A Multi-Expert System for Movie Segmentation

Colace, F.; De Santo, M.; Molinara, M.; Percannella, G.; and Vento, M.

Abstract— ***In this paper we present a system for movie segmentation based on the automatic detection of dialogue scenes.***

The proposed system processes the video stream directly in the MPEG domain: it starts with the segmentation of the video footage in shots. Then, a characteri-zation of each shot between dialogue and not-dialogue according to a Multi-Expert System (MES) is performed. Finally, the individuated sequences of shots are aggregated in dialogue scenes by means of a suitable algorithm. The MES integrates three experts, which classifies a given shot on the basis of very complementary descriptions; in particular an audio classifier, a face detector and a camera motion estimator have been built up and employed.

The performance of the system have been tested on a huge MPEG movie data-base made up of more than 15000 shots and 200 scenes, giving rise to encouraging results.

Index Terms— ***MPEG, Multi-Expert Systems, multimedia database, video analysis***

1. *Introduction*

MORE and more videos are generated every day, mostly produced and stored in analog form. In spite of this, the trend is toward the total digitization of movies and video products given that the effective use of them is hindered by the difficulty of efficiently classifying and managing video data in the traditional analog format.

In the past few years, several algorithms have been presented in the scientific literature to allow an effective filtering, browsing, searching and retrieval of information in video databases [1]. It is generally accepted that the first step toward an effective organization of the information in video databases consists in the segmentation of the video footage in shots that are defined as the set of frames obtained through a continuous camera recording. Anyway, even if the individuation of shots represents a fundamental step, it is clear that this approach does not allow an effective non linear access to the video information. This is evident from at least two points of view: firstly, humans usually remember different events after they watched a movie, and hence they also think in terms of events during the retrieval process; secondly, a modern movie contains more than 2000 shots on average, which means that an intelligent video analysis program needs to process 2000 frames per movie to give a coherent representation.

Consequently, it is necessary to define units for accessing the video footage obtained by grouping semantically correlated shots. Scene is the term most used in the scientific literature to call this semantic unit. First approaches for detecting scenes (see for example [2, 3]) operate by simply clustering the shots according to the visual content of the most representative frames (also called key-frames). Anyway the quoted techniques do not take into account any model for the scene, so the results are not always semantically coherent. In fact, it is worth noting that the way the shots are grouped in a scene generally depends on the type of scene under analysis as well as on the video genre. The scenes of a TV-news program are different from the scenes of a talk-show, of a documentary, of a movie. Hence, it is important to aggregate shots also considering a model for the scene. Several recent papers try to define models for scene detection, mainly in the field of TV-news, where effective and simple models can be defined. For example, in [4] a method based on a Hidden Markov Model to segment TV-news at various semantic levels it is presented, while in [5], Bertini et al. describe the use of multiple features for content based indexing and retrieval of TV-news.

The same problem is much more complex when the movies domain is faced: there are much more different scene types and for each kind of scene different styles can be adopted depending on the Movie Director. Interestingly enough, although scene analysis can be very useful for several purposes (think for example to video abstraction and automatic classification of the video genre) only few papers have been presented on the problem of detection and characterization of scenes in movies [15][16].

Manuscript received April 11, 2005.

Francesco Colace is with the DIIIE, University of Salerno, Italy (e-mail: fcolace@unisa.it)

Massimo De Santo is with the DIIIE, University of Salerno, Italy (e-mail: desanto@unisa.it)

Mario Molinara is with the DAEIMI, University of Cassino, Italy (e-mail: m.molinara@unicas.it)

Gennaro Percannella is with the DIIIE, University of Salerno, Italy (e-mail: pergen@unisa.it)

Mario Vento is with the DIIIE, University of Salerno, Italy (e-mail: mvento@unisa.it)

Among those few, an interesting one is [6] where Saraceno et al. define some simple rules to group the shots of a movie according to some semantic types.

In this paper we present a system for video segmentation based on the automatic detection of dialogue scenes within movies. The detection of dialogue scenes is a task of particular interest given the special semantic role played by dialogue based scenes in the most part of movies. The proposed system starts with the segmentation of the video footage in shots. Then, it operates a characterization of each shot as dialogue or not-dialogue according to a multi-expert approach, where each decision system (expert, hereinafter) classifies a given shot on the basis of a particular description while employing the most appropriate decision technique. The final result is obtained by combining the single decisions through suitable rules [7]. In this way, if the utilized experts consider different and complementary aspects of the same decision problem, the combination of the single decisions provides a performance that is better than that of any single expert. Finally, the individuated sequences of shots are aggregated in dialogue scenes by means of an appropriate algorithm.

In order to improve the computational efficiency of the whole process, we analyze the video footage directly in the MPEG coded domain.

While the general approach of multiple experts is not new (see for example [8, 9]), its application to this specific problem is interesting and quite novel, and the obtained results on a huge MPEG movie database are encouraging.

2. The Proposed System

As stated in the introduction, the proposed method starts with the segmentation of the video footage in shots. Then, a characterization of each shot between dialogue and not dialogue according to a Multi-Expert System (MES) is performed. Finally, the individuated sequences of shots are aggregated in dialogue scenes by means of a suitable algorithm. This approach can be justified on the basis of the following considerations: i) a scene is a group of semantically correlated shots; ii) almost all the shots belonging to a dialogue scene can be characterized as dialogue shots; and iii) the shots belonging to the same dialogue scene are temporally adjacent.

Therefore, it follows that the proposed system can be structured according to three successive stages, as depicted in Fig.1:

- Stage 1 - shot boundaries detection
- Stage 2 - dialogue / not dialogue shot classification
- Stage 3 - shot grouping

A short description of each of the quoted stages is given in the following.

Shot boundaries detection: the problem of automatic detection of shot boundaries has been widely investigated in recent years; hence, the scientific literature is rich of papers discussing approaches which allow us to reliably segment videos in shots both in the un-compressed and in the MPEG coded domain. For the purposes of this paper, we have implemented the technique described in [10] that is characterized by good performances both in terms of correct detection and of low computational requirements, since it operates directly on the compressed stream.

Dialogue - not dialogue shot characterization: this classification is performed through the use of a multi-expert system. The rationale lies in the assumption that, by suitably combining the results of a set of experts according to a rule (combining rule), the performance obtained can be better than that of any single expert. The successful implementation of a multi-expert system (MES) implies the use of the most complementary experts as possible, and the definition of a combining rule for determining the most likely class a sample should be attributed to, given the class to which it is attributed by each single expert.

Therefore, for the purpose of shot classification as dialogue or not, we introduce the following set of experts:

1. Face detection,
2. Camera motion estimation,
3. Audio classification

which are integrated within the whole system as shown in Fig.1.

Each expert can be viewed as constituted by a sensor and a classifier. Each expert of the system has two inputs: the MPEG video or audio stream and the complete list of the shots boundaries. The latter information is used by the sensor to access and characterize the MPEG data at shot level. The output of the sensor is used by the classifier to perform the dialogue / not dialogue shot classification. In our system we have integrated three experts whose sensors implement the algorithms described in [11] for face detection, in [12] for camera motion estimation and in [13] for audio stream classification, all working directly in the video/audio coded domain. It is worth noting that the output of the first sensor is correlated in a simple way to the output of the corresponding expert; in fact, the presence (absence) of a face implies a dialogue (not dialogue) shot. On the contrary, the sensor for camera motion estimation provides three estimates respectively for the zoom, tilt and pan rate for each P frame. Then, the average and the standard deviation of the zoom, tilt and pan rate over each shot constitute the features vector used by a neural network to perform the shot classification. Finally, the sensor for audio classification uses the same feature vector defined in [13], but in our case the classification is realized through a neural network trained to recognize only the dialogue and not dialogue shot classes.

Then, the outputs of the single experts are combined according to a suitable combining rule (for a review of the most common used rules see [7]).

Shot grouping: the final stage of our approach provides to group in dialogue scenes the shots classified in the previous stage. The rationale of the algorithm for shot grouping derives from the consideration that the shots belonging to a dialogue scene are temporally adjacent. However, the shot grouping algorithm has to properly handle also the possible classification errors generated at stage 2. In fact:

- **a false alarm** (i.e. a not dialogue shot classified as dialogue) might cause the declaration of an inexistent short dialogue scene, and
- **a missed detection** (i.e. a dialogue shot classified as not dialogue) might cause the partitioning of a dialogue scenes in two scenes.

Thus the shot grouping algorithm implements the following rule: a transition from a dialogue scene to a not dialogue scene (and vice versa) is declared when a sequence of at least N not dialogue (dialogue) shots occurs. In Fig. 2, there are depicted examples of scene transitions in case of N = 3.

3. EXPERIMENTAL RESULTS

In order to assess the performance of the proposed system we used a large and significant database of video footages obtained from 10 movies. It results in about 20 hours, corresponding to more than 15000 shots and 228 dialogue scenes. In the construction of this movie database particular care was taken to include a representative of the major movie genres (action, comedy, drama, science fiction, fantasy, etc) so that to reproduce the high variability of the video characteristics over the different genres. More details on the chosen movies are given in [14].

In order to setup the proposed system and to assess its performance, we extracted two disjoint sets of samples from the database: a training set (henceforth **TRS**) and a test set (**TS**). The **TS** has been built by choosing continuous sequences of *L* shots from each movie, where *L* was obtained as approximately 30% of the total number of shots in that movie. The choice of using temporally adjacent shots is motivated by the fact that such sequences have to be used to test the Stage 3 of the system in the detection of dialogue scenes. The **TRS** were built by randomly choosing among the remaining part of the database a number of samples corresponding to 50% of the whole dataset. Note that the remaining 20% of samples of the database were used for building the validation set (**VS**); this set was required for training the neural classifiers of the 2nd stage, as it will be clarified in the next subsections.

3.1 PERFORMANCE EVALUATION OF THE STAGE 1

This stage provides the segmentation of the video stream in shots, by mean of the technique described in [10]. In order to assess the performance of this algorithm we carried out a comparison between the algorithm output and the ground truth. Such a comparison consists in the evaluation of the numbers of missed detections (MD, i.e. cut frames which were not detected by the algorithm) and false alarms (FA, i.e. non-cut frames which were declared as cuts from the algorithm). Then the overall performance is usually expressed in terms of *Recall* and *Precision*, which represent the fraction of correctly detected cuts with respect to the true cuts and the total number of detected cuts, respectively. They are defined as it follows:

$$Recall = \frac{CD}{CD+MD} \quad \text{and} \quad Precision = \frac{CD}{CD+FA} \tag{1}$$

where *CD* is the number of correctly detected cuts.

The algorithm implemented in this stage of our system required a tuning phase in order to select a suitable threshold that maximized the performance. In order to take into account both *Precision* and the *Recall*, we used a unique performance index defined as the sum of the preceding indexes; in this way we are able to weigh equally both indexes. The tuning phase required us to select the value of the threshold that maximized the performance on the **TRS**. Once completed the tuning phase, we tested the algorithm on the **TS**, obtaining the following performance: *Recall* = 0.96 and *Precision* = 0.94. These results confirm how the selected algorithm is able to perform very accurately, even if it is interesting noting that the performance of the implemented cut detection algorithm is lower with respect to what the authors declare in [10]. In fact, they report no missed detections and only one false alarm on their test set composed of only 27000 frames and 269 cuts.

3.2 PERFORMANCE EVALUATION OF STAGE 2

The camera motion expert and the audio expert are built by using a neural network - namely, a Multi-Layer Perceptron (MLP) - for their classification modules.

The architecture of the neural classifier of the audio and the camera motion experts has been chosen after a preliminary optimization phase on the **TRS**. In particular, the MLP net adopted for the audio expert is made up of 35 hidden neurons, while the net for the camera motion expert has 25 neurons in the hidden layer.

In Table 1 there are the *confusion matrices* obtained on the **TS** by the best audio, camera motion and face experts.

The face expert required a different experimentation since it employs a naive classifier. It simply associates the presence/absence of a face in the central I-frame of the shot to a dialogue/not-dialogue shot. Anyway, this expert also required a training phase in order to setup some parameters of the face detection algorithm [11], with particular reference to the skin color module. In this case we used the same training, validation and test set defined for the other two experts.

After having assessed the performance of each expert, their results have to be fused together in the combiner. In Table 2, it is represented the *Coverage Table* evaluated on the **TS**, which reports the joint behavior of the three experts with respect to the shot classification task. In particular, the rows of this table represent the percentage of samples in the test set for which respectively three, two, one or zero experts performed the correct classification.

From Table 2 it is readily available the recognition rate achievable by employing a majority voting rule: it is given by the sum of the recognition rates of the first two rows of the quoted table. Hence, by using this simple rule it is possible to achieve a recognition rate of 83.97% (not dialogue shots) and 86.2% (dialogue shots) for the 2nd stage of the system.

It is worth noting that the multi-expert approach allows to obtain a relative overall improvement of about 8% with respect to the best single expert (the Audio one – about 79% correct classification). In Table 3, we have reported the relative improvements obtained by using the MES with respect to each single expert.

3.3 Performance evaluation of Stage 3

In the 3rd stage of the system the shots, classified in the 2nd stage, are aggregated in dialogue and not dialogue scenes. This is realized by the simple shot grouping algorithm described in Section 2.

Before going into the details of the tests that we carried out in order to assess the performances of this stage, it is worthwhile to dwell upon the set of indexes which we are going to estimate. It is important that such set is able to give a correct representation of the actual performances of the system.

We decided to provide a description of the overall performance of our technique in terms of **Correct Detection** (***CD***) and **False Alarms** (***FA***), which respectively account for the actual dialogue scenes which were detected and the dialogue scenes which were detected without being actually present in the movie.

These two parameters are defined as follows:

$$CD = \frac{CDS}{DS}\% \quad FA = \frac{FDS}{NDS}\% \qquad (2)$$

where:

- *CDS* is the number of actual dialogue scenes, which were detected;
- *FDS* is the number of dialogue scenes which were detected, but not actually present in the movie;
- *DS* is the number of actual dialogue scenes;
- *NDS* is the number of actual not dialogue scenes.

To this aim, we declare that an actual dialogue scene has been correctly detected if at least one of its shots is present in a detected dialogue scene. Anyway, it can be simply devised how the indexes introduced before provide only a rough description of the real performances of the system: no information about the “quality” of the detection is given. In fact, such indexes do not account for scenes which are only partially detected and/or split and/or merged. In order to cope with such a problem we introduce two other sets of indexes: *overlap percentages* and *split/merged scenes percentages*.

The first set of indexes has been introduced in order to give a condensed view of how much the detected dialogue scenes coincide with the actual dialogue scenes. Hence, we define the **percentage of correct overlap** (***CO***) and the **percentage of false overlap** (***FO***), given by:

$$CO = \frac{DSF}{ADSF}\% \qquad FO = \frac{NDSF}{ADSF}\% \qquad (3)$$

where:

- *DSF* is the number of frames of the detected dialogue scenes which overlap to the real dialogue scenes;
- *NDSF* is the number of frames of the detected dialogue scenes which do not overlap to the real dialogue scenes;
- *ADSF* is the number of frames of the actual dialogue scenes which have been detected by the system. The rationale inspiring the choice of excluding the undetected actual dialogue scenes relies on the fact that with such set of parameters we want to give a measure only of the quality of the detected scenes.

The set of indexes about split/merged scenes has been introduced in order to take into account the errors occurring when an actual dialogue scene is split into two or more dialogue scenes or vice versa when two or more dialogue scenes are merged together. To this aim we define the **percentage of merged dialogue scenes** (***MS***) and the **percentage of the split dialogue scenes** (***SS***) as it follows:

$$MS = \frac{AS}{DDS}\% \quad SS = \frac{DiS}{DDS}\% \qquad (4)$$

where:

- *AS* is the number of the detected dialogue scenes, which were merged into a single scene;
- *DiS* is the number of the detected dialogue scenes, which were divided into two or more scenes;
- *DDS* is the number of the detected dialogue scenes.

Note that according to the previous definitions it might occur also the situation of a detected dialogue scene that is both merged and divided. In such case this scene is considered for the computation of both *MS* and *SS*.

After the definition of these indexes, we can evaluate the results of the experimental campaign carried out on the video sequences of the **TS**. It is worth recalling that the **TS** has been built by considering a continuous sequence of *L* shots from each movie, where *L* was obtained as approximately the 30% of the total number of shots in that movie.

The experimentation of the 3rd stage of the system required to set only the parameter *N* that was defined in Section 2, representing the minimum number of adjacent shots that allows switching among the two different types of scenes. We tested the system for different values of the parameter *N*. In Table 4, there are reported the experimental results obtained by setting $N = 3$ and 4.

The first conclusion which can be drawn is that the dialogue scene segmentation is significant only in the cases of $N = 3$ and 4. In fact, a higher value gives rise to under-segmentation: many scenes are merged together; conversely with $N = 2$ over-segmentation occurs.

The experimental results are very appealing as in both cases we obtained about 90% in the detection of the dialogue scenes. Furthermore also the results about overlap are quite good with about 80% of correct overlap and only 10% of false overlap. The results about scene overlap are important since they represent an index of the quality of the detection. Low values of *CO* accompanied by high values of *FO* would be misleading, in the sense that they could not allow a user to perceive the true semantic content of the scene.

In order to evaluate the improvement in the overall performance introduced by the use of sensor fusion approach with respect to the case of a single expert, we have tested our system using in the 2nd stage only the best expert (audio). We have reported in Table 5 the experimental results obtained with this expert in case $N = 3$ together with the relative improvement introduced by the use of the MES.

From the experimental results reported in Table 5 it is evident the improvement yielded by the employment of a MES. The advantages are considerable not only in the percentage of correct detection, but also for the other indexes. The use of information about face presence and camera motion allows improving the overall quality of the segmentation.

4. Conclusion

In this work we have analyzed the problem of movie segmentation. The proposed approach is based on the detection of dialogue scene by means of a Multi-Expert System (MES). The MES is constituted by three different experts which analyze the video and audio tracks of the movie directly in the MPEG coded domain. Although each expert is not characterized by optimal performances in the classification of the shots (this is due both to the errors of the sensor and of the classifier which constitute each expert), their combined use gives good performances even when a very simple combining rule is used. This confirms our initial hypothesis that the utilized experts consider different and complementary aspects of the same decision problem. Current research is focused on improving the overall performance of the system by implementing the experts as classifiers able to yield also some information about the reliability of the classification, and by using more sophisticated combining rules. Actually, we are also exploring the possibility to extend the proposed approach to detect action scenes within movies.

References

[1] M.R. Naphade, T.S. Huang, "A Probabilistic Framework for Semantic Video Indexing, Filtering and Retrieval", IEEE Transactions on Multimedia, Vol. 3, No. 1, pp 141-151, 2001

[2] M. M. Yeung, B. Liu, "Efficient matching and clustering of video shots", in Proc. IEEE ICIP'95, vol II, pp. 260-263.

[3] A. Hanjalic, R. Lagendijk, J. Biemond, "Automated high-level movie segmentation for advanced video-retrieval systems", in IEEE Trans. on Circuits and Systems for Video Technology, vol. 9, No. 4, June 1999, pp. 580-588.

[4] S. Boykin, A. Merlino, "Machine learning of event segmentation for news on demand", in Communications of the ACM, Feb. 2000, vol. 43, No. 2, pp. 35-41.

[5] M. Bertini, A. Del Bimbo, P. Pala, "Content-based Indexing and Retrieval of TV-news", in Pattern Recognition Letters, 22, 2001, 503-516.

[6] C. Saraceno, R. Leopardi, "Identification of Story Units in Audio-Visual Sequences by Joint Audio and Video Processing", in Proc. ICIP'98, pp. 363-367, 1998.

[7] L. P. Cordella, P. Foggia, C. Sansone, F. Tortorella and M. Vento, Reliability Parameters to Improve Combination Strategies in Multi-Expert Systems, Pattern Analysis & Applications, Springer-Verlag, vol. 2, pp. 205–214, 1999.

[8] T.K. Ho, J.J. Hull, S.N. Srihari, "Decision Combination in Multiple Classifier Systems", IEEE Transactions on Pattern Analysis and Machine Intelligence 1994; 16(1): 66-75.

[9] J. Kittler , J. Hatef, R.P.W. Duin, J. Matas, "On Combining Classifiers", IEEE Trans. on PAMI, vol 20 n.3 March 1998.

[10] S.C. Pei, Y.Z. Chou, "Efficient MPEG compressed video analysis using macroblock type information", in IEEE Trans. on Multimedia, pp. 321 – 333, Dec. 1999, Vol. 1, Issue: 4.

[11] H. Wang, S.F. Chang, "A Highly Efficient System for Automatic Face Region Detection in MPEG Video", IEEE

Trans. on Circuits and Systems for Video Technology, vol. 7, no. 4, August 1997, pp. 615-628.

[12] Y.P. Tan, D.D. Saur, S.R. Kulkarni, P.J. Ramadge, "Rapid Estimation of Camera Motion from Compressed Video with Application to Video Annotation", IEEE Trans. on Circuits and Systems for Video Technology, vol. 10, no. 1, February 2000, pp. 133-146.

[13] M. De Santo, G. Percannella, C. Sansone, M. Vento, "Classifying Audio of Movies by a Multi-Expert System", Proc. of the 11th ICIAP, pp. 386-391, 2001.

[14] M. De Santo, G. Percannella, C. Sansone, M. Vento, "Dialogue Scenes Detection in Mpeg Movies: a Multi-Expert Approach", LNCS, vol. 2184, pp. 192-201, Sept. 2001.

[15] Liu, L.; Fan, G.; "Combined Key-Frame Extraction and Object-Based Video Segmentation", Circuits and Systems for Video Technology, IEEE Transactions on Volume 15, Issue 7, , July 2005, Page(s):869 – 884

[16] Xingquan Zhu; Xindong Wu; Elmagarmid, A.K.; Zhe Feng; Lide Wu; Video data mining: semantic indexing and event detection from the association perspective, Knowledge and Data Engineering, IEEE Transactions on Volume 17, Issue 5, May 2005 Page(s):665 - 677

TABLES

	Audio		Camera Motion		Face	
	ND	D	ND	D	ND	D
ND	**77.57%**	22.43%	**58.43%**	41.57%	**76.43%**	23.57%
D	20.01%	**79.99%**	23.59%	**76.41%**	30.45%	**69.55%**

Table 1. The confusion matrix obtained on the **TS** by the best audio, camera motion and face expert, where **ND** and **D** stand for *Not-dialogue* and *Dialogue shot*, respectively.

# Correct classification	Not-Dialogue	Dialogue
3	**32.18%**	**45.17%**
2	**51.79%**	**41.03%**
1	12.31%	8.38%
0	3.72%	5.42%

Table 2. The coverage table evaluated on the **TS** by considering the outputs of the three experts.

	Not-Dialogue	Dialogue
Audio	*8.3%*	*7.8%*
Camera motion	*43.7%*	*12.8%*
Face	*9.9%*	*23.9%*

Table 3. The relative improvements obtained by using the MES with respect to each single expert in the dialogue/not-dialogue classification of the shots.

	N = 3	*N* = 4
Correct Detections (*CD*)	**90.83%**	**88.78%**
False Alarms (*FA*)	7.89%	6.88%
Split Scenes (*SS*)	16.32%	13.39%
Merged Scenes (*MS*)	14.68%	17.89%
Correct overlap (*CO*)	82.64%	76.80%
False overlap (*FO*)	11.54%	6.67%

Table 4. There are reported the experimental results obtained by setting N = 3 and 4.

	Audio	*Relative improvement*
Correct Detections (*CD*)	*81.27%*	*11.8%*
False Alarms (*FA*)	*19.38%*	*59.3%*
Split Scenes (*SS*)	*23.31%*	*30.0%*
Merged Scenes (*MS*)	*21.34%*	*31.2%*
Correct overlap (*CO*)	*75.70%*	*9.2%*
False overlap (*FO*)	*17.72%*	*34.9%*

Table 5. The experimental results obtained by using the audio expert in the 2nd stage of the system with *N*=3; the relative improvement introduced by the use of the MES are also reported.

FIGURES

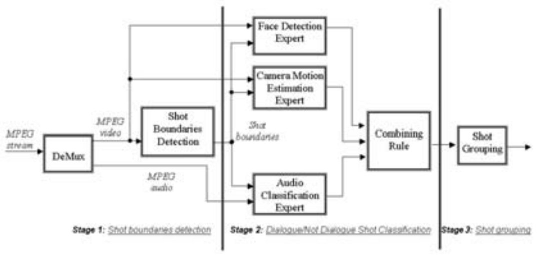
MPEG stream
DeMux
MPEG video
MPEG audio
Shot Boundaries Detection
Shot boundaries
Face Detection Expert
Camera Motion Estimation Expert
Audio Classification Expert
Combining Rule
Shot Grouping
Stage 1: Shot boundaries detection
Stage 2: Dialogue/Not Dialogue Shot Classification
Stage 3: Shot grouping

Fig. 1. Block diagram of the system for automatic detection of dialogue scene.

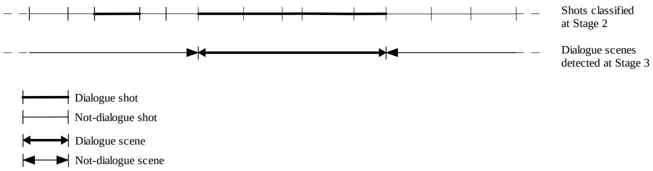
Shots classified at Stage 2
Dialogue scenes detected at Stage 3
Dialogue shot
Not-dialogue shot
Dialogue scene
Not-dialogue scene

Fig. 2. Examples of scene transitions in case of N = 3 are depicted.