AUTOMATIC CLASSIFICATION OF FIELD OF VIEW IN VIDEO

Maria Zapata Ferrer, Mauro Barbieri, Hans Weda

Philips Research Europe
High Tech Campus 34, 5656AE Eindhoven, The Netherlands
Email: {maria.zapata.ferrer, mauro.babieri, hans.weda}@philips.com

ABSTRACT

Automatic systems are needed for audiovisual databases to efficiently index, browse, summarize and retrieve, because the amount of stored data is increasing tremendously. Historically film production techniques, have developed, in part, to convey e.g. meaning or atmosphere to the viewer. By studying these techniques, established guidelines for conveying meaning may be incorporated into automated tools for video analysis. In the current paper we present an approach in this area to classify different shot types, such as long shots, medium shots and close ups, which are important elements of video production. Based on a set of features calculated from the audiovisual content (e.g. presence of camera motion and size of detected faces), a Bayesian classifier distinguishes between six different shot types. The performance of this novel generic field of view classifier in terms of precision and recall is promising.

1. INTRODUCTION

Recent advances in computing, communications and data storage have led to a tremendous growth of large digital archives in both the professional and the consumer environment. Because these archives are characterized by a steadily increasing capacity and content variety, finding efficient ways to index, browse, summarize and retrieve stored information of interest is of crucial importance.

Traditional content analysis aims at automating these tasks by using computer vision and signal processing techniques to understand video at a semantic level. However, in professionally created video, semantics are also influenced by filming techniques and editing operations. Professional video production uses certain common conventions often referred to as “the grammar of film”. Among the various film grammar rules used by content producers to convey meaning through video, the *field of view* plays a very important role [1][2]. It is determined by the size of a subject in relation with the overall frame, which depends on the distance of the camera from the subject, and the focal length of the lens used. Based on field of view, shots can be classified into different *shot types* usually labelled by how big and how near an object appears to the viewers: for example, *long* shot, *medium* shot, and *close-up* shot (see Figure 1).

Long shots are employed to establish all elements in a scene so that viewers will know the actors involved and where they are located. Medium shots, which represent the bulk of most productions, are used to emphasize what the subjects are doing, while still allowing the audience to see their facial expressions. Close-up shots are used to play up narrative highlights such as important dialogues, subjects’ actions or reactions, and focus attention on a person’s feelings [1].

From these examples it is clear that the simple knowledge of the shot type allows a level of understanding of the video that otherwise would be very difficult to achieve. Multimedia applications such as video indexing and summarization could certainly benefit from an algorithm capable of automatically classifying video shots based on the field of view.

The problem of automatic classification of field of view has not been addressed in a generic way. In summarizing soccer and baseball games, Ekin et al. [3][4] have used the dominant colour of the playing field to classify shots into three types. Kumano et al. [5] have presented a method to automatically classify shots depending on the field of view based on strict assumptions on the structure of the video that are in practice rarely verified if no restrictions are made on the genre of the video. These assumptions are translated in fixed rules that the system uses to classify shots in close-up, medium and long.

The importance of considering media production rules for building effective indexing, searching and browsing multimedia applications is recognized in the emerging computational media aesthetics research area [6][7]. Although field of view is one of the most important production elements, the problem of shot type classification has not yet been addressed.

We propose a new generic algorithm for classifying video shots according to field of view that is applicable to any video genre and does not require specific assumptions. A set of features is extracted from the audiovisual signal and used to automatically classify video shots into six different categories.

The rest of the paper is organized as follows. In Section 2 we define the different shot types and we present a set of potentially discriminating features. In Section 3 we describe the implemented method and system for automatic classification of video into shot types and in Section 4 we discuss the results. We devote Section 5 to conclusion and introduction of future work.

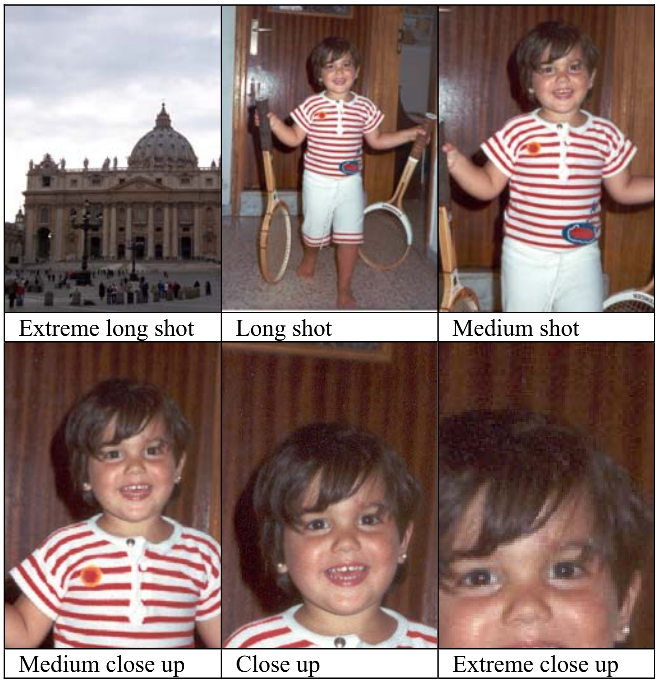
2. FIELD OF VIEW

2.1. Shot types definition

Regarding the field of view, different categories and terminology are used in the film industry and across literature. Generally the shots are classified by how big an (arbitrary) object appears in the frame and how near it appears to the viewers [1][2]. Although shot types can be defined in such a general way, for practical reasons we base our definition on shots that contain people. We define six different shot types with respect to which part of a person's body they contain:

- *Extreme long shot (ELS*): showing twice a person's body length or more, for example a house or an entire block of houses.
- *Long shot (LS)*: showing a person's entire body length.
- *Medium shot (MS)*: showing a person from the knees up, or from the waist up.
- *Medium close up (MCU)*: showing a person's head plus shoulders or upper arms.
- *Close-up (CU)*: showing just a person's head or face.
- *Extreme close-up (ECU)*: showing a part of the face or less, such as just an eye or a person's fingertips on a keyboard.

Examples of each shot type are shown in Figure 1.



Extreme long shot
Long shot
Medium shot
Medium close up
Close up
Extreme close up

Figure 1: *Different shot types***.**

2.2. Features

Apart from the part of the body that is shown in the frame, many other features characterize shot types. We performed an empirical study of Hollywood films and analysed existing film production theory [1][2] to elicit the common characteristics within a shot type. Based on this study we present below an initial set of potentially discriminating features and how they have been computed from the audiovisual signal:

- *Face size:* if a shot shows persons, then the face size is a strong discriminating feature. We perform face detection for each video frame [8] and we compute the ratio between the area of the detected faces and the size of the frame. In case multiple faces are detected, we consider only the biggest face.
- *Number of faces:* number of faces detected in a video frame, obtained from the face detector [8].
- *Face distance from the centre:* Euclidean distance from the centre of the detected face to the centre of the video frame. In case of multiple detected faces, we consider only the biggest face. When the distance is large, the shot is more likely to be a LS.
- *Shot duration*: the type of a shot tends to influence the typical duration of a shot. For example, ELS usually lasts longer than CU shot because they contain more details and require more time for being assimilated. We perform shot cut detection using a histogram-based method [9] and we consider the duration of each shot in seconds.
- *Camera motion:* based on the observation that camera motion is more common in certain shots (e.g. LS) than in others (e.g. CU), we perform camera motion estimation using the luminance projection correlation method [10]. We use as features the horizontal and vertical panning factors that are the displacements in pixels between successive frames.
- *Entropy:* ELS and LS usually present more visual details than CU or ECU shots; therefore we estimate the visual complexity of a video frame by calculating its entropy from the luminance histogram.
- *Motion statistics:* knowing if the objects of a scene are moving can be useful to characterize long shots since they are usually employed to depict action scenes. We take object motion into consideration by calculating the standard deviation and the average of the magnitude of the motion vectors for each input frame, assuming the input material is in MPEG format.
- *Audio features:* since music is often used in LS or ELS whereas speech is more common for MS, MCU and CU, we perform audio classification using the statistical classifier developed in [11] and we consider as features the probabilities of silence, music, speech, noise (background noises) and crowd noise (e.g. applauding or cheering) for each shot.

3. METHOD

3.1. System design

Based on the list of discriminating features in the previous section, we propose a system that can automatically identify the field of view of each shot. Such a system generally consists of two main building blocks: the feature extraction and the classification. The feature extraction computes a set of features from the audiovisual content. This set is used as input of a statistical classifier that classifies video shots into one of the mentioned categories. A general overview of the system is shown in Figure 2.

Some of the extracted features mentioned in the previous section might be correlated. This means that these features characterize the different shot types in the same way. It is important to find the correlated features because they unnecessarily complicate the classifier, and more data is needed for the training.

We have determined the correlation values across all extracted feature vectors. Considering these correlation values and some tests with a preliminary version of the classifier, several features were discarded. It was also decided to merge the horizontal and the vertical panning factor into a binary feature that only indicates if there is panning or not. Eventually the feature vector used in the experiments contained 9 features (see Figure 2): *face size*, *face distance from the centre*, *shot duration*, *entropy*, *panning*, *standard deviation of the motion vectors (MV σ)*, *probability of silence*, *probability of speech*, and *probability of noise*. More details on the system components shown in Figure 2 are discussed in the next sections.

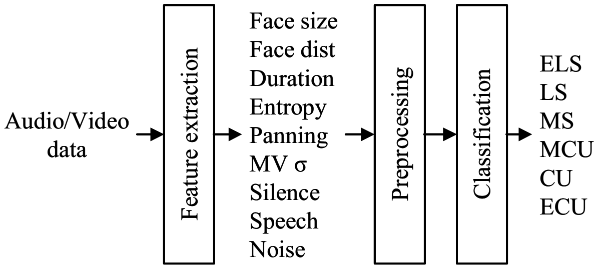


Figure 2: General pattern classification model applied to the combined shot type classification.

3.2. Preprocessing

All of the features, except for the *shot duration*, are calculated on a frame basis, which means that results are given for every frame. As the classifier has to work on shot level, and not on frame level, the frame-level features have to be converted to shot-level features. In most cases simply averaging results per frame within a shot does this. However, the motion vectors are not available for each frame, and faces might not be present or be detected in each frame of a shot. Therefore, the conversion of the *size of the faces*, the *face distance from the centre* and the *standard deviation of the motion vectors*, is calculated by averaging only the results of the frames in which these features are present.

In case of the *camera motion*, the binary value 1 was assigned to each shot with at least one frame with non-zero panning factors; in any other case 0 was assigned.

3.3. Classifier

The classifier uses the feature vector normalized in the previous block to assign the input data to the different classes. A Bayesian classifier using Gaussian densities was chosen to perform the classification. A Bayesian classifier is a supervised learning system, which has to be trained with data to correctly set the involved internal parameters. Though other methods based on neural networks or support vector machines [12] could have been selected, we chose this model for simplicity and because it has been widely used to address similar pattern recognition problems.

Two different approaches have been investigated and implemented using the Bayesian method: a single classifier to distinguish among all classes (combined classifier), and six individual classifiers, each able to identify a different class (individual classifier). Instead of having only one classifier with six Gaussians, in the second approach, the system is built by combining six individual classifiers with one Gaussian for each of them (Figure 3). Single Bayesian classifiers have simpler decision boundaries and they can potentially improve performances with a given limited training set. Furthermore, this second approach allows optimising each individual classifier by selecting the most differentiating set of features for each different class.

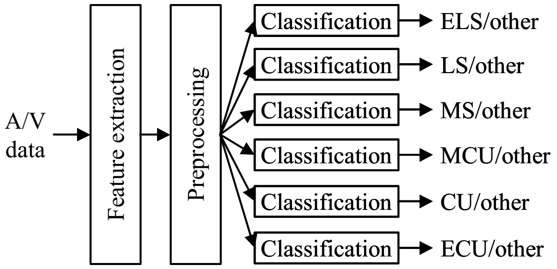


Figure 3: General pattern classification model applied to the individual shot type classification.

4. RESULTS

The performance of the combined and individual classifiers have been evaluated using 90 minutes of video content, of which 65 minutes are from the film "Charlie's Angels" and the rest are news and TV advertisements. The video content is manually annotated for shot types; the number of shots for each shot type is shown in Table 1. The difference in the amount of shots for the shot types is taken into account in the Bayesian classifier by setting the prior probabilities accordingly [12].

Descriptor	Annotated shots
Extreme long shot	127
Long shot	237
Medium shot	325
Medium close up	670
Close up	172
Extreme close up	44
Total	1575

Table 1. Annotated video samples.

Each experiment consisted of 10 runs. In each run the classifier was trained using randomly chosen samples. These training samples consisted of 90% of the annotated shots. Subsequently the classifier was tested using the rest of the shots. To be able to compare all the tests in the same conditions, all the experiments were run with the same 10 sets of randomly chosen samples. The final result of the classification was obtained by averaging over the 10 runs. By looking at the spread over the 10 runs the consistency of the classifiers can be estimated.

Table 2 shows the classification results in precision and recall for both the combined and the individual classifiers. Since the spread in the results over the 10 runs is below 10% in all cases, the classifier is reasonably consistent. The overall performance of this first generic classifier is promising. In particular when faces are detected in the video material, the results are good.

Regarding precision, both methods perform similarly whereas the *individual classifier* achieves higher values of recall for almost every case, especially for *ELS*, *LS* and *MS*. For the *individual classifier*, *ELS*, *LS* and *ECU* present very high recall (over 80%), but rather low precision (less than 30%), which means that many of these shot types are detected but also a lot of other shots are misclassified into these classes. On the other hand *MCU* have rather high precision but lower recall, so when they are detected, they are almost always correctly detected. Finally *MS* and *CU* give better average results, because both precision and recall are neither high nor low.

	Precision (%)		Recall (%)	
Shot type	**Combined**	**Individual**	**Combined**	**Individual**
ELS	18	15	45	86
LS	40	26	38	82
MS	39	38	37	63
MCU	76	63	34	20
CU	39	40	44	47
ECU	12	10	73	76

Table 2. Results in precision and recall for the combined and individual classifiers.

The classifier tends to classify many shots *approximately* right, for example *MCU* is often classified as *MS* or *CU*. Therefore, merging classes may improve the performance, but decrease the differentiating power. The results for the merged classes are shown in Table 3.

	Precision (%)		Recall (%)	
Shot type	**Combined**	**Individual**	**Combined**	**Individual**
ELS/LS	37	38	87	89
MS/MCU	89	82	48	45
CU/ECU	39	35	39	53

Table 3. Results in precision and recall for the combined and individual classifiers in case of merged classes.

Most results improve with respect to the results obtained for the separate classes. Precision and recall are even approaching 90%. Only the results on the merged *CU/ECU* class do not reach the same level of performance. The used set of features apparently does not reflect the similarities in the *CU* and *ECU* shot type.

5. CONCLUSIONS AND FUTURE WORK

By acquiring insight in film grammar rules, advances can be made in automating summarization, browsing and retrieving information in large video collections. Field of view is an important element for such an approach. In the current paper we have presented a novel generic algorithm for field of view classification. We have shown that a Bayesian classifier, based on a set of numeric features extracted from the audiovisual signal, can distinguish among six different shot types. Additionally we have shown that specific shot types can be distinguished using individual classifiers.

Our results are promising. In general the individual classifier performs better than the combined classifier. When shot types are merged, the precision and recall approach 90%, although the distinguishing power obviously decreases.

This novel generic field of view classifier can be improved in several ways. Incorporating new features and enlarging the training set will improve the performance. As the face size is an important feature, and we have used a detector that works only for frontal faces, employing an omni directional face detector is expected to improve the performance significantly.

6. ACKNOWLEDGEMENTS

We would like to thank Martin McKinney, Jeroen Breebaart, Fabian Ernst, and Vasanth Philomin from Philips Research for their assistance in the feature extraction software.

7. REFERENCES

[1] J. V. Mascelli, *The Five C's of Cinematography – Motion Pictures Filming Techniques*, Silman-James Press, Los Angeles, USA, 1965.

[2] W. H. Phillips, *Film – An Introduction*, Bedford St. Martin's, USA, 1999.

[3] A. Ekin, A. Murat Tekalp, and R. Mehrotra: "Automatic Soccer Video Analysis and Summarization", *Proc. of the Int. Conf .on Electronic Imaging*, Santa Clara, USA, pp. 339-350, 2003.

[4] A. Ekin and A. Murat Tekalp, "Shot type classification by dominant color for sports video Segmentation and summarization", *Proc. of the Int. Conf. on Acoustic, Speech and Signal Processing, ICASSP 2003*, Hong Kong, Vol. 3, pp. 173-176, 2003.

[5] M. Kumano, Y. Ariki, K. Tsukada, K. Shunto, "Automatic shot size indexing for a video editing support system", *Proc. of the Third Int. Workshop on Content-Based Multimedia Indexing, CBMI 2003*, Rennes, France, September 2003.

[6] C. Dorai and S. Venkatesh, *Media Computing - Computational Media Aesthetics*, Kluwer Academic Publishers, 2002.

[7] C. Dorai and S. Venkatesh, "Bridging the Semantic Gap with Computational Media Aesthetics", *IEEE Multimedia*, pp. 15-17, April-June 2003.

[8] P. Viola and M. Jones, "Rapid object detection using a boosted cascade of simple features", *Proc. of the Int. Conf. on Computer Vision and Pattern Recognition, CVPR 2001*, Kauai, USA, pp. 511-518, 2001.

[9] R. Lienhart, "Comparison of Automatic Shot Boundary Detection Algorithms", *Proc. of Storage and Retrieval for Image and Video Databases VII*, San Jose, USA, vol. 3656, pp. 290-301, 1999.

[10] K. Uehara, M. Amano, Y. Ariki, M. Kumano, "Video shooting navigation system by real-time useful shot discrimination based on video grammar", *Proc. of the Int. Conf. on Multimedia & Expo, ICME 2004*, Taipei, Taiwan, Vol. 1, pp. 583-586, 2004.

[11] M.F. McKinney and D.J. Breebaart, Features for audio and music classification, *Proc. of the Int. Conf. on Music Information Retrieval, ISMIR 2003*, Washington DC, USA, 2003.

[12] R. O. Duda, R. E. Hart, D. G. Stork, *Pattern classification*, John Wiley & Sons, 2001.