the semantic content of a video sequence [19]. The HMM models the time-varying structure of a video sequence. It is characterized in terms of its component shots, as depicted in Figure 5(a) and is used to classify each shot of the sequence into one among three categories represented by HMM states. The *Dialogue* state represents self-repetitive shots that reoccur over a temporal window, while the *Progression* state encompasses the shots introducing new camera setups. The *Misc* state accounts for miscellaneous entries, not included in the two other states. The HMM used to model the dialogue state is illustrated in Figure 5(b). The *Est* state represents an establishing shot, used to determine the location of the action, whereas the *Master* state refers to master shots which provide a view of all characters in the scene. The states *1-Shot* and *2-Shot* correspond to shots including the respective number of people.
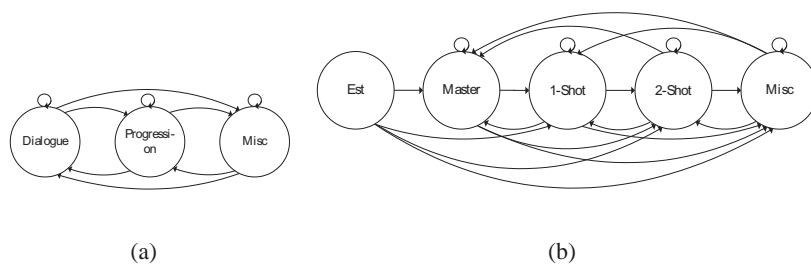

(a)   (b)

Fig. 2.5. HMMs proposed in [19]: (a) HMM for characterizing video sequences (b) HMM for dialogue sequences.

Each shot of the video sequence is characterized by a single feature vector given as input to the HMMs. The necessary features include the normalized distance of the median histograms of two successive shots, the normalized pixel differences between the last frame of a shot and the first frame of its immediate successor, and the normalized distance between the direction histograms of the last few frames of a shot and the first few frames of its neighbor. The direction histogram is comprised from the orientations of the individual motion vectors. Furthermore, shot duration, shot activity, as well as shot transition type (cut, fade or dissolve) are incorporated in the feature vector. After the feature vectors are computed for each shot, the Baum-Welch algorithm is employed in order to train the HMMs, and shot labeling is performed using the Viterbi algorithm.

Table 2.2. Results for visual-only and audio-only dialogue detection experiments.

Reference	Recall	Precision	F_1
Aoki (dialogue detection - news) [5]	86.0%	94.0%	0.898
Aoki (dialogue detection - variety) [5]	100.0%	100.0%	1.000
Chen et al. (dialogue detection) [9]	94.8%	87.4%	0.909
Chen et al. (action scene detection) [9]	82.3%	78.6%	0.804
Kotti et al. (dialogue detection) [22]	85.7%	100.0%	0.922
Lehane et al. (action sequence detection) [24]	92.6%	59.4%	0.533
Lehane et al. (dialogue detection) [25]	86.0%	77.8%	0.816
Sundaram et al. (dialogue detection) [46]	86.0%	95.0%	0.903

Reference	Hit Rate	Miss Rate	False Hit Rate
Pfeiffer et al. (dialogue detection) [36]	84%	12%	3.9%

The results achieved for visual dialogue detection techniques we have reviewed, are summarized in Table 2.2. When the authors provide results for each movie or TV program separately, the average results, measured over the total number of dialogue and action scenes in all movies, have been included in Table 2.2. In addition, we have computed the F_1 metric for all the methods described.

2.5 Audiovisual Dialogue and Action Scene Detection

In this section, methods are discussed, which exploit both the video and audio information, for efficient detection of dialogue and action scenes. Some methods are extensions of those described in Section 2.4, incorporating the information contained in both the video and the audio channels. The techniques for audiovisual dialogue and action scene detection are classified as deterministic [10, 26, 29, 53] and probabilistic [4, 2, 3, 28, 50], like in Section 2.4. While the deterministic methods usually cluster consecutive shots by utilizing appropriate measures, most probabilistic approaches use HMMs representing the semantic events in their states. The deterministic meth-

ods are presented in Section 2.5.1, whereas the probabilistic methods are described in Section 2.5.2.

2.5.1 Deterministic Approaches

Dialogue detection using audiovisual cues is performed in [29], where three types of events are identified: *2-speaker dialogues*, *multiple-speaker dialogues*, and *hybrid events*, which are defined as events containing less speech and more visual action. The framework proposed by Li is depicted in Figure 2.6. At first, shot detection is employed using a color histogram-based method [27]. Visually related shots, that are close to each other, are grouped into *shot sinks*. The similarity between two shots is determined by the Euclidean distance or the histogram intersection between the color histograms of the two shot keyframes.
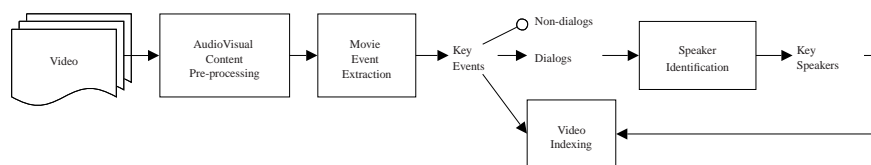

Fig. 2.6. Movie analysis framework proposed by Li (adapted from [29]).

In the next stage, each sink is assigned into one of three predefined classes: *periodic*, *partly-periodic*, and *nonperiodic*. The categorization of each sink is based on the so-called *shot repetition degree*, which is determined by the distance between each pair of neighboring shots. Therefore, a *distance sequence* is determined for each sink. Intuitively, a distance sequence corresponding to a periodic class would exhibit a smaller standard deviation than the one belonging to a nonperiodic class. The k-means algorithm is employed to group all sinks into the 3 classes based on the distance sequences characteristics.

All the temporally overlapping sinks are grouped into one event. During the event grouping procedure, a boundary between two events is declared, when a *progressive scene* appears that consists of some sequential, nonrepetitive shots. The events extracted are organized into 2-speaker dialogues, multiple-speaker dialogues, and hybrid events based on the number of periodic, partly-periodic, and nonperiodic shot sinks included in the event. In addition, two more features are computed for each event in order to validate the aforementioned classification: the event length, which should exceed a certain threshold, and the temporal variance, which is defined as the average variance of the color histogram of all shots within the event. The temporal variance indicates the amount of motion included in an event.

In order to reduce the errors inherent in the deterministic approaches, a post-processing step is included, where audio and face characteristics are incorporated. 5 audio features, namely the short-time energy, short-time average zero-crossing rate,

fundamental frequency, energy band ratio, and silence ratio are extracted. A rule-based heuristic procedure incorporating these audio features is performed, aiming at classifying the shots into one of the following classes: silence, speech, music, and environmental sounds. An event is confirmed as a dialogue, if at least 40% of its shots contain speech. The facial analysis includes the detection of frontal faces. A simple face tracking system is employed, that retains only the faces appearing in several consecutive frames. A 2-speaker dialogue is considered as not having more than one face in most of its component shots. Hence, when more faces are detected it is relabeled as multiple-speaker dialogue. The system was evaluated with encouraging results in three movies, containing 80 events in total. When audio and facial cues were integrated, the false alarms were eliminated, yielding a precision rate of 100%, and a recall rate higher than 83% in all movies. However, the amount of heuristic rules and employed thresholds requires a large validation set in addition to the test set in order to experimentally verify the rules and the corresponding thresholds associated to the rules.

A deterministic FSM for classifying video scenes is employed in [53]. 3 different categories of scenes are identified: conversation, suspense, and action. The proposed method exploits the structural information of the scenes based on shot motion and audio energy as well as mid-level features, i.e., person identity based on face detection [48]. The weighted sum of the extracted low-level features constitutes the *activity intensity* parameter, which is considered to admit low values in conversation scenes. The activity intensity parameter is used as an input to the FSM. The other input, *person identity*, stems from the face detection process. The middle frame of each shot is selected as its keyframe and the face detector is applied, which is expanded in order to include the torso of the detected person. The similarity between two shots is measured by the color histogram intersection between the detected bodies. The shots are then clustered based on the body similarity using the k-means algorithm.
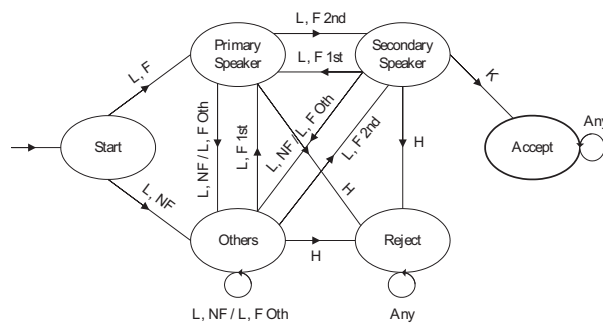

Fig. 2.7. FSM for conversational scene. L – low activity intensity; H – high activity intensity; F – facial shot; NF – non-facial shot; 1st, 2nd, Oth – speaker clusters; K – acceptance condition satisfied; Any – any shot (adapted from [53]).
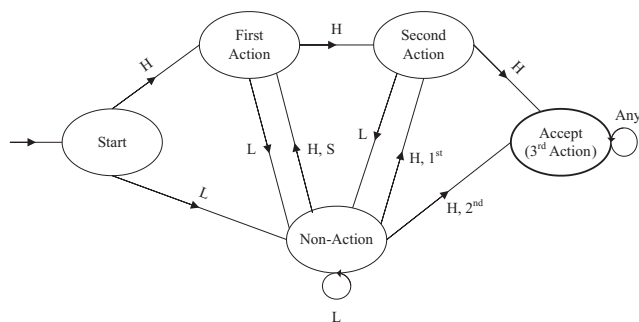
Fig. 2.8. FSM for action scene. L – low activity intensity; H – high activity intensity; S, 1st, 2nd - pre-state values to determine which transition should be taken from start 'Non-action'; Any – any shot (adapted from [53]).

The FSM for classifying conversational scenes is shown in Figure 2.7. The character having the largest cluster is denoted as *Primary Speaker* and the character with the second largest cluster is the *Secondary Speaker*. The transitions of the FSM are determined from the feature values of the shots in the scene. The state *Accept* of the FSM is reached and a *Conversation scene* is declared, when there are at least two main speakers with more than three appearances in the scene. Similar structures are proposed for the FSMs defining the other types of scenes. The FSM for classifying action scenes is depicted in Figure 2.8. To classify a scene as an action scene, the scene must contain a certain number of shots with action intensity greater than a defined threshold level. The FSMs for conversational, suspense, and action scene detection have been tested in a number of movies and TV shows, where a total of 35 conversational, 16 suspense, and 33 action scenes were included. The dialogue scene detection method yielded a recall rate of 94.3% and a precision rate of 97.1%. The precision and recall rates for the suspense scenes were 100% and 93.7%, respectively, whereas the action scenes exhibited precision and recall rates equal to 91.4% and 97%, respectively.

Lehane et al. extended their work [25] described in Section 2.4, by incorporating audio analysis [26]. Low-level audio features are extracted: high zero-crossing rate, silence ratio and short-time energy. A filter determines if an audio clip contains only silence by using the silence ratio and the short-time energy. Afterwards, in order to detect the presence of speech or music, a Support Vector Machine (SVM) that uses the zero crossing rate and the silence ratio is employed. Audio information is fed to an audio-only FSM and color and motion information is input to a video-only FSM. The output of the two FSMs is combined in order to classify the scenes. The combined system delivered a recall rate of 96.5% and a precision rate of 81.33%. The average precision using the combined audio and visual system is 3% lower than the average precision of the visual system, but there is a 12.5% improvement in recall. However, the performance evaluation assumed that a correct decision was taken when either only a part of the dialogue sequence was identified or a manually marked dialogue scene was split into two separate conversations.

Chen et al. have also extended their work [9] to dialogue and action scene extraction by incorporating audio cues in their system presented in Section 2.4 in order to improve accuracy [10]. The underlying model is an FSM coupled with audio features that are determined using an audio classifier. The audio features employed are the zero-crossing rate variance, the silence ratio, and the harmonic ratio. An SVM is trained to classify the audio channel as either speech with environmental sound or music encountered in dialogue scenes or as environmental sound mixed with music encountered in action scenes. Hence, if the audio channel of a scene has more speech segments than enviromental/music segments, then the corresponding scene will be considered as a dialogue scene. The experiments performed in 2 movies (cf. Table 2.1). The dialogue scenes exhibited a recall rate equal to 96.60% and 90.51% for the 2 movies respectively, whereas the corresponding precision rates were 93.4% and 86.11%. The recall rate for action scenes was 100% in both movies and the precision rates were 86% and 81.08%, respectively.

2.5.2 Probabilistic Approaches

An approach for multi-modal dialogue detection using HMMs has been proposed by Alatan et al. [4, 2, 3]. Each shot is classified into speech, silence, or music based on the audio content and at the same time face occurrences and location changes are detected by analyzing the video content. Face analysis is limited to declaring the existence or not of a face in the shot, whereas the location analysis uses histogram-based methods. Each shot is assigned a token based on the analysis of the audio-visual features, i.e. 'SFC' for *silence*, *face existence* and *location change*. The tokens are used to identify dialogue scenes. More specifically, they are used as input, in order to obtain the state sequence that is most likely to have generated that sequence of tokens. At the output of the HMM, each shot of the input sequence is labeled according to the type of scene that best fits it. The block diagram of the system is depicted in Figure 2.9.
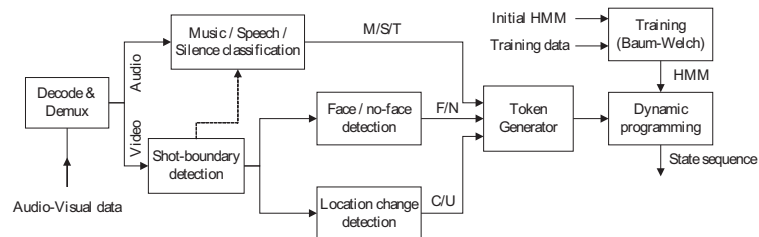

Fig. 2.9. Block diagram of the proposed system in [3]. T – speech; S – silence; M – music, F – face existence; N – no face; C – location change; U – location unchanged

Two different topologies for the HMM are proposed, as shown in Figures 2.10a and 2.10b. The left-to-right topology (Figure 2.10a) includes three state types, called

establishing scene, dialogue scene, and transitional scene. The circular topology (Figure 2.10b) has only two states, the dialogue scene and the non-dialogue scene. The left-to-right topology requires the knowledge of the number of scenes in the content as a prerequisite; hence, its practical use is in doubt, since this information is not usually available a priori. The HMMs are trained by a video data set to determine the state-transition probabilities.
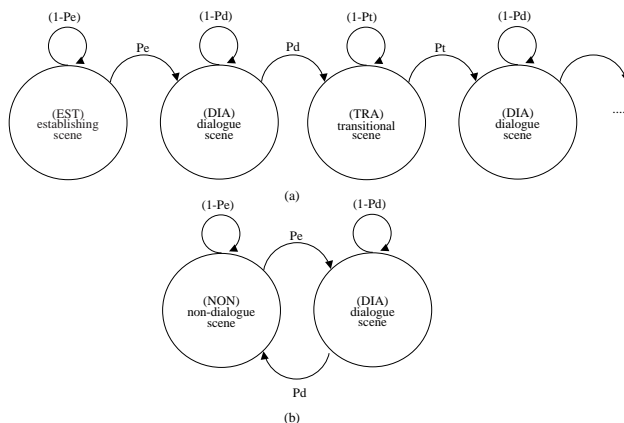

Fig. 2.10. (a) Left-to-right and (b) circular HMM state diagram for dialogue scenes in movies (adapted from [4]).

Two TV sitcoms and one movie were used to compare the two different HMM topologies. The ground truth was obtained by manually assigning every shot to a scene type (establishing, dialogue and transitional, or dialogue non-dialogue), depending on the HMM topology. Furthermore, the audio-visual features, used to produce the tokens, were also manually obtained. The system performance was evaluated using the *shot accuracy* measure. The left-to-right topology performed better compared to its circular counterpart, obtaining a shot accuracy measure for each video sequence equal to 92%, 98%, 99% against 71%, 82%, 94% respectively. It is worth mentioning that the input data in the left-to-right topology had to be manually pre-segmented, so that they contained one establishing scene, one dialogue scene and one transitional scene. Otherwise, it is not possible to use the left-to-right topology. Obviously, this process is not feasible in practice.

As a next step, different observation and training sets are applied to the circular topology, in order to further examine its performance. In addition to the shot accuracy measure, a *scene accuracy* measure was introduced, which was defined as the ratio of correct scene assignments to the total number of scenes being either dialogue or non-dialogue ones. Three different sets of observation symbols were used, audio only, audio and face, and audio, face, and location. These different data sets were also tested for different training data. The best results (scene accuracy around 91%) were obtained when face and audio were the observed features. The location change

detection had no impact or even negative impact to the system. Furthermore, when the training data were not included in the test data, the system performance decreased considerably. Additionally, the system was unable to distinguish between dialogue and monologue scenes, since it does not incorporate any information about the occurrences of the detected face, i.e., if a face has appeared before in the sequence.

Another work on movie scene segmentation was performed by Yaşaroğlu et al. [50]. In particular, an algorithm for automatic multimedia content summarization by segmenting a video into semantic scenes using HMMs was proposed. Two different content types with different properties are defined: dialogue-driven content and action-driven content. Several visual and audio descriptors are extracted, such as face detection descriptors using simple heuristics in the YUV color space and audio features including the zero-crossing rate and the autocorrelation function. In addition, location change analysis is performed using a windowed histogram comparison method. Finally, frame motion vectors are analyzed for detecting motion activity. The variance of magnitudes of these vectors is calculated for each frame and variances are averaged for each shot. The HMM, which has a 2-state topology (the states are labeled as "Dialogue" and "Non-dialogue"), is trained using the Baum-Welch algorithm and the above low-level features as input. Experiments were performed on TV series and family movies yielding recall and precision rates 95% and 80%, respectively.

The results obtained by the reviewed deterministic and probabilistic audiovisual dialogue detection methods are summarized in Table 2.3, along with the performance measure used.

Table 2.3. Results on audiovisual dialogue detection experiments.

Reference	Recall	Precision	F_1
Chen et al. (dialogue detection) [10]	92.9%	88.9%	0.909
Chen et al. (action scene detection) [10]	100%	82.5%	0.904
Lehane et al. (dialogue detection) [26]	96.5%	81.3%	0.882
Li et al. (dialogue detection) [29]	94.2%	100.0%	0.970
Yaşaroğlu et al. (scene segmentation) [50]	95.0%	80.0%	0.868
Zhai et al. (suspense scene detection) [53]	93.7%	100%	0.967
Zhai et al. (action scene detection) [53]	97.0%	91.4%	0.941

Reference	Shot Accuracy R_1
Alatan (dialogue detection - Left-to-Right) [4] [3]	0.96
Alatan (dialogue detection - Circular) [4][3]	0.82

2.6 Conclusions

As the amount of multimedia content available in the web, broadcast data streams or personal collections grows exponentially, multimedia data management becomes

an indispensable tool for efficient and user-friendly browsing and retrieval of such data. Dialogue and action scene detection techniques aim at segmenting a video into semantically meaningful units with respect to this particular semantic concept, i.e., the existence or not of a dialogue or an action scene in movies or TV programs. This process can lead to a more sophisticated navigation, browsing and searching of the video document.

Low and mid-level features, extracted from visual and audio analysis, are exploited. The predominant approach is to classify temporally close shots that demonstrate similar low level features and search for repetitive shot patterns. However, this strategy may cause semantically unrelated shots to be clustered together, based on their "low-level similarity". In addition, visually dissimilar shots that are commonly inserted in semantically coherent scenes, introduce a non-deterministic nature to these scenes. Hence, statistical models, employing HMMs, have also been applied. It has been observed that techniques integrating visual and audio information, using either low or mid-level features, yield more accurate dialogue and action scene detection than classifiers that employ video only or audio only information. In addition, probabilistic techniques exhibit improved performance over deterministic classifiers.

Generally speaking, limited research has been performed in the field of dialogue and action scene detection. In addition, a universal and commonly accepted definition of a "dialogue scene" or an "action scene" does not exist, and most authors introduce their own perspective. Nor does a common, annotated database for the performance evaluation of the proposed methods exists; every method is tested in a different relatively small data set, where the ground truth is subjectively defined. Hence, the comparison of the presented results can not lead to a safe conclusion. In general, dialogue and action scene detection are promising techniques for the segmentation of a video document into semantically meaningful units, but much work remains to be done in order to devise robust and efficient methods.

Thus, the creation of a common annotated database for scene analysis and dialogue detection experiments that would enable comparative evaluation of different methods is necessary. This database could include the movies and TV shows enlisted in Table 2.1. A standardization of the experimental protocols and figures of merit will also help to establish a common ground for method comparison and evaluation.

Acknowledgement

References

1. Mpeg-7 overview (version 9). Technical report, ISO/IEC JTC1/SC29/WG11 N5525, March 2003.

2. A. A. Alatan. Automatic multi-modal dialogue scene indexing. In *Proc. Int'l Conf. Image Processing*, volume 3, pages 374–377, 2001.

3. A. A. Alatan and A. N. Akansu. Multi-modal dialog scene detection using hidden-markov models for content-based multimedia indexing. *J. Multimedia Tools and Applications*, 14:137–151, 2001.

4. A. A. Alatan, A. N. Akansu, and W. Wolf. Comparative analysis of hidden markov models for multi-modal dialogue scene indexing. In *Proc. Int'l Conf. Acoustics, Speech, and Signal Processing*, volume IV, pages 2401–2404, 2000.

5. H. Aoki. High–speed dialog detection for automatic segmentation of recorded tv program. In *Int'l Conf. Image and Video Retrieval*, pages 49–58, 2005.

6. H. Aoki, S. Shimotsuji, and O. Hori. A shot classification method of selecting effective key-frames for video browsing. In *ACM Conf. Multimedia*, pages 1–10, 1996.

7. F. Beaver. *Dictionary of Film Terms*. Twayne Publishing, New York, 1994.

8. D. Bordwell and K. Thompson. *Film Art: An Introduction*. McGraw–Hill, Inc., 4th ed., New York, 1993.

9. L. Chen and M. T. Özsu. Rule-based scene extraction from video. In *Proc. Int'l Conf. Image Processing*, volume 2, pages 737–740, 2002.

10. L. Chen, S. Rizvi, and M. T. Özsu. Incorporating audio cues into dialog and action scene extraction. In *Proc. IS&T/SPIE's 15th Annual Symp. Electronic Imaging - Storage and Retrieval for Media Databases*, pages 252–264, 2003.

11. T.F. Cootes, G.J. Edwards, and Taylor C.J. Active appearance models. *IEEE PAMI*, 23(6):681–685, 2001.

12. C. Cotsaces, N. Nikolaidis, and I. Pitas. Video shot detection and condensed representation: A review. *IEEE Signal Processing Magazine*, 23(2):28–37, March 2006.

13. M. S. Crouse, R. D. Nowak, and R. G. Baraniuk. Wavelet-based statistical signal processing using hidden markov models. 46(4):886–902, 1998.

14. A. P. Dempster, N. M. Laird, and D. B. Rubin. Maximum Likelihood from incomplete data via the EM algorithm. *J. of Royal St. Soc. (B)*, 39(1):1–38, 1977.

15. L. Deng, J. Dropo, and A. Acero. Dynamic compensation of HMM variances using the feature enhancement uncertainty computed from a parametric model of speech distortion. *IEEE TSAP*, 13(3):412–421, 2005.

16. V. Digalakis, J.R. Rohlicek, and M. Ostendorf. ML estimation of a stochastic linear system with the EM algorithm and its application to speech recognition. *IEEE TSAP*, pages 431–442, 1993.

17. S. Dupont and J. Luettin. Audio-visual speech modeling for continuous speech recognition. *IEEE Tr. on Mult.*, 2(3):141–151, 2000.

18. P. Felzenszwalb and D. Huttenlocher. Pictorial structures for object recognition. *Int'l J. of Comp. Vis.*, 61(1):55–79, 2005.

19. M. Ferman and A. M. Tekalp. Probabilistic analysis and extraction of video content. In *Proc. Int'l Conf. Image Processing*, volume 2, pages 91–95, 1999.

20. B.J. Frey, T. Kristjansson, L. Deng, and A. Acero. Learning dynamic noise models from noisy speech for robust speech recognition. In *Proc. NIPS*, volume 8, pages 472–478, 2001.

21. H. Glotin, D. Vergyri, C. Neti, G. Potamianos, and J. Luettin. Weighting schemes for audio-visual fusion in speech recognition. In *Proc. ICASSP*, 2001.

22. M. Kotti, C. Kotropoulos, B. Ziolko, I. Pitas, and V. Moschou. A framework for dialogue detection in movies. *Lecture Notes in Computer Science*, 4105:371–378, 2006.

23. P. Král, C. Cerisara, and J. Klečková. Automatic dialog acts recognition based on sentence structure. In *Proc. Interspeech 2005*, pages 825–828, 2005.

24. B. Lehane, N. O'Connor, and N. Murphy. Action sequence detection in motion pictures. In *Proc. European Workshop Integration of Knowledge, Semantics and Digital Media Technology*, 2004.

25. B. Lehane, N. O'Connor, and N. Murphy. Dialogue scene detection in movies using low and mid-level visual features. In *Proc. Int. Conf. Image and Video Retrieval*, pages 286–296, 2004.

26. B. Lehane, N. O'Connor, and N. Murphy. Dialogue sequence detection in movies. In *Proc. Int. Conf. Image and Video Retrieval*, pages 286–296, 2004.

27. Y. Li and C. C. J. Kuo. Real-time segmentation and annotation of mpeg video based on multimodal content analysis i & ii. Technical report, Univ. Southern California, Los Angeles, Technical Report, 2000.

28. Y. Li, S. Narayanan, and C. C. J. Kuo. Identification of speakers in movie dialogues using audiovisual cues. In *Proc. Int'l Conf. Acoustics, Speech, and Signal Processing*, volume 2, pages 2093–2096, 2002.

29. Y. Li, S. Narayanan, and C. C. J. Kuo. Content-based movie analysis and indexing based on audiovisual cues. *IEEE Trans. Circuits and Systems for Video Technology*, 14(8):1073–1085, 2004.

30. J. Luettin, G. Potamianos, and C. Neti. Asynchronous stream modeling for large vocabulary audio-visual speech recognition. In *Proc. ICASSP*, 2001.

31. I. Matthews, T. F. Cootes, J. A. Bangham, S. Cox, and R. Harvey. Extraction of visual features for lipreading. *IEEE PAMI*, 24(2):198–213, 2002.

32. A. Morris, A. Hagen, H. Glotin, and H. Bourlard. Multi-stream adaptive evidence combination for noise robust ASR. *Speech Communication*, 34:25–40, 2001.

33. A.V. Nefian, L. Liang, X. Pi, X. Liu, and K. Murphy. Dynamic bayesian networks for audio-visual speech recognition. *EURASIP Journal on Applied Signal Processing*, 11:1–15, 2002.

34. E. K. Patterson, S. Gurbuz, Z. Tufekci, and J. N. Gowdy. CUAVE: A new audio-visual database for multimodal human-computer interface research. In *Proc. ICASSP*, 2002.

35. A. Pentland, B. Moghaddam, and T. Starner. View–based and modular eigenspaces for face recognition. In *Proc. IEEE Conf. Computer Vision and Pattern Recognition*, pages 84–91, 1994.

36. S. Pfeiffer, R. Lienhart, and W. Effelsberg. Scene determination based on video and audio features. *Multimedia Tools and Applications, Kluwer Academic Publishers*, 15:59–81, 2001.

37. A. Potamianos, E. Sanchez-Soto, and K. Daoudi. Stream weight computation for multi-stream classifiers. In *Proc. ICASSP*, 2006.

38. G. Potamianos, C. Neti, G. Gravier, and A. Garg. Automatic recognition of audio-visual speech: Recent progress and challenges. *Proc. of the IEEE*, 91(9):1306–1326, 2003.

39. W. Press, S. Teukolsky, W. Vetterling, and B. Flannery. *Numerical Recipes*. Cambridge Univ. Press, 1992.

40. L. Rabiner. A tutorial on hidden markov models and selected applications in speech recognition. *Proc. of the IEEE*, 77(2):257–286, 1989.

41. L. Rabiner and B.-H. Juang. *Fundamentals of Speech Recognition*. Prentice-Hall, NJ, USA, 1993.

42. Z. Rasheed and M. Shah. Scene detection in hollywood movies and tv shows. In *Proc. IEEE Conf. Computer Vision and Pattern Recognition*, volume 2, pages 343–348, 2003.

43. R. C. Rose, E. M. Hofstetter, and D. A. Reynolds. Integrated models of signal and background with application to speaker identification in noise. *IEEE TSAP*, 2(2):245–257, 1994.

44. H. A. Rowley, S. Baluja, and T. Kanade. Human face detection in visual scenes. *Advances in Neural Information Processing Systems, The MIT Press*, 8:875–881, 1996.

45. J. Sietsma and R. Dow. Creating artificial neural networks that generalize. *Neural Networks*, 4:67–79, 1991.

46. H. Sundaram and S. F. Chang. Computable scenes and structures in films. *IEEE Trans. Multimedia*, 4(4):482–491, 2002.

47. A. Vassiliou, A. Salway, and D. Pitt. Formalizing stories: sequences of events and state changes. In *Proc. 2004 IEEE Int. Conf. Multimedia and Expo*, volume 1, pages 587–590, 2004.

48. P. Viola and M. Jones. Robust real-time face detection. *Int. J. Compututer Vision*, 57(2):137–154, 2004.

49. W. Wolf. Hidden markov model parsing of video programs. In *Proc. Int'l Conf. Acoustics, Speech, and Signal Processing*, pages 2609–2611, 1997.

50. Y. Yaşaroğlu and A. A. Alatan. Summarizing video: content, features & hmm topologies. In *Proc. Int. Workshop Very Low Bitrate Video Coding*, pages 101–110, 2003.

51. M. M. Yeung and B.-L. Yeo. Video visualization for compact presentation and fast browsing of picturial content. *IEEE Trans. Circuits Systems for Video Technology*, 7:771–785, 1997.

52. N.B Yoma and M. Villar. Speaker verification in noise using a stochastic version of the weighted viterbi algorithm. *IEEE TSAP*, 10(3):158–166, 2002.

53. Y. Zhai, Z. Rasheed, and M. Shah. Semantic classification of movie scenes using finite state machines. *IEE Proc. - Vision, Image, and Signal Processing*, 152(6):896–901, 2005.