SHOT TYPE IDENTIFICATION OF MOVIE CONTENT

Ines Cherif, Vassilios Solachidis, Ioannis Pitas

Department of Informatics, Aristotle University of Thessaloniki
Thessaloniki 54124, Greece Tel, Fax: +30-2310996304
e-mail: pitas@aiia.csd.auth.gr

ABSTRACT

This paper aims at providing a quantitative description of shot types commonly used in movie productions. Only qualitative descriptions are available in the literature and even these are subject to various naming conventions. A vocabulary is fixed and human body-based rules are defined to extract the shot types. A database was generated with a set of samples labeled by cinematography experts. The proposed approach was tested on the set of samples providing promising results.

I. INTRODUCTION

Due to the increasing amount of digital video data and the intense need to access rapidly the useful information, many methods were proposed for video content description. One fundamental unit of the video stream is the video shot. It represents a set of frames acquired with a unique camera without interruption. The shot can be described in terms of various shot grammar elements, one of them being the apparent camera-to-subject distance. This distance is related to the field of view with respect to a character identified as the primary subject. Extracting such information about the shot composition is very useful, since specific shot types are selected to tell the stories, convey emotions or express a point of view. As will be shown throughout this paper, the shot type extraction is a rather difficult task. Only qualitative descriptions are available and different terminologies can be found in the literature on movie production. Moreover, the description can be limited to 3 broad categories (close up, medium shot and long shot) or extended to 9 more specific ones.

While a large amount of work was done on shot genre extraction [1], [2] most of the approaches dealing with shot type detection were directed towards sport applications. This is understandable, since, for example, prior information about the field geometry for example could be exploited [3]-[5]. For instance, soccer video shots were classified into four categories: long shot, in-field medium, close-up and out of field shot[3], giving two low-level features: the grass pixel ratio and a 3:5:3 section-decomposition, as input to a Bayesian classifier. Similarly, a classification was done based on the field ratio, the head area, the texture and the object scale feature [4]. Still in sport, a method based on the camera motion estimation was proposed to differentiate between wide shots and close up shots in basketball videos [6].

The present paper is structured as follows. First the terminology used for shot type analysis is given and a qualitative description of each type is provided. Then thresholds are proposed to distinguish between the different shot types and a process for primary subject selection is presented. Section IV describes the experiments carried out and the results obtained. Finally, conclusions are drawn in section V.

II. QUALITATIVE DESCRIPTION OF THE SHOT TYPES

One way of describing the video content is by classifying the video shots according to how far the camera seems to be from a primary subject, being an object or an actor. Since the definition is not based on the physical distance but rather on a personal perception of the space, the description of the shot type becomes a subjective concept.

A common qualitative description of shot types is based on the human body information. Therefore subjects are limited to actors [7].

Moreover the number of shot type variations can differ from one description to another one. Three basic shot types are widely used: close-up, medium shot and long shot. However, some definitions can be more detailed by introducing intermediate types, up to 9 variations.

In this study, we will retain 7 of those variations, defined as follows [8]. In the *eXtreme Long Shot* (XLS), the majority of the frame is taken up by the scene. Actors can/cannot be seen in the field of view. If there are ones, then they are small and unrecognizable. The *Long Shot* (LS) shows the whole body of an actor to appear with some of the surroundings. The subject is now recognizable. The *Medium Long Shot* (MLS) is a shot where the subject body is not viewed entirely in the video frame, and the human body is framed from the knees up. In the *Medium*

Shot (MS), the upper part of the body is visible until the waist and hand gestures can be seen. The *Medium Close Up* (MCU) shows the head and shoulders of an actor. In a *Close Up* (CU), the head is shown in great detail and fills the screen. A part of the neck can be visible as well. This kind of shot conveys a high degree of intimacy to the characters. For an *eXtreme Close Up* (XCU), less than the complete face is visible. The XCU is usually employed to convey intense emotions in love or agressive scenes.

Since the definitions adopted here are based on the human body information, we will consider only shots where actors are present.

III. SHOT TYPE DETECTION

III-A. The golden ratio in the human body

The golden ratio, also called golden section or golden proportion, is represented by ϕ and equals $\frac{1+\sqrt{5}}{2} \simeq 1.618$. It has been widely used in art and architecture and is said to play a role in human perception of beauty, defining a natural balance between symmetry and asymmetry. This ratio can also be found in the human body (head, arms, legs,...) [9]. ϕ links the height of the head to the height of other body parts (see Fig. 1). Therefore, it will be used as basis for our quantitative description. Even though it appears in numerous natural proportions, this ratio is of course not a precise measure of the body structure. Nevertheless, it gives as we will see in the forthcoming sections, a good approximation of the body geometry. It can also be mentioned that, for our specific application, the approximation of the proportions by ϕ is convenient, since the boundaries between successive shot types are rather fuzzy.

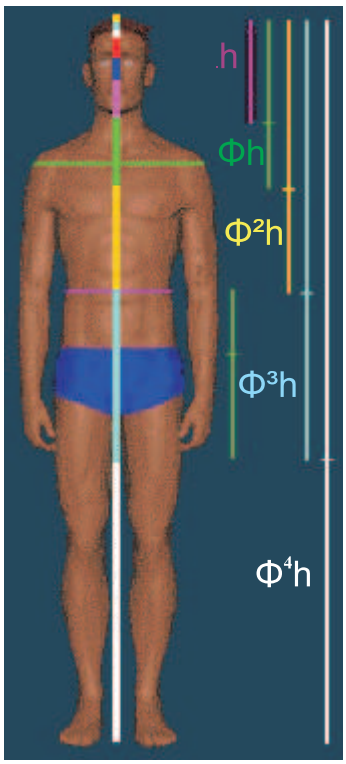

Fig. 1. The golden ratio in the human body

III-B. Quantitative description of the shot types

Starting from the qualitative description given in section II, mathematical expressions based on information about the character head location and size are proposed to allow an automatic extraction of the shot types. For the head size evaluation, the height h will be used. In fact, this feature was proven to be more robust than either the width or the area, because it is invariant to face pan rotation, which occurs more frequently than head tilts. The second feature used is the distance d from the bottom of the head bounding box to the bottom of the frame. This feature is very relevant since it indicates the frame space that can be used to frame the subject's body (see Fig. 2).
After multiple visualizations of video sequences and a study of the filming techniques[8],[7], several observations, relative to the subject size and position, on the screen, were made and led to a set of rules listed below:

1) The XLS shot can contain faces that are not recognizable, thus h should not exceed $0.05H$, where H is the height of the video frame.
2) The LS will be composed by faces bigger than $0.05H$, since the subjects have to be recognizable. There might also be enough space under the face to frame the whole body, hence the value of $d+h$ should be bigger than $\phi^4 h$ which approximates the height of the whole human body (see Fig. 1).
3) The subjects belonging to an MLS, MS and MCU shot should have a face height smaller than $0.65H$. Further more, the value of $d+h$ should be in the range $[\phi^3 h, \phi^4 h]$, $[\phi^2 h, \phi^3 h]$ and $[\phi h, \phi^2 h]$ for MLS, MS and MCU respectively. We also noticed, on one hand, that if h is bigger than $0.65H$ and smaller than $0.9H$, while $d+h$ is bigger than ϕh, we are still in the case of an MCU shot.
4) At the same height range, the shot becomes a CU one when $d+h < \phi h$.
5) when $h > 0.9H$, the shot is an XCU one, whatever the position of the face on the screen is. All the rules previously described are summarized in a logic diagram (see Fig. 3).

III-C. Determining the primary subject

Generally, a video frame contains numerous actors with various postures possible at different distances from the camera. The assignment of a single type to a video frame is based on a character considered as the most significant

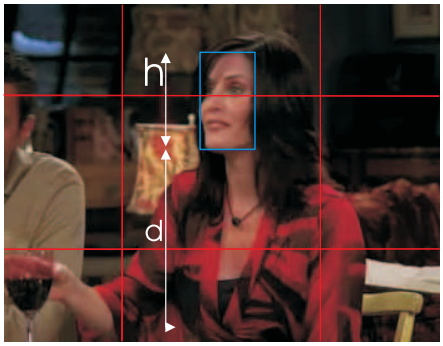

Fig. 2. Representation of the two low-level feature d and h and the Golden section.

one in the video frame. This significance can be translated into observer attention (focus of attention) to this subject, due to its location, motion or illumination. Obviously, the head and body locations are the main criteria, for defining the focus of attention.

To identify the center of focus in the video frame, saliency maps were proposed [10]. This approach, even though attractive, is rather heavy computationally speaking. Thus, we based our selection on the classic rule of the Golden Section dividing the video frame into 3:5:3 parts along both dimensions [3], [4]. This rule suggests that the center of the camera's attention is the middle frame section (see Fig. 2).

The character, located in this particular area of the frame and whose head height has the highest value, will represent the primary subject. If no character is screened in this area, then the one with the dominant head height in the whole frame will be selected as a reference for the forthcoming shot classification.

III-D. From the frame level to the shot level

The classification is performed on a frame-basis. For each frame, the most-significant character is selected and according to its head height and position, a label is assigned to the video frame. However, the shot types presented in section II are defined for the entire shot, and not for each individual frame. Therefore, a decision should be taken at a higher level. The weighted majority voting of the shot type status over the frames in a video shot is a straightforward process to get the shot type for the entire shot[3],[5]. Other similar approaches can be used as well.

IV. EXPERIMENTAL PART

Since the shot type detection is based on the head size and position, the process required the detection of faces

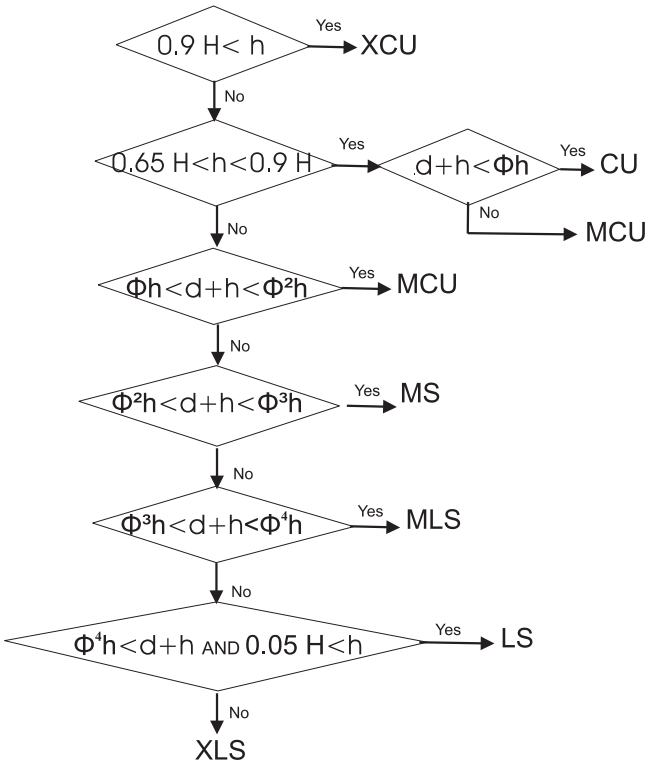

Fig. 3. Set of rules for shot type definition

in the video frames. This task was performed using a semi-automatic face tracker [11]. The output face bounding boxes were adjusted manually when needed.

IV-A. Ground truth data

A video shot database was created. 66 shots were extracted and manually labeled as: XCU, CU, MCU, MS, MLS, LS or XLS. For each shot, the faces were detected. For each face, the information about its height and position were stored. Therefore, the database contains 4606 detected faces with several values of d and h (see Fig. 4). It can be noticed that when the value of h increases, the shots go towards XCU and CU shots, while when h decreases and d increases the shots are more likely to be LS or XLS.

IV-B. Results

The approach described above was applied on the database samples. The confusion matrices of the classification are provided in Table I and Table II. The first one shows results at a frame level while the second one shows results at a shot level. These results point out that the maximum values are obtained on the diagonal of the matrix, therefore maximizing the good classification rate. We can also notice that in case of misclassification the label assigned to the shot or frame is close to the ground

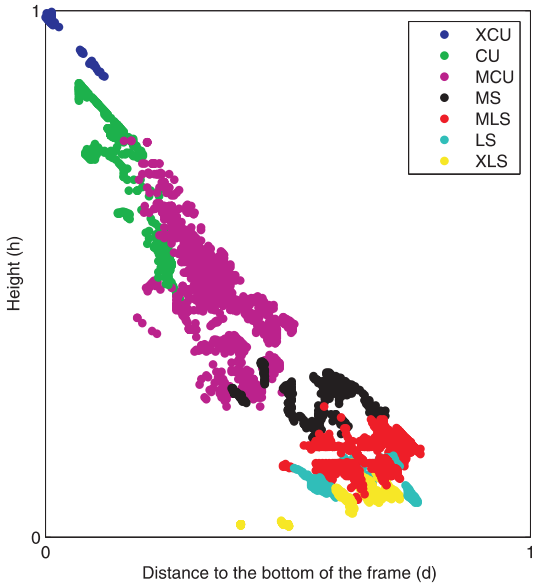

Fig. 4. Representation of the ground truth samples in the feature space.

Table I. Frame-based confusion matrix.

	XCU	CU	MCU	MS	MLS	LS	XLS
XCU	50	9	0	0	0	0	0
CU	0	349	160	0	0	0	0
MCU	0	84	1952	11	0	0	0
MS	0	0	48	307	3	0	0
MLS	0	0	0	86	992	12	0
LS	0	0	0	0	93	248	0
XLS	0	0	0	0	28	141	33

truth label, i.e non-zero entries occur close to the main diagonal of the respective tables.

A classification accuracy metric A is defined to evaluate the algorithm performance over the whole set of shots. It is formulated as follows:

$$A = \frac{N_{CC}}{N_{GT}} \quad (1)$$

where N_{CC} is the number of shots or frames correctly classified. N_{GT} is the number of ground truth shots or frames, respectively. The frame-based accuracy obtained is 85.35% while the shot-based accuracy reaches 90.91%. Support Vector Machine algorithms were also employed to solve the multi-class shot type classification problem, but the accuracy obtained was lower using these data sets.

V. CONCLUSIONS

In this paper, we presented a quantitative description of the 7 shot types most often encountered in movies. The approach used is based on two main features, namely head height and head distance to the bottom of the frame. When tested on a set of shots the algorithm provided very good results. This approach uses human head-based features. Therefore, no decision can be taken when no actors are screened on the frame.

Table II. Shot-based confusion matrix.

	XCU	CU	MCU	MS	MLS	LS	XLS
XCU	2	0	0	0	0	0	0
CU	0	5	2	0	0	0	0
MCU	0	0	28	0	0	0	0
MS	0	0	1	5	0	0	0
MLS	0	0	0	0	17	0	0
LS	0	0	0	0	1	2	0
XLS	0	0	0	0	0	2	1

ACKNOWLEDGEMENT

The work presented was developed within NM2 (New Media for a New Millenium), a European Integrated Project (http://www.ist-nm2.org), funded under the European Commission IST FP6 programme.

VI. REFERENCES

[1] L. Chaisorn and T. Chua, "The segmentation and classification of story boundaries in news video," in *Proceedings of the IFIP TC2/WG2.6 Sixth Working Conference on Visual Database Systems*. Deventer, The Netherlands, The Netherlands: Kluwer, B.V., 2002, pp. 95–109.

[2] A. Ferman and A. Tekalp, "A fuzzy framework for unsupervised video content characterization and shot classification," *Journal of Electronic Imaging*, vol. 10, no. 4, pp. 917–929, October 2001.

[3] A. Ekin, A. Tekalp, and R. Mehrotra, "Automatic soccer video analysis and summarization," *IEEE Transactions on Image Processing*, vol. 12, pp. 796–807, July 2003.

[4] H.-Q. L. H.-L. J. X.-F. Tong, Q.-S. Liu, "Shot classification in sports video," in *Proceedings of the 7th International Conference on Signal Processing(ICSP 2004)*, vol. 2, Beijing, China, 31 Aug.-4 Sept 2004, pp. 1364–1367.

[5] C. X. J. Wang, E. Chng, "Soccer replay detection using scene transition structure analysis," in *Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing(ICASSP 2005)*, vol. 2, Pennsylvania, USA, 18-23 March 2005, pp. 433–436.

[6] S. K. Y.-P. Tan, D.D. Saur and P. Ramadge, "Rapid estimation of camera motion from compressed video with application to video annotation," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 10, no. 1, pp. 133–146, 2000.

[7] D.Ablan, *Digital cinematography & directing*, 1st ed. New Riders Press, December 2002.

[8] R.Thompson, *Grammar of the Shot*. Focal Press, June 1998.

[9] "The golden ratio in the human body," http://goldennumber.net/body.htm.

[10] C. K. L. Itti and E. Niebur, "A model of saliency-based visual attention for rapid scene analysis," *IEEE Transactions on Pattern Analysis and Machine Intelligence (PAMI 1998)*, vol. 20, no. 11, pp. 1254–1259, 1998.

[11] N. N. G. Stamou and I. Pitas, "Object tracking based on morphological elastic graph matching," in *Proceedings of the IEEE International Conference on Image Processing (ICIP 2005)*, Genova, Italy, September 2005.