VISILAB at MediaEval 2013: Fight Detection

Ismael Serrano, Oscar Déniz, Gloria Bueno
VISILAB group, University of Castilla-La Mancha
E.T.S.I.Industriales, Avda. Camilo José Cela 3,
13071 Ciudad Real, Spain.
Ismael.Serrano@uclm.es, Oscar.Deniz@uclm.es, Gloria.Bueno@uclm.es

ABSTRACT

Fight detection from video is a task with direct application in surveillance scenarios like prison cells, yards, mental institutions, etc. Work in this task is growing, although only a few datasets for fight detection currently exist. The Violent Scene Detection task of the MediaEval initiative offers a practical challenge for detecting violent video clips in movies. In this working notes paper we will briefly describe our method for fight detection. This method has been used to detect fights within the above-mentioned violent scene detection task. Inspired by results that suggest that kinematic features alone are discriminative for at least some actions, our method uses extreme acceleration patterns as the main feature. These extreme accelerations are efficiently estimated by applying the Radon transform to the power spectrum of consecutive frames.

Keywords

Action recognition, violence detection, fight detection

1. INTRODUCTION

In the last years, the problem of human action recognition from video has become tractable by using computer vision techniques [5]. Despite its potential usefulness, the specific task of violent scene detection has been comparatively less studied. The annual MediaEval evaluation campaign introduced this specific problem in 2011. For an overview of this year‘s task please see [4].

2. PROPOSED METHOD

The presence of large accelerations is key in the task of fight detection ([6], [2] and [1]). In this context, body part tracking can be considered, as in [3], which introduced the so-called Acceleration Measure Vectors (AMV). In general, acceleration can be inferred from tracked point trajectories. However, extreme acceleration implies image blur (see for example Figure 1), which makes tracking less precise or even impossible.

Motion blur entails a shift in image content towards low frequencies. Such behaviour allows building an efficient acceleration estimator for video. Our proposed method works with sequences of 50 frames; therefore we need to divide the shots in 50-frame clips. First, we compute the power spectrum of two consecutive frames. It can be shown that, when there is a sudden motion between two consecutive frames, the power spectrum image of will depict an ellipse. The orientation of the ellipse is perpendicular to the motion direction, the frequencies outside the ellipse being attenuated. Most importantly, the eccentricity of this ellipse is dependent on the acceleration. Basically, the proposed method aims at detecting the sudden presence of such ellipse.

Our objective is then to detect such ellipse and estimate its eccentricity, which represents the magnitude of the acceleration. Ellipse detection can be reliably performed using the Radon transform, which provides image projections along

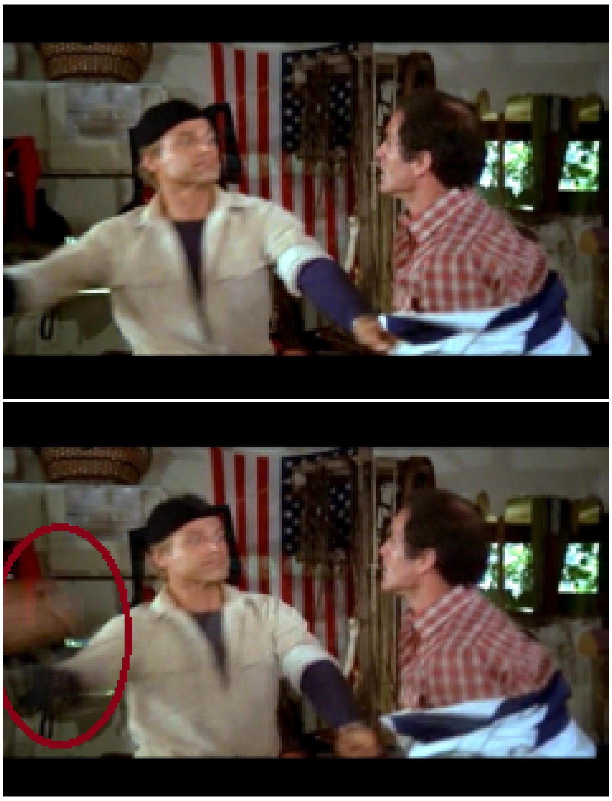

Figure 1: Two consecutive frames in a fight clip from a movie. Note the blur on the left side of the second frame.

Copyright is held by the author/owner(s).
MediaEval 2013 Workshop, October 18-19, 2013, Barcelona, Spain

Table 1: AED precision, recall and F-measure at video shot level

Run	AED-P	AED-R	AED-F
Run1-classifier-knn	0.1178	0.6265	0.1982
Run2-classifier-svm	0.1440	0.4482	0.1700

Table 2: Mean Average Precision (MAP) values at 20 and 100

Run	MAP at 20	MAP at 100
Run1-classifier-knn	0.1475	0.1343
Run2-classifier-svm	0.1350	0.1498

lines with different orientations.

For each pair of consecutive frames, we compute the power spectrum using the 2D Fast Fourier Transform (in order to avoid edge effects, a Hanning window is applied before computing the FFT). Let us call these spectra images P_{i-1} and P_i. These images are divided, i.e. $C = P_i/P_{i-1}$.

When there is no change between the two frames, the power spectra will be equal and C will have a constant value. When motion has occurred, an ellipse will appear in C. Ellipse detection can be reliably performed using the Radon transform.

After applying the Radon transform to image C, its vertical maximum projection vector is obtained and normalized (to maximum value 1). Next when there is an ellipse, this vector will show a sharp peak, representing the major axis of the ellipse. The kurtosis of this vector is therefore taken as an estimation of the acceleration.

Note that kurtosis alone cannot be used as a measure, since it is obtained from a normalized vector (i.e. it is dimensionless). Thus, the average value per pixel P of image C is also computed, taken as an additional feature. Without it, any two frames could lead to high kurtosis even without significant motion.

Deceleration was also considered as an additional feature, and it can be obtained by reversing the order to consecutive frames and applying the same algorithm explained above. For every short clip, we compute histograms of these features, so that acceleration/deceleration patterns can be used for discrimination.

Once we have features for every clip, for classification we use two different classifiers: the well-known K-Nearest Neighbours and Support Vector Machine (SVM) with a lineal kernel.

The method described requires training clips that contain fights. Thus, when fight sequences are given, they may have to be first evaluated for fight subsequences.

3. EXPERIMENT

Sometimes, within a violent segment non-fight clips may appear. For each violent segment (as provided by MediaEval organizers), we manually removed clips without fighting action. Moreover, random clips of 50 consecutive frames, taken outside the violent segments, were selected for the non-fight class. We trained the two classifiers for the fight concept. Then we apply the two trained classifiers to every 50 consecutive frames of the test set.

We submitted 2 runs and the details of performances are as follows. Table 1 reports AED [4] precision, AED recall and AED F-measure values, whereas Table 2 shows the evaluation results for the submitted runs, MAP at 20 and 100.

4. CONCLUSIONS

Based on the observation that kinematic information may suffice for human perception of various actions, in this work a novel fight detection method is proposed which uses extreme acceleration patterns as the main discriminating feature.

In experiments with other datasets we obtained accuracies above 90%, and processing times of a few milliseconds per frame. The results on the MediaEval dataset are, however, very poor. We suppose it could be due to the test ground truth (used by the organizers to obtain the performance measures). The category 'violence' is more general because includes violent scenes that could include explosions, shots, car chases, fights, etc. Although we had 'fight' labelled training videos, that label was not available for test videos. What is more, the definition of 'violence' in MediaEval is "physical violence or accident resulting in human injury or pain", so we were able to detect only part of the violence: fights.

On the other hand, there are a number of practical aspects that are not taken into account. In many surveillance scenarios, for example, we do not have access to color images and audio, and the typical forms of violence are fights and vandalism, instead of explosions, car chases, etc. The processing power needed for running detection algorithms is also an important issue in those applications.

5. ACKNOWLEDGMENTS

This work has been supported by Project TIN2011-24367 from Spain's Ministerio de Economía y Competitividad.

The authors also thank the MediaEval organizers for inviting us to participate.

6. REFERENCES

[1] G. Castellano, S. Villalba, and A. Camurri. Recognising human emotions from body movement and gesture dynamics. *Affective Computing and Intelligent Interaction*, 4738(1):17–82, 2007.

[2] T. J. Clarke, M. F. Bradshaw, D. T. Field, S. E. Hampson, and D. Rose. The perception of emotion from body movement in point-light displays of interpersonal dialogue. *Perception*, 34(1):1171–1180, 2005.

[3] A. Datta, M. Shah, and N. D. V. Lobo. Person-on-person violence detection in video data. *Pattern Recognition. Proceedings. 16th International Conference*, 1(1):433–438, 2002.

[4] C. H. Demarty, C. Penet, M. Schedl, B. Ionescu, V. Quang, and Y. G. Jiang. The 2013 Affect Task: Violent Scenes Detection. In *MediaEval 2013 Workshop*, Barcelona, Spain, October 18-19 2013.

[5] R. Poppe. A survey on vision-based human action recognition. *Image and Vision Computing*, 28(6):976–990, 2010.

[6] M. Saerbeck and C. Bartneck. Perception of affect elicited by robot motion. *In Proceedings of the 5th ACM/IEEE international conference on Human-robot interaction*, 10(1):53–60, 2010.