Dialog Detection in Narrative Video by Shot and Face Analysis

B. Kroon[ab], J. Nesvadba[a], A. Hanjalic[b]
[a]Philips Research Europe, High Tech Campus 34, 5656 AE Eindhoven, The Netherlands
[b]Delft University of Technology, Mekelweg 4, 2628 CD, Delft, The Netherlands

ABSTRACT

The proliferation of captured personal and broadcast content in personal consumer archives necessitates comfortable access to stored audiovisual content. Intuitive retrieval and navigation solutions require however a semantic level that cannot be reached by generic multimedia content analysis alone. A fusion with film grammar rules can help to boost the reliability significantly. The current paper describes the fusion of low-level content analysis cues including face parameters and inter-shot similarities to segment commercial content into film grammar rule-based entities and subsequently classify those sequences into so-called shot reverse shots, i.e. dialog sequences. Moreover shot reverse shot specific mid-level cues are analyzed augmenting the shot reverse shot information with dialog specific descriptions.

Keywords: fusion, parallel shot, face detection, shot reverse shot, film grammar, dialog detection

1. INTRODUCTION

The explosive growth of commercially produced audiovisual content and its increasingly distribution through Internet portals, search engines, *Internet Protocol TV* (IPTV) and private or community networks did not only boost the interest for the content itself, but also in its associated content-awareness, hence, metadata. The latter are, in general, captured for commercially produced contents in scripts, which describe the story and which define the subdivision of the story into its semantic scene and, furthermore, into its narrative elements according to film grammar rules, as defined in [1] and [2]. These script based production rules result in production decisions, which are documented and archived, but unfortunately the current business models of content creators do not include sharing those data with third parties.

One of the major narrative elements in commercial content are narrative elements containing dialogues between two or more persons, in film grammar referenced as *shot reverse shots* SRS. The authors aimed to elaborate film grammar knowledge to boost the robustness for shot reverse shot sequence detection, a.k.a. dialogue sequence detection. In the next section the authors presented a state-of-the art survey about this topic and specified film grammar related rules for shot reverse shots. Here after, they presented in short methods for the detection of interleaved narrative events, statistics about shot reverse shots and a set of features applied for shot reverse shot detection, i.e. an omni-directional face detector with pose estimation. In section 4 they, subsequently, presented the results of the shot reverse shot detector using an audiovisual corpus and the paper was finalized by the conclusions in section 5.

2. SURVEY AND FILM GRAMMAR FOR SHOT REVERSE SHOT DETECTION

2.1 Film grammar for shot reverse shot sequences

The production of broadcast TV and cinema content underlie a handful of common conventions often referenced as *film grammar* [1]. Fortunately, almost every producer or director commits himself to follow the film grammar conventions during the production of multimedia content, which influence their production decisions. One of the techniques derived from film grammar is called *Mise-en-Scene*, i.e. French for 'putting into the scene', which covers from a cinematographic point of view all what a viewer sees, i.e. spatial compositions, settings, camera position, make-up, light settings and space-time relations.

Furthermore, *cinematography*, literally the *writing in movement*, provides the director with tools to manipulate the viewer's experience and to create (non-)uniform impressions, using e.g. range of tonalities, speed of motion, perspective relations and transformation of the perspective. *Perspective relations* can be applied to achieve e.g. the impression of 3-D spaces.

Multimedia Content Access: Algorithms and Systems, edited by Alan Hanjalic, Raimondo Schettini, Nicu Sebe,
Proc. of SPIE-IS&T Electronic Imaging, SPIE Vol. 6506, 65060K, © 2007 SPIE-IS&T · 0277-786X/07/$15

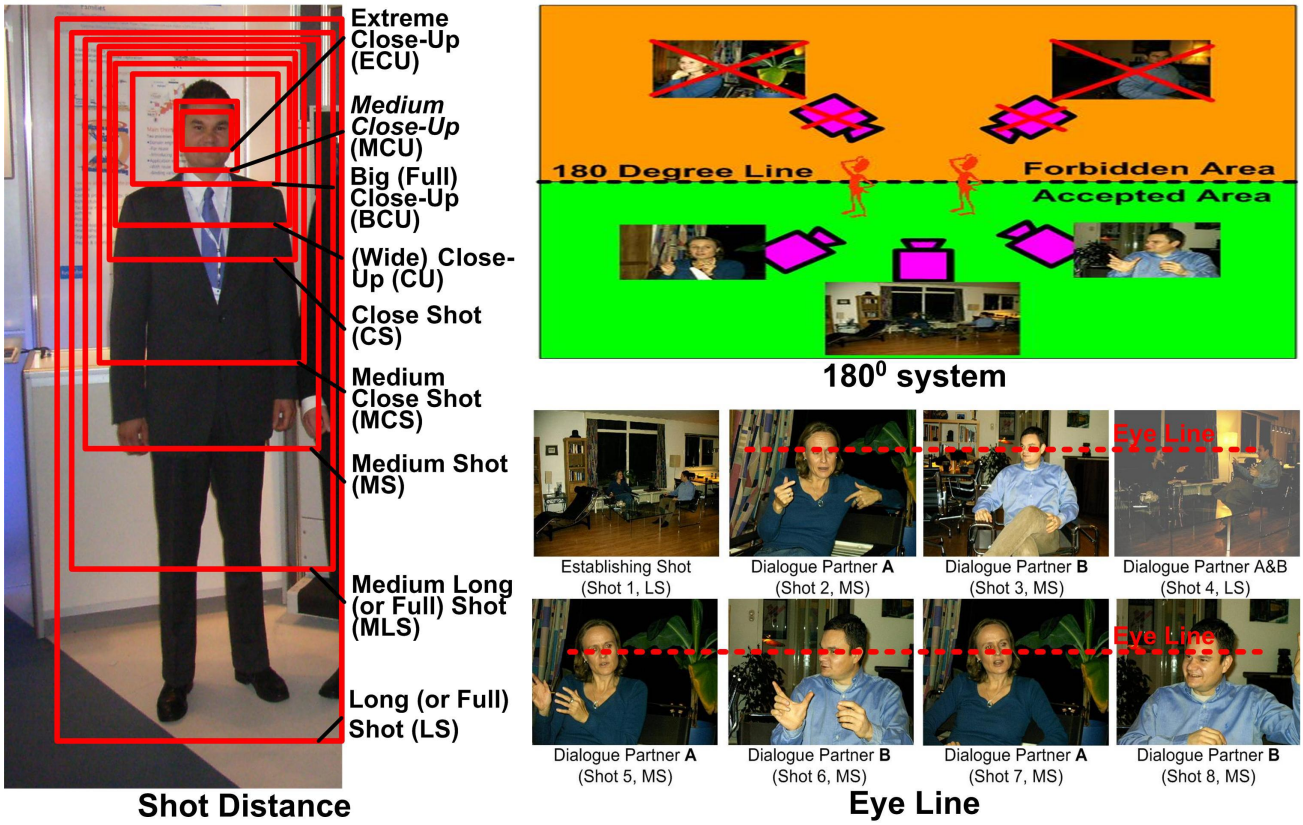

Figure 1. Film grammar based 180° system, eye line and shot distance.

Those relations deal with the spatial relation of objects/subjects inside a setting. Perspective relations are created e.g. with properly selected camera angles, heights and *shot distances*, i.e. the position from which a setting is captured in relation to the setting. The latter, i.e. the distance from which the shot is captured, is used to separate shots into *Long Shots* (LS), *Medium Long Shots* (MLS), *Medium Shots* (MS), *Medium Close Shots* (MCS), *Close-Ups* (CU), *Big Close-Ups* (BCU) and *Extreme Close-Ups* (ECU), as visualized in Figure 1 (left). For dialogue sequences mainly the following shot distance types are relevant. In *Medium Shots* (MS) object or subject of interest and its surrounding setting share equal frame areas, e.g. in the case of a standing actor the lower frame line passes through his/her waist, providing sufficient space to follow his/her gestures. Reducing further the distance leads to *Medium Close Shot* (MCS) level, wherein the lower frame line passes e.g. through the chest of an actor often used for a tight presentation of two persons. *Close-Ups* (CU) are covering extreme close distances, showing only e.g. the character's face and its shoulders in great detail so that it fills the screen. Those shots abstract the subject from the context. *Big Close-Ups* (BCU) show only an actor's forehead and chin, focusing the attention of the viewer on a person's feelings and reactions. They are sometimes used in interviews to show participant's emotional excitement state, grief or joy.

A viewer also expects uniformity what concerns *spatial relation*, hence, it is of importance to create first a sense of the scenery by means of an establishing shot or shots covering the scene-relevant locations and / or actors. Here after, spatial relations are secured following the *180° system*, as presented in Figure 1 (upper right). While watching a certain activity in a setting the viewer expects certain uniformity in terms of the camera's location, i.e. the action should take place along a so-called *axis of action* also referenced as *180° line*. Furthermore, it is important that the objects of interest, in our example the dialogue partners, have to be positioned spatial-conform inside the captured frame, i.e. spatially left located partners have to be present in the left side of the frames which captured them and the other way around. In addition, they have to face each other spatial-conform, i.e. facing the centre of the frame to support the impression of talking to each other. Furthermore, the convention for e.g. news programs is, that the camera is on eye-level with the anchorperson (*Eye-Level*). Particularly dialogues make use of this technique to maintain the eye-level when switching between speaker A and speaker B (*eye-line match)*, as published by Boggs in [3]. An example hereof was captured in Figure 1 (lower right).

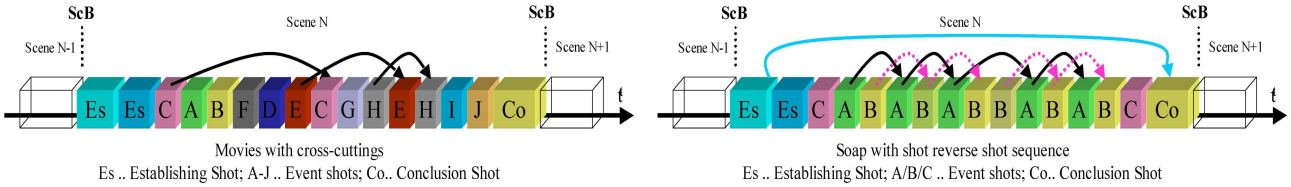

Figure 2. Interleaved narrative events – parallel shots.

But film grammar also specifies the setup of story elements, i.e. the composition and concatenation of shots representing sequential or parallel events. Normally each *narrative element* represents one individual event consisting of strongly related but not necessarily connected shots. Usually two or more narrative elements are interleaved with each other, forming *parallel shots*. The latter can be clustered into two groups, i.e. cross-cuttings and shot reverse shots. *Cross-cuttings* visualize, in general, either (a) time-wise correlated, location-wise disjoined parallel running narrative events, i.e. same time, but different location were interaction is not obligatory, or (b) time-wise uncorrelated events such as one event and a flash-back, i.e. different time at same or different location, as shown in Figure 2 (left). *Shot reverse shots* are used to visualize events such as a dialogue between to actors, i.e. same time at same location, captured from two or more camera positions and rendered in an interleaved manner, e.g. A-B-A-B, as shown in Figure 2 (right). In-between the interleaved sequences distant shots are used, e.g. an AB shot, to introduce spatial relations. Here after, A and B shots follow the spatial relation rules. Essential for *shot reverse shots* is to fulfill the earlier described eye-line match.

2.2 Existing dialog detection methods

In the past various attempts to develop dialog detection methods have been made and the methods differ a lot. Some were single modal while others were multimodal and within each modality, apparently a wide range of features could be applied to solve the dialog detection problem. The classification methods differed as well and varied in complexity from heuristic rule based approaches to finite state machines with many states. Table 1 provides a structured overview of the previous work found. An overview of basic features and definitions is available in [4].

Method	Modalities	Information available from pre-analysis	Classification model
Yoshitaka [5]	Visual	Shot length, shot dynamics, shot similarity	Rule based
Lienhart [6]	Visual	Clusters of faces by location and visual similarity	Rule based
Sundaram [7]	Visual	Shot similarity, shot length	Rule based
Chen [8]	Visual	Shot boundaries, faces (annotated)	Finite state model
Ying Li [9]	Auditory, Visual	Clusters of shots, parallel shot length	Rule based
Zhai [10]	Auditory, Visual	Motion intensity, audio energy, clusters of faces	Finite state model

Table 1. Overview of dialog detection methods.

In [5] an early system based on film grammar rules is presented. The paper described how low complexity algorithms were applied to detect cuts, shot lengths, shot dynamics, shot similarity and repetition. In [6] dialogs were detected in the video domain only by detecting and clustering faces and finding interlinked face groups. According to [7] in dialogs every second shot will be more similar than every adjacent or third one. Sundaram exploited this fact by designing a dialog detector based on shot color similarity and shot length only. A method that applied a finite state machine is [8], but shots and faces were marked manually. A rule-based method that applied both visual and auditory clues is [9]. In [10] a system for movie scene classification was proposed based on finite state machines. Features used were motion, audio energy and body similarity. Shots were compared by body color similarity instead of frame similarity. A "conversation" was defined as having low activity intensity, medium audio energy and multiple speakers.

The first papers focused on visual analysis of the shot structure. We have realized that although a dialog has a specific shot structure, this information is not sufficient to conclude that a parallel shot is a dialog, because there are also other non-dialog scenes that have a similar link structure. It is therefore important to find a different and complementary source of information to improve the quality of the dialog detection. Based on our knowledge of film grammar we concluded that the most important complementary source of information is face-related. Hence, the difference to methods like those published in [8][10] is the choice of the authors to apply exclusively more reliable low-level face-metadata, i.e. position, size and orientation. In this way we attempted to design a straightforward method for dialog detection.

3. DESCRIPTION OF THE METHOD

Figure 3 displays the system overview of the method. The main components were a shot linker that detected the parallel shot sequences, a pose-estimating face detector and a dialog detector component that computed all parallel shot-wide features and then classified parallel shots as dialog or not.

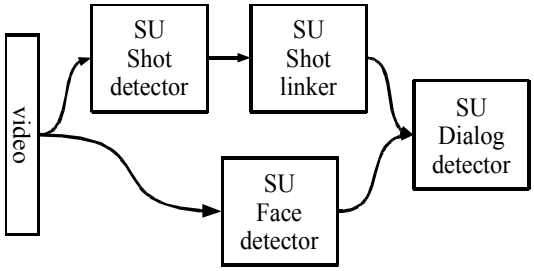

Figure 3. System overview at service unit [11] level.

All operations were performed in a streaming fashion [11] with the output known after a delay of several shots. The video encoder outputted video and audio frames. The face detector processed each frame and outputted for every face a list of properties. The shot linker selected a set of key frames for every shot and compared them with key frames from successor shots, as described next in 3.1. In the dialog detector component, the frame and shot-based information was processed to parallel shot-wise statistics, summarized in 3.2 and 3.3. On basis of these statistics the dialog detector classified the current parallel shot as either dialog (shot-reverse-shot) or not, described in 4. To allow evaluation of various features sets the actual amount of features extracted was initially bigger than finally applied for the solution.

The remainder of this section described the major components of the system in more detail. The first subsection described the shot linker that detected the parallel shots. The second one discussed the face detector and pose estimator and the final subsection provided a description of the set of features that the system extracted.

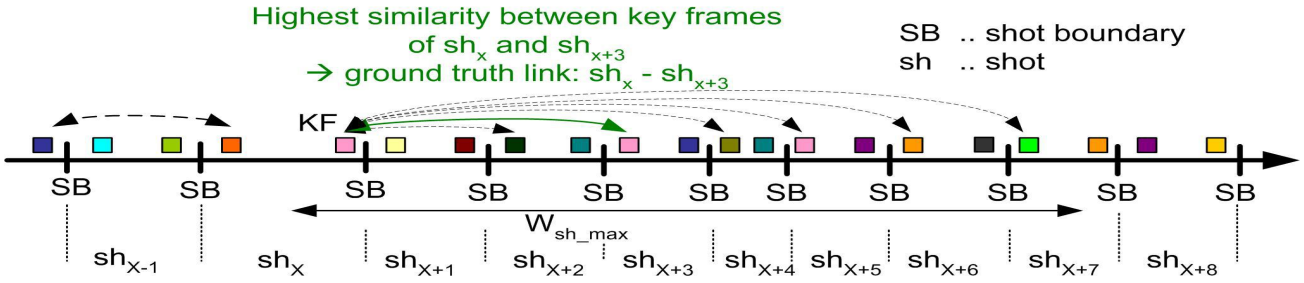

Figure 4. Parallel shot detection specific shot linker from [12].

Content	# of SRSs	# of shots in SRSs	# of GT SRS links	# of CCs	# of shots in CCs	# of GT CC links
Series total	92	871	631	29	185	81
Movies total	153	2271	1682	189	1390	661

Table 2. Ground truth statistics of SRSs and CCs of series and movies of AV corpus from [12].

Genre	Average length of SRS	Average length of CC	Ratio SRS links : number of SRS shots	Ratio CCS links : number of CC shots
Series	9.5 shots	6.4 shots	0.7	0.4
Movies	14.8 shots	7.4 shots	0.74	0.48

Table 3. Parallel shot statistics for series and movies from [12].

3.1 Parallel shot detection: ground truth and statistics for this work

The development of the algorithms described in this work required an objective benchmark set, i.e. a ground truth data set, which was derived from [12]. A corpus of 10 hours broadcast content, specified in [12], consisting of two genres, i.e. series and movies, were recorded, and, there after, non-content related inserts, i.e. commercial- and channel adds, were automatically indexed [11] and non-relevant boundary content, i.e. content before and after the series or movies, were identified by *electronic program guide* EPG time stamps. Both, inserts and boundary contents, were excluded from further analysis. Moreover, a shot boundary detector, described in [12] and [13] was applied to segment the remaining parts into individual shots *sh* with its shot boundaries *SB*, as shown in Figure 4 and summarized in Table 2. Subsequently, rules of film grammar were applied to cluster intentionally interleaved shots of two or more narrative events, i.e. dialogues, a.k.a. *shot reverse shots* SRS, and *cross-cuttings* CC, manually together to gain the ground truth for parallel shots. Knowing from film grammar that directors (editors) establish bridges between related, but not consecutive, shots through a continuation of the audiovisual story flow, the rule was established to visually compare a set of key frames per shot with a set of key frames of a set of successor shots, which were within a certain window of shots (W_{sh}). The analysis unveiled that producers avoided links longer than 6 shots to enable the viewer to recall the connection, hence a window length of $W_{sh}=7$ was applied For the manual annotation shots were linked together, whose key frame pairs exhibited the highest similarity within $W_{sh}=7$. For the manual ground truth annotation the cinematographic rule based key frame pair similarity, specified in [12], was applied, enabling an 'objective' analysis. The similarities were specified as, (a) strong correlation of foreground, i.e. region of interest, and/or (b) similarity of background, i.e. in film grammar terms the setting, with similar spatial layout. Because the analysis in [12] unveiled that virtual camera movements, i.e. non-captured camera zooming or panning, were happening between connected shots, the annotation rules covered as well links between key frame pairs exhibiting such virtual camera actions. Here after, interleaved links were clustered together, as in the case of the AB sequence in Figure 2 (right), into parallel shot sequences and indexed either as shot reverse shot, i.e. the presence of a dialogue between two or more actors, or cross-cutting. The *ground truth* GT results and statistics covering the series/movies content corpus were summarized in Table 2 and Table 3. In the latter columns two and three showed that in series about 60% of all shots were member of *shot reverse shot* SRS sequences (dialogues with an average duration of 9.5 shots) and ~13 % of *cross-cuttings* CC (average duration of 6.4 shots). On contrary, in movies only ~46% of all shots were members of shot reverse shots (with an high average duration of 14.8 shots) and almost ~30% of cross-cuttings (average duration of 7.4 shots). Finally, columns four and five in Table 3 unveiled that the ratio of parallel shot links compared to the number of member shots, further referenced as link fraction, was for both, series and movies, about ~0.7 for shot reverse shots and ~0.45 for cross-cuttings, i.e. link bridges in movies were almost twice the size in terms of shots compared to those in series.

The concept of linking individual shots together has been introduced in the late 90's, e.g. in [14] and [15] by clustering shots into story units by means of visual similarities and/or temporal rules. In [12] landmark point, i.e. Harris points and *scale invariant feature points* SIFT, similarity based methods were elaborated and benchmarked with various additional color based methods. The detection rates in [12] achieved for parallel shot detection recall and precision levels of about 85%/85%. In the remainder of this paper, the authors applied the manual annotated parallel shot ground truth for the subsequent analysis enabling a clear separation between the performance results of the parallel shot detector and the dialogue classifier.

3.2 Omni-directional face detection and pose estimation

Here after, several face-related mid-level features were generated applying a face detector and pose estimator analysis unit. By such we could determine the relative size, position and pose of the actors per frame and per shot. These frame properties were chosen by the director and are part of the film format as described in section 2.1.

The face detector we developed [16] is based on the successful Viola and Jones [17] method that combines the AdaBoost algorithm [18] [19] with easy to compute Haar-wavelet like features and a cascaded detector structure. The Haar-wavelet like features are efficient, overcomplete and in classification only slightly better than random. These features are used with the contemporary machine-learning algorithm AdaBoost that is well suited to form a strong learner out of many weak ones. Furthermore, by cascading the detector, the method was able to discard non-face patches after evaluating only part of the features. This increased the computational efficiency without lowering the performance much. The main benefits of this appearance-based method were its efficiency, effectiveness, and relative ease of training compared to other approaches.

Disadvantage was that the method is designed to detect frontal upright faces only. Many approaches have been published to circumvent this shortcoming [20] [21] [22] [23] [24]. A common approach was to combine multiple detectors into one multiview detector. [20] described how to use a binary classification tree to estimate the pose prior before choosing one of many pose-specific Viola and Jones-based detectors. We chose to concentrate on this method because we thought the other methods mentioned were less efficient and more difficult to train. This method however only delt with the twist angle (also called rotation in plane). In [20] the 360º twist range was quantized into twelve 30º bins, forming twelve classes. We extended [20] by additionally classifying five azimuth rotations: left/right profile, left/right semi-profile and frontal, creating a total of 5 × 12 = 60 pose classes. We also substituted the AdaBoost [18] learning method with GentleBoost [25]. Nevertheless we learned that even with reduction of the number of classes from 5 × 12 = 60 to 3 × 3 = 9 classes, it was infeasible to train the decision tree for profile faces.

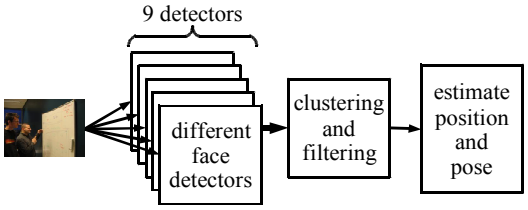

Figure 5. Face detector system overview.

Hence for this paper we decided to circumvent the decision tree by using all nine detectors in parallel without decision tree. Figure 5 shows the system overview of this detector system. For every image that is processed, first all single-view detectors were applied. The outputs of these detectors were a list of face bounding boxes where one face typically generated multiple responses all slightly offset. Next all bounding boxes from all detectors were collected and then clustered by size and position. As small clusters were likely to be false alarms, clusters with a size below a fixed threshold were removed. The remaining clusters all represented one face. By averaging the position and size of the bounding boxes within a cluster the final bounding box was retrieved. Furthermore, by taking into account which detectors provided the bounding boxes, it was possible to average the angles and get an estimate of the face azimuth and twist.

3.3 Suitable features for shot reverse shot detection

For classification of parallel shots as dialog or not the amount of information available was overwhelming. By finding statistical properties in this data we achieved to condense the necessary information in a small feature set that was suitable for our task. To find good features we tested several inspired by our knowledge from film grammar and our experience with manual video analysis. The features could be split into describing properties of faces and shots. All features were computed for entire parallel shots but most features were statistical properties of frame or shot-based observations. Table 4 provides an overview of the selected features.

Focusing on face information first, a film property that directly relates to face information is the shot distance classification as described in section 2.1. This is the relation between the size of the most prominent face in relation to the screen size and was described by fuzzy classifications such as “close up” and “medium long shot”. The shot distance classification was an important tool of the director. Dialogs were expected to have on average shorter shot distances than for instance action movies. We encapsulated this film grammar property in a feature *relative size of the subject (RS)* shown in equation (1) where $f_D(frame)$ was the dominant (biggest) face in the frame and width and height were measured in screen units.

$$RS(frame) = \begin{cases} \dfrac{width_{f_D(frame)} height_{f_D(frame)}}{width_{frame} height_{frame}} & frame\ has\ faces \\ 0 & otherwise \end{cases} \tag{1}$$

$$f_D(frame) = \arg\max_{face} \{ width_{face} height_{face} \mid face \in faces\ in\ frame \} \tag{2}$$

We selected the dominance face $f_D(frame)$ in a frame because we assumed that the intensions from the director were more clear by focusing on this face only. In a frame with multiple small faces and one big one, the smaller ones are further away and are a less important aspect of the scene. We took further advantage of the dominant face by computing

relative vertical and *horizontal position of the subject* (RP_x and RP_y) as shown in (3) respectively with subscripts x and y indicating horizontal and vertical position.

The horizontal position of the subjects in a shot is often chosen by the director to reflect their actual position in the studio. A camera aimed at a subject at the left side of the scene is aimed such that the subject appears on the left-side of the screen, and the other way around. This relates to the so-called 180º-rule as described in section 2.1 and visualized in Figure 1. To transform this frame feature into a parallel shot feature, we created a histogram with three bins (left, right and middle) and called this feature p_{RPy}. The vertical position RP_y of the subjects is important because in a dialogue according to the eye-line rule, see Figure 1, this should be the same for all shots. We expected the standard deviation of RP_y (σ_{RPy}) to be low for dialogs and high otherwise.

$$RP_x(frame) = \begin{cases} \frac{x_{f_D(frame)}}{width_{frame}} & frame\ has\ faces \\ 0 & otherwise \end{cases} \qquad RP_y(frame) = \begin{cases} \frac{y_{f_D(frame)}}{height_{frame}} & frame\ has\ faces \\ 0 & otherwise \end{cases} \tag{3}$$

The last property we extracted from our face detector was the azimuth *A,* which is the horizontal out-of-plane rotation of the face. In dialogs there was a high correlation between RP_x and A because most faces were rotated towards the center of the video frame. Exceptions were when people nod "no" or (act to) look away in disgust. With a motivation similar to before we decided to create a three bin histogram p_A.

Concerning shot information, we knew that dialogs have a clear ABAB pattern sometimes disturbed by unrelated interleaved shots. Causes for disturbances were directing choices and mistakes from the parallel shot algorithm. One would expect that in dialogs the distance between similar shots (*link distance*) is most often two. In sequential shots the dominant value for the link distance was one because consecutive shots are most similar. For other patterns in video there are either no links or the links were more irregular. The shot-reverse-shot structure was easy to spot but was not always used in a strict sense. The ABAB structure was often interleaved with a small number of other shots and our measures have to be robust to these fluctuations. Because we expected that in a dialog most link distances are two we design a feature (p_{LD}) by making a histogram with three bins expressing links of 1, 2 and 3 shot length. To test if this is the best solution we also designed some alternative representations. Comparing dialogs with non-dialogs we expected that dialogs have more links and are more ordered. We encapsulate this observation in a measure called the *link fraction* (*LF*), which indicates the amount of links in the parallel shot compared to the maximum amount (4). The third link-based feature measures how well-structured the parallel shot is by measuring the standard deviation of the link distance (σ_{LD}). In dialogs the standard deviation should be low.

$$LF = \begin{cases} \frac{number\ of\ links}{number\ of\ shots - 1} \end{cases} \tag{4}$$

Besides the shot links the duration of the shots could also be indicative for dialogs. We therefore measured the dominant value (peak in the distribution) of the shot duration (SD_{dom}) and the conditional variance (SD_{cv}) to describe the distribution of the shot lengths. We estimated the dominant value with the position of the highest peak of an histogram where we chose the number and position of the bins automatically to match the data. To describe the size of a parallel shot we could both use the parallel shot duration (*PSD*) or the shot count (*SC*).

Abbreviation	Description
LF	Link fraction [0, 1] is the fraction of the shots in a parallel shot that have links
$cv_{LD}/\mu_{LD}/p_{LD}$	Conditional variance (σ/μ)/mean/rel. frequencies of the link distances (in nr. of shots) in a parallel shot
SD_D/cv_{SD}	Dominant value/conditional variance of the shot duration [s] in a parallel shot
ED	Duration [s] of the entire parallel shot
SC	Number of shots {3, 4…} in a parallel shot
$G_D/\mu_G/p_G$	Dominant value/mean/relative frequencies of the nr. of faces per frame {0, 1…} in a parallel shot
p_{RPx}	Relative frequencies of the relative horizontal position [0, 1] of the dominant subject in a parallel shot
σ_{RPy}	Standard deviation of the relative vertical position [0, 1] of the dominant face in a parallel shot
RS_D/cv_{RS}	Dominant value/conditional variance for the relative size [0, 1] of the dominant face in a parallel shot
p_A	Relative frequencies (-30°, 0°, 30°) of the face azimuth [°] of all dominant faces in a parallel shot

Table 4. Overview and short descriptions of the features taken into consideration

4. EVALUATION AND RESULTS

We evaluated our system on the dataset described in section 3.1. The evaluation strategy is twofold. First we compared different feature combinations in order to discuss the quality of the features and possible combinations. Second we performed additional tests on the preferred feature combination.

We tried various classifiers but for most combinations the linear discrimant classifier performed better than the others. We also tried the nearest mean classifier, the first and second order polynomial support vector classifier, the nearest neighbor classifier (1-NN), the quadratic discrimant classifier, a decision tree with various options for pruning, and even AdaBoost [18]. The reason that the more complex algorithms failed is probably because although the footage has a total length of several hours, the number of parallel shots (samples) in the dataset was too small to design complex decision boundaries. It surprised us however that a decision tree classifier performed poorly because most of the previous work utilizes rule-based classification mechanisms.

For this paper all results published were therefore based on a linear discrimant classifier. All our tests are performed with PRTools [22].

We tested multiple combinations of features and tested them using leave-one-out cross validation on the entire dataset. This provided the most accurate classification error estimation possible. We tested all features separately and then we manually included feature combinations based on single-feature performance and our expectation of how complementary features would be. Table 5 shows the result of this first test. We omitted solutions with a high classification error to shorten the table.

In all tests (Table 5) and pre-final experiments we found that the link fraction was the most valuable individual feature with the conditional variance of the link distance a good second with classification errors of 0.238 and 0.241 respectively. This proved that the shot link information is the most important source of information for a dialog detector, even in such a diverse dataset as ours. We expected that the addition of face information would improve the classification result and we especially expected p_{RPx} and p_A to be the best contributions but the latter was not the case with the best two-feature combination containing one of these features having a classification error of 0.227. In contrary to our assumptions, just counting faces provided the best contribution as $LF+p_G$ was the best two-feature combination for which we found a classification error of 0.174. Table 5 shows that the addition of p_{RPx} provided a small improvement but we expected this improvement to be insignificant and in favor of not tuning on our dataset we selected $LF+p_G$ as our preferred solution. In Table 7 we displayed the coefficients of the linear discrimant classifier that were determined. The coefficients form a linear mapping such as used in [22].

We further investigated the preferred combination by measuring precision and recall for various thresholds. Figure 6 shows the precision-recall curves for all sequences, all series, and all movies respectively. The small dots are threshold values and the single thick dot in every graph is the equal error rate threshold.

Table 8 is a compilation of results of the surveyed papers. None of the datasets used were equal and the definition of the samples varied too. We like to note that all papers tested with smaller datasets and besides [10] all papers published precision/recall for separate content items and did not test on series. From our experience the characteristics of series and movies are different and limits the range of acceptable solutions.

5. CONCLUSIONS

The combination of the parallel shot detector and a face detector provided a straightforward classification system of which the results were easy to interpret. This was an advantage of our system over the referenced work with more difficult to interpret systems like [8]. We evaluated our system on an 10 hour dataset with narrative video of different genres. In comparison to other papers our precision and recall seem to be lower. However all referenced papers test with smaller datasets, and under different assumptions and providing less details on the steps actually taken.

The previous work discussed seemed to have similar definitions of a dialog as did this paper. However we found that the definition of a dialog may have to be refined. In long parallel shots only part of the parallel shot was actually a dialog. This happened for instance if a parallel shot was a crosscutting of an action scene and a shot-reverse-shot dialog. A more detailed annotation and an algorithm that provides a more detailed (shot-based) analysis can provide new insights. Such an algorithm should be able to describe crosscutting, sequential shots, shot-reverse-shot, and other elements within one parallel shot and how they are combined.

Feature	Error	Feature	Error	Combination	Error	Combination	Error
G_D	0.236	μ_G	0.325	$LF + p_G + p_{RPx}$	0.172	$SC + p_G$	0.236
LF	0.238	SC	0.331	$LF + p_G$	0.174	$LF + SC + p_{RPx}$	0.236
LD_{cv}	0.241	ED	0.364	$LF + SC + p_G + p_{RPx}$	0.179	$LF + SC$	0.241
p_{LD}	0.245	σ_{RPy}	0.413	All features	0.183	$SC + p_G + p_{RPx}$	0.252
pG (5 bins)	0.245	p_A	0.428	$LF + SC + p_G$	0.185	$p_G + p_{RPx}$	0.258
pG (3 bins)	0.247			$LF + p_{RPx}$	0.227		

Table 5. Leave one out cross validation result for a linear discriminant classifier.

Selection	Genre	Main language	Parallel shots	Recall	Precision	F1-test
All sequences	Mixed	Mixed	453	0.849	0.83	0.84
All series	Mixed	Mixed	113	0.941	0.80	0.87
All movies	Mixed	Mixed	340	0.812	0.84	0.83
Serie 1	Family/comedy	English	40	1.000	0.77	0.87
Serie 2	Drama/romance	German	12	1.000	0.67	0.80
Serie 3	Drama/romance	German	23	1.000	0.92	0.96
Serie 4	Comedy	Dutch	21	0.667	0.40	0.50
Serie 5	Drama/romance	Dutch	17	0.857	0.86	0.86
Movie 1	Drama/sci-fi	English	110	0.588	0.91	0.71
Movie 2	Drama/sci-fi	English	74	0.913	0.91	0.91
Movie 3	Comedy	German	71	0.778	0.93	0.85
Movie 4	Drama/family	German	26	0.78	1.00	0.86
Movie 5	Drama	English	59	0.95	0.69	0.80

Table 6. Precision and recall with separate training and test-sets (no cross validation).

	LF	G: P(0 faces)	G: P(1 face)	G: P(2 or more faces)
mean vector	0.7497	0.2886	0.4948	0.2166
covariance matrix	0.0174	-0.0044	0.0044	0.0000
	-0.0044	0.0427	-0.0188	-0.0239
	0.0044	-0.0188	0.0233	-0.0044
	0.0000	-0.0239	-0.0044	0.0284

Table 7. $LF + p_G$ linear discrimant classifier coefficients

Paper	Dataset	P	R	F1	H	M	FA
Sundaram [7]	5 movies	95	85	90	75	13	4
Li [9]	3 movies	100	96	98	49	2	0
Chen [8]	3 movies	81—92	91—97	85—94	—	—	—
Lienhart [6]	2 movies	95	87	91	1364	203	65
Zhai [10]	80 scenes	97	94	96	—	—	—
Yoshitaka [5]	3 movies	77	90	83	—	—	—

Table 8. Score of other papers on precision and recall on own datasets

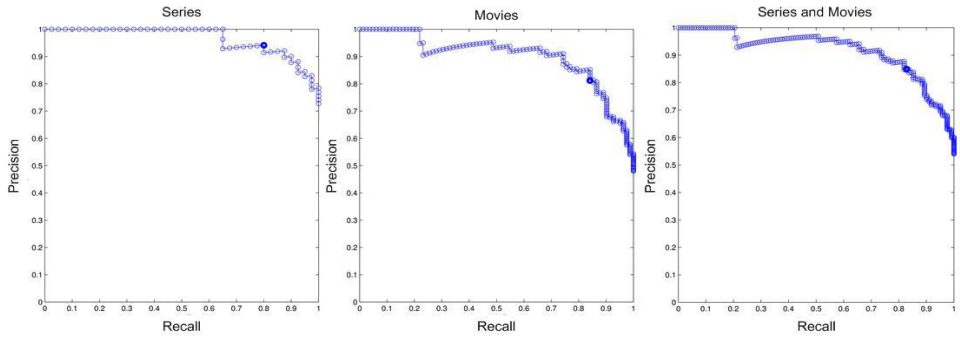

Figure 6. Recall and precision results for dialogue detection.

REFERENCES

1. D. Bordwell, and K. Thompson, “Film Art: An Introduction”, McGraw-Hill, ISBN 0-07-248455-1, 2004.
2. F. Beaver, ‘Dictionary of Film Terms: The Aesthetic Companion to Film Analysis’, Twayne-McMillan, ISBN: 0805793348, 1994.
3. J.M. Boggs, D.W. Petrie, ‘The Art of Watching Films’, Manfield Publishing Company, 5th edition, 2000.
4. Y. Wang, Z. Liu, and J.-C. Huang, “Multimedia Content Analysis”, IEEE Signal Processing Magazine, November, 2000.
5. A. Yoshitaka, T. Ishii, M. Hirakawa, and T. Ichikawa, “Content-based retrieval of video data by the grammar of film”, IEEE Symposium on Visual Languages, 1997.
6. R. Lienhart, S. Pfeiffer, and W. Effelsberg, “Scene determination based on video and audio features”, in *Proceedings of IEEE Conference on Multimedia Computing and Systems*, Florence, Italy, 1999.
7. H. Sundaram, and . Chang, “Determining computable scenes in films and their structures using audio-visual memory models”, in *Proceedings of Eighth ACM International Conference on Multimedia*, October, 2000.
8. L. Chen, and M. Tamer Özsu, “Rule-based scene extraction from video”, in *Proceedings of IEEE International Conference on Image Processing*, 2002.
9. Y. Li, S. Narayanan, and C.-C.J. Kuo, “Content-based movie analysis and indexing based on audio visual cues”, in *IEEE Transaction on Circuits and Systems for Video Technology* 14(8), 1073—1085 (2004).
10. Y. Zhai, Z. Rasheed, and M. Shah, “Semantic classification of movie scnenes using finite state machines”, in *IEE Proceedings of Vision, Image and Signal Processing* 152(6), 896—901 (2005).
11. J. Nesvadba, et al., ‘Real-Time and Distributed AV Content Analysis System for Consumer Electronics Networks’, Proc. Int. Conf. for Multimedia and Expo (ICME 2005), Amsterdam, The Netherlands, June 6-8, 2005.
12. J. Nesvadba, ‘Semantic Segmentation of Audiovisual Content’, PhD thesis, Labri, Univ. of Bordeaux, 2007.
13. J. Nesvadba, F. Ernst, J. Perhavc, J. Benois-Pineau, L. Primaux, ‘Comparison of Shot Boundary Detectors’, Poster , Int. Conf. for Multimedia and Expo (ICME 2005), pp 788-791, Amsterdam, The Netherlands, June 6-8, 2005.
14. M.M. Yeung, B.-L. Yeo, Video Visualization for Compact Presentation and Fast Browsing of Pictorial Content, IEEE Transactions on Circuits and Systems for Video Technology, 7(5), 771—785, October 1997.
15. A. Hanjalic, R.L. Lagendijk, J. Biemond, Automated High-Level Movie Segmentation for Advanced Video-Retrieval Systems, IEEE Transactions on Circuits and Systems for Video Technology, 9(4), 580—588, June 1999.
16. J. Nesvadba, A. Hanjalic, H. Broers, E.A. Hendriks, B. Kroon, H. Çelik, P. Fonseca, “Towards a realtime, distributed system for face detection, pose estimation and gaze recognition”, *International Conference on Methods and Techniques in Behavioral Research*, Wageningen, July 2005.
17. P. Viola, M. Jones, “Rapid object detection using a boosted cascade of simple features”, in *Proceedings IEEE CVPR*, Hawaii, December, 2001.
18. R. Schapire, “The Strength of Weak Learnability” Machine Learning, Vol. 5, No. 2, pp 197-227, 1990.
19. Y. Freund, “Boosting a Weak Learning Algorithm by Majority”, Information and Computation, Vol. 121, No. 2, pp 256 – 285, 1995.
20. M.J. Jones, P. Viola, "Fast Multi-view Face Detection", IEEE Conference on Computer Vision and Pattern Recognition (CVPR), June 2003.
21. S. Li, L. Zhu, Z. Zhang, H. Zhang, “ Learning to Detect Multi-View Faces in Real-Time”, Proc. Int. Conf. On Development and Learning, Washington DC, June 2002.
22. C. Huang, H.Z. Ai, Y. Li, S.H. Lao, “ Vector Boosting for Rotation Invariant Multi-view Face Detection”, IEEE International Conference on Computer Vision (ICCV), vol. 1, pp. 17-20, October 2005.
23. Y.Y. Lin, T.L. Liu, “Robust Face Detection with Multi-Class Boosting”, IEEE Conference on Computer Vision and Pattern Recognition (CVPR), vol. 1, pp.680-687, June 2005.
24. S. Li, Z. Zhang, “FloatBoost Learning and Statistical Face Detection”, IEEE Transaction on Pattern Analysis and Machine Intelligence, Vol. 26, No. 9, 2004.
25. R. Lienhart, A. Kuranov, V. Pisarevsky, “Empirical Analysis of Detection Cascades of Boosted Classifiers for Rapid Object Detection”, Technical Report, MRL, Intel Labs, 2002.
26. PRTools, available at http://www.prtools.org.