SVM-based Shot Type Classification of Movie Content

Ioannis Tsingalis, Nicholas Vretos, Nikos Nikolaidis, Ioannis Pitas

Department of Informatics Aristotle University of Thessaloniki, Thessaloniki, Greece

E-mail: pitas@aiia.csd.auth.gr

Abstract: **In this paper, we propose a Support Vector Machine (SVM) based shot classification method for movies. This method classifies shots into seven different classes, namely eXtreme Long Shot (XLS), Long Shot (LS), Medium Long Shot (MLS), Medium Shot (MS), Medium Close Up (MCU), Close Up (CU) and eXtreme Close Up (XCU). The proposed method uses two features. The first one is the ratio of the height of the actor's facial image to the height of video frame. The second one is the ratio of the corresponding widths. These two ratios constitute the 2-D feature vectors which are fed into the SVM. A ground truth labeled shot database was created in order to experimentally test the proposed method performance. The corresponding results are very promising.**

Keywords: Shot type classification, Feature extraction, Support Vector Machines

1 INTRODUCTION

Due to the flourish of the movie industry during the last decades the automatic analysis, description, indexing and retrieval of video content became an urgent necessity. In these applications, shot type classification is undoubtedly one of the most useful techniques for analysing, characterizing and subsequently retrieving video. A shot is a continuous filming from a single camera for a certain period of time [5]. Shots constitute the main building blocks of film editing process. Some of the shot types used in cinematography can be found in [6], [3] and [5]. As far as research in shot type classification is concerned, a vast amount of work has been done, mostly in sports and news videos. For example, S.F.Chang et. al. [13] classify shots into CU, LS and MS using grass-ratio, and based on this classification they categorize sports video in play or break segments using a heuristic rule. Other shot classification methods for sports content that use a ratio on the apparent grass of the football field, the so called grass-ratio, can be found in [6] and [4]. Another interesting approach for shot type analysis in tennis videos is presented by X. Yu. et. al. [14] and is based on MPEG motion vectors and other features. Moreover, in [9] and [11] histogram color information is used for shot type classification. Unlike sports whose shot types are limited and more restricted, movie shots are more diverse over the various film types and genres. In [1] and [12], information from saliency maps, geometric composition of the scene, color and motion distribution are considered in order to classify shots. Other alternative approaches for determining the shot types work by estimating either the *absolute* or *relative* depth of the scene. The former technique is based on the actual distance between the filmed object and the camera, whereas the latter is based on the estimation of texture gradients [8], shape from shading, fractal dimensions [7] and other features.

The method in [3] is based on a combination of the height of the bounding box that frames the actor's face and the distance between the bottom of the bounding box and the bottom of the frame. However, this algorithm does not perform very well since it is based on a simple thresholding of the derived value that involves the human body golden ratio. Figure 1 shows the basic features of the approach described in [3]. These two features are capable to represent the dominance of the actor on the frame and thus provide the shot type classification.

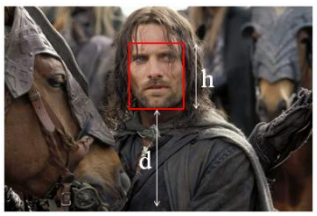

Figure 1: Low level features used in [3]

In this paper, a shot classification method for movies is presented. The major assumption in that method is that an actor is present in each shot and the bounding box of the actors face is available through the application of a face detection and/or tracking algorithm. The dimensions of the bounding boxes are used for the extraction of the features that are in the proposed classification method. The proposed method is not restricted to specific movie genre, like in [1] that can operate only in action movies, but covers different genres. The rest of the paper is organized as follows: in Section 2 the main shot types are introduced. In Section 3 the proposed method is presented. Experimental results are drawn in Section 4. Section 5, concludes the paper.

2 MOVIE SHOT TYPES

There are seven major types of video shots in movies, also known as field sizes [5].

eXtreme Close Up Shot (XCU): A part of the actor's face is visible. In this type the frame contains no information for the background.

Close Up Shot (CU): In this category the head of the actor is visible from the top of the actor's hair till the top of the shoulder. Sometimes it is called a "head shot".

Medium Close Up Shot (MCU): The human body is framed from the elbow joint and above.

Medium Shot (MS): In this type the human is visible down to the waist level and hand gestures are visible.

Medium Long Shot (MLS): In this shot type, the actor's body is usually framed from the knees and up.

Long Shot/Wide Shot (LS/WS): In this case almost the entire body of the actor is visible and is usually considered as a full "body shot".

eXtreme Long Shot (XLS): It is usually used as an establishing shot and unlike previous types that can occur either in indoor or outdoor scenes, it usually occur in outdoor scenes. In this case the background is the most dominant element in the frame.

It should be noted that the previous shot types are not equiprobably in movies. For example, eXtreme Close Up and eXtreme Long Shot rarely appear. On the other hand, shots such as Close Up and Medium Close Up are used more frequently. Figure 2, shows the type of shots.

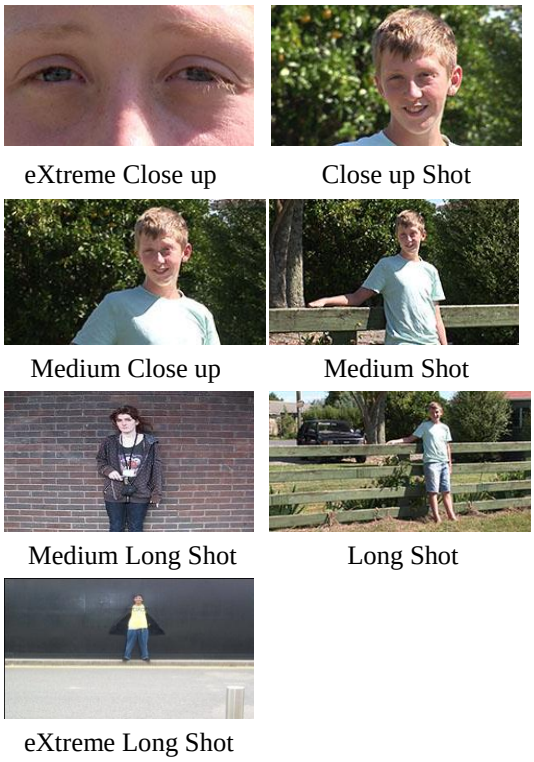

Figure 2: Type of Shots

3 SVM-BASED SHOT TYPE CLASSIFICATION

In the proposed method the height and the width of the facial bounding box are used, combined with the height and the width of the video frame. More specifically the extracted features are:

- Let the height of the face bounding box be h_{bb} and the height of the video frame H_F. The first feature evolved in the proposed method is the ratio h_{bb}/H_F.
- The second feature is $w_{bb}/\ W_F$. Where w_{bb} and W_F are the width of the bounding box and the video frame respectively.

Figure 3 shows the abovementioned features. These two features are the elements of the feature vectors. Figure 4, illustrates the value of the features for the ground truth data.

The feature vector, $\left(\frac{h_{bb}}{H_F}, \frac{w_{bb}}{W_F}\right)$ of each frame is fed to an appropriate trained Support Vector Machine. SVM consists of the following optimization problem: let $(\mathbf{x}_i, \mathbf{y}_i)$, where $\mathbf{x_i} \in R^n$ and $\mathbf{y}_i \in \{0,1,\dots,m\}$ is the sample label index. The optimization problem is formulated as:

$$\min_{w,b,\xi} \frac{1}{2}\mathbf{w}^T\mathbf{w} + \mathrm{C}\sum_{i=1}^{n}\xi_i \quad (1)$$

Subject to: $\mathrm{y}_i(\mathbf{w^T}\phi(\mathbf{x_i}) + \mathrm{b}) \geq 1 - \xi_i,\ i = 1,\dots,N \geq 0$ (2)

$$\xi_i \geq 0,\ i = 1,\dots,\mathrm{N} \quad (3)$$

where C is the tuning parameter used to balance the margin and training error and ξ_i are called the slack variables that are related to the soft margin. In this work in order to find the best parameterization of the SVM "grid-search" along with cross validation is applied. The applied SVM uses the RBF kernel. Thus, we need to define in the same time the parameter C of the optimization problem, as well as the kernel parameter. In our case we use the RBF kernel, i.e. $K(x_i, x_j) = \exp\left(-\gamma\|x_i - x_j\|^2\right)$, $\gamma > 0$. In order to apply the SVM-based experiments the library libSVM [2] is used.

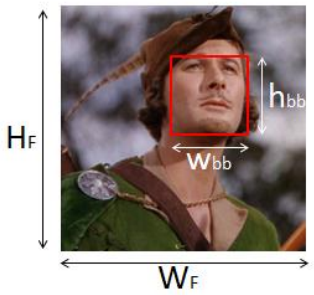

Figure 3: Example of face tracked by the tracking algorithm in [18]

4 EXPERIMENTAL EVALUATION

In this section experimental results are provided. Two types of results are presented. The frame based results refer to the classification of each frame of the shot in one of the shot types, whereas the shot-based results refer to the classification of the entire shot based on the classification of the individual frames

of the shot. In order to decide on the type of an entire shot majority voting on the derived labels of the frames contained in the shot is applied similar to [3], [6] and [10]. In other worlds, the majority of the labeled frames, that compose a shot, characterize the type of the entire shot.

In order to calculate the classification accuracy at the frame/shot level we use the same metric as in [3]. That is,

$$A = \frac{N_{CC}}{N_{GT}}$$

where N_{CC} is the frames/shots correctly classified to a specific shot type and N_{GT} is the total number of the frames/shots

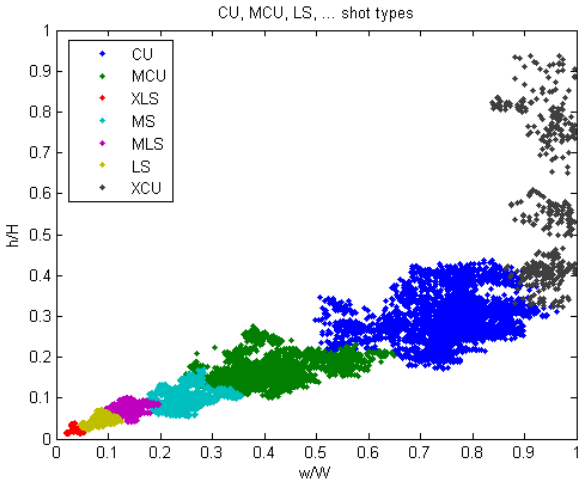

Figure 4: Feature values for frames from various shot types in the ground truth data

labeled in the ground truth data as belonging to this shot type. In this Section confusion matrixes are also evaluated.

Our dataset is composed of 173 shots and 12178 frames, compiled from different movie genres by different directors. In each frame an actor is present. Most of the actors are looking towards the camera. Adhering to the definitions in Section 2, we have labeled each shot manually to construct the ground truth data (shot labels). The database includes only shots where only one actor is depicted so as to simplify it's construction. However, the proposed method can easily generalized to work on shots depicting more than one actor, by selecting the most "dominant" one. This can be done for example by using a rule similar to the one proposed in [3] which decides on the dominant object from all depicted objects.

For the extraction of facial bounding boxes, we manually mark the face bounding box in the first frame of each shot and then we apply the tracking algorithm described in [15] in order to track the face in the remaining frames. A common problem in tracking is the occlusion of the tracked object, in our case the face, that might lead to loss of target. To cope with this problem one can reinitialize the tracked region or periodically use an effective face detector. In our database, for reasons of simplicity, we exclude shots where occlusions occur since solving tracking problems is beyond the scope of this paper.

The method in [3] was applied to the database described above in order to obtain frame-based and shot-based experimental results. Frame-based results are shown in the confusion matrix in Table 1.

Table 1: Frame-based results according to algorithm in [3]

	XCU	CU	MCU	MS	MLS	LS	XLS
XCU	0.92	0.08	0	0	0	0	0
CU	0.01	0.86	0.13	0	0	0	0
MCU	0	0.08	0.92	0	0	0	0
MS	0	0	0	0.9	0.1	0	0
MLS	0	0	0	0	0.88	0.12	0
LS	0	0	0	0	0.007	0.96	0.033
XLS	0	0	0	0	0	0.12	0.88

Based on this metric the frame-based accuracy is 90,3%. The shot-based overall classification accuracy for this method is 91,4% and the corresponding confusion matrix is presented in Table 2.

Table 2: Shot-based results according to algorithm in [3]

	XCU	CU	MCU	MS	MLS	LS	XLS
XCU	1	0	0	0	0	0	0
CU	0	0.78	0.22	0	0	0	0
MCU	0	0	0.94	0.06	0	0	0
MS	0	0	0	1	0	0	0
MLS	0	0	0	0	0.85	0.15	0
LS	0	0	0	0	0	1	0
XLS	0	0	0	0	0	0.17	0.83

When the proposed method was applied on the same dataset, the frame-based overall accuracy was 98,3% and the corresponding confusion matrix is presented in Table 3, whereas the shot based accuracy was 100%.

Table 3: Frame-based classification results of the proposed method

	XCU	CU	MCU	MS	MLS	LS	XLS
XCU	0.97	0.03	0	0	0	0	0
CU	0.004	0.99	0.006	0	0	0	0
MCU	0	0.003	0.98	0.017	0	0	0
MS	0	0	0.019	0.976	0.005	0	0
MLS	0	0	0	0.023	0.97	0.007	0
LS	0	0	0	0	0.004	0.996	0
XLS	0	0	0	0	0	0	1

Finally, in order to implement the experimental results 5-fold cross validation method was applied in the sample dataset described.

5 CONCLUSION

Movie shot type classification is a challenging problem. In this paper, we propose a method based on the height and width of the facial image in combination with the corresponding height and width of the video frame. The ratios are fed in a properly trained SVM classifier obtaining 100% accuracy at the shot type level. The main drawback of our method is that it is based on facial images, and thus, it is capable of shot type classification only in video that contain faces.

Acknowledgment

The research leading to these results has received funding from the European Union Seventh Framework Programme (FP7/2007-2013) under grant agreement n 287674 (3DTVS). The publication reflects only the authors' views. The EU is not liable for any use that may be made of the information contained therein.

References

[1] S. Benini, L. Canini, and R. Leonardi. Estimating cinematographic scene depth in movie shots. In Proceedings of the IEEE International Conference on Multimedia and Expo (ICME), pages 855 –860, July 2010.

[2] C. C. Chang and C. J. Lin. LIBSVM: A library for support vector machines. ACM Transactions on Intelligent Systems and Technology, 2:27:1–27:27, 2011. Software available at http://www.csie.ntu. edu.tw/_cjlin/libsvm.

[3] I. Cherif, V. Solachidis, and I. Pitas. Shot type identification of movie content. In Proceedings of Signal Processing and Its Applications (ISSPA), 2007.

[4] S.Chen, M.Shyu, C.Zhang, L.Luo, and M.Chen. Detection of soccer goal shots using joint multimedia features and classification rules. In Proceedings of the Fourth International Workshop on Multimedia Data Mining (MDM/KDD), pages 36–44, 2003.

[5] D.Arijon. Grammar of the Film Language. Silman-James Press, 1991

[6] A. Ekin, A.M. Tekalp, and R. Mehrotra. Automatic soccer video analysis and summarization. In the IEEE Transactions on Image Processing, vol.12, page 796 – 807, July 2003.

[7] J. M. Keller, R. M. Crownover, and R. Y. Chen. Characteristics of natural scenes related to the fractal dimension. In the IEEE Transactions on Pattern Analysis and Machine Intelligence, vol.9, no,5, page 621-627, September 1987.

[8] B. J. Super and A. C. Bovik. Shape from texture using local spectral moments. In the IEEE Transactions on Pattern Analysis Machine Intelligence, vol. 17, no.4, page 333–343, April 1995.

[9] T. Xiaofeng, L .Qingshan, and L. Hanqing. Shot classification in broadcast soccer video. Electronic Letters on Computer Vision and Image Analysis, vol.7, no.1, 2008.

[10] C. J. W. Engsiong, and X. Changsheng. Soccer replay detection using scene transition structure analysis. In the Proceedings of IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP), vol.2, pages 433 – 436, March 2005.

[11] L. Wang, M. Lew, and G. Xu. Offense based temporal segmentation for event detection in soccer video. In Proceedings of the 6th ACM SIGMM International Workshop on Multimedia information retrieval, MIR, pages 259–266, New York, USA, 2004.

[12] M. Xu, J. Wang, M. A. Hasan, X. He, C.Xu, H. Lu, and J.S. Jin. Using context saliency for movie shot classification. In the Proceedings of 18th IEEE International Conference on Image Processing (ICIP), pages 3653-3656, September 2011.

[13] P. Xu, L. Xie, S.F. Chang, A. Divakaran, A. Vetro, and H. Sun. Algorithms and system for segmentation and structure analysis in soccer video, In the Proceedings of IEEE International Conference on Multimedia and Expo (ICME), pages 721 - 724, August 2001.

[14] X. Yu, L. Duan and Q. Tian. Shot classification of sports video based on features in motion vector field. In the Proceedings of 3rd IEEE Pacific-Rim Conference on Multimedia Proceedings, pages 235-260, December 2002.

[15] S. Zhou, R. Chellappa, and B. Moghaddam. Visual tracking and recognition using appearance-adaptive models in particle filters. In the IEEE Transactions on Image Processing, vol.13, page 1434–1456, 2004.