MTM at MediaEval 2013 Violent Scenes Detection: through acoustic-visual transform

Bruno do Nascimento Teixeira
Universidade Fedederal de Minas Gerais
Belo Horizonte, Brazil
bruno.texeira@dcc.ufmg.br

ABSTRACT

This paper describes the team MTM participation in the MediaEval 2013 campaign. We submitted one run at shot level that explores spatial correlation between acoustic-visual features. The motion features are computed to represent the video.The Mel Frequency Cepstral Coefficients (MFCC) of the acoustic signal, and their first and second order derivatives are exploited to represent audio. One main issue in designing movie shot classification is considered. This issue is "there is a correlation between velocity and acceleration and the acoustic features". Our approach relies in find canonical bases, using Canonical Correlation Analysis (CCA), in order to represent video. We also add spatial information using frame regions. We evaluate the performance of our proposed method on MediaEval 2013 Violent Scenes Detection in film data.

Keywords

MediaEval 2013, violent scenes detection, canonical correlation analysis, Bayesian networks

1. INTRODUCTION

Two classes of violence can be considered: objective and subjective violence. Objective violence is defined as "physical violence or accident resulting in human injury or pain". More details about the violence detection task can be found in [1].
We focus on objective violence and assume there is non-trivial correlation between acoustic features and motion in objective violence scenes. In this case, we explore the correlation between acoustic and visual features. Canonical Correlation Analysis (CCA) proposed by [3] maximizes the correlation between two multivariate random vectors by finding a linear transforms w_x and w_y. CCA is employed for identification and segmentation of moving-sounding objects [4].

2. METHOD

The goal of the proposed work is to combine visual and acoustic features by computing the canonical base vectors. A grid (see Figure 1 (a)) is used to segment the frame and capture the spatial information and build an acoustic transform map.

2.1 Video Representation

For each grid segment, S_x is computed using optical flow, where $x^j = (x_1^j, x_2^j)$ and x_1^j, x_2^j average velocity and acceleration magnitude respectively for all pixels belonging to the same region. For each audio frame, 12 MFCCs and their first and second derivates are computed to build an acoustic vector $y^j = (y_1^j, y_2^j, ..., y_{36}^j)$.

2.2 Canonical Correlation Analysis

Consider a multivariate random vector of the form (x, y) and a sample of instances $S = ((x^1, y^1), ..., (x^n, y^n))$ of (x, y), we can project x and y onto directions w_x and w_y ($x \to \langle w_x, x\rangle, y \to \langle w_y, y\rangle$) to maximizes the correlation ρ between $S_x w_x$ where $S_{xw_x} = (\langle w_x, x^1\rangle, ..., \langle w_x, x^n\rangle)$, and $S_y w_y$ where $S_{yw_y} = (\langle w_y, y^1\rangle, ..., \langle w_y, y^n\rangle)$:

$$\begin{aligned} \rho &= \max_{w_x, w_y} corr(S_x w_x, S_y w_y) \\ &= \max_{w_x, w_y} \frac{\langle S_x w_x, S_y w_y\rangle}{||S_x w_x|| ||S_y w_y||}. \end{aligned} \quad (1)$$

The coordinate system that optimizes the correlation between corresponding coordinates is found by solving the generalized eigenvectors $Ax = \lambda Bx$ [2, 3].

2.3 Early Fusion

Early fusion is performed by compute visual and acoustic representation T based on the canonical basis w_x for each region of the frame, using S_x and S_y to maximizes the correlation ρ:

$$T = [w_x^1 w_x^2 ... w_x^{25}], \quad (2)$$

where T the feature vector composed, w_x^r is visual linear transformation and r is the r-th region or grid position.

2.4 Bayesian Network

Bayes Net Toolbox for Matlab (BNT) [5] is used to train the network with 2 nodes: class node C (violence and non-violence) and the observed node T composed by $T = [w_x^1 w_x^2 ... w_x^{25}]$ (see Figure 1 (b)).

3. RESULTS

Table 1 shows the global result of our approach on the MediaEval 2013 affect test set. We obtain for each film of the test set the following precision values, which range from 0.009 to 0.216: Fantastic Four 0.216, Fargo 0.185, Forrest Gump 0.184, Legally Bond 0.009, Pulp Fiction 0.101, The God Father 0.070, The Pianist 0.124 (see Table 2). We plotted an FA (false alarm) curve (see Figure 2) for classification analysis.

4. CONCLUSIONS

We have developed a method based on canonical basis vectors to represent the video. Our method uses acoustic-visual transform

MediaEval 2013 Workshop, October 18-19, 2013, Barcelona, Spain

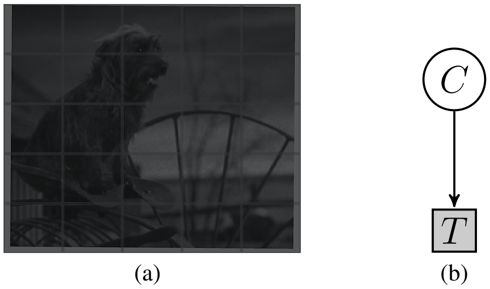

Figure 1: (a) A grid segments the frame in regions to compute velocity and acceleration. (b) Bayesian network with class node C (violence and non-violence) and the observed node $T = [w_x^1 w_x^2 ... w_x^{25}]$.

Table 1: Performance of MTM run (MediaEval 2013 Affect task) - Objective violence - Global Results

	Precision	Recall	F-measure	MAP-AT100
run#1	0.102	0.685	0.178	0.073

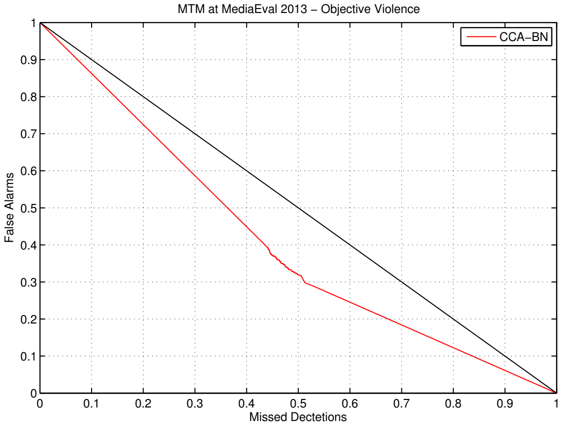

Figure 2: False alarms vs. missed detection for run# 1.

and spatial grid. It builds a transform map, used as representation, by using the assumption that motion features from violence scenes are correlated with acoustic features.

Analyzing each film result separately, high true positive and false alarm rates demonstrate that the transformation map alone can not distinguish all violence and non-violence shots and generalize the violence concept. It relies on variability of types of violence in movies and uncorrelated grid segments, which are not audio sources and must be discarded from the map. Event as explosions, screams and gunshots must present a visual acoustic pattern and are located in a specific frame region. In many scenes, few grid segments contribute with the audio dynamics.

Table 2: Performance of MTM run (MediaEval 2013 Affect task) - Objective violence

	Average Precision	Precision at 100
FantasticFour1	0.216	0.12
Fargo	0.174	0.15
ForrestGump	0.184	0.13
LegallyBlond	0.009	0.01
PulpFiction	0.101	0.02
TheGodFather	0.070	0.01
ThePianist	0.124	0.08

Possible directions for future work include region filtering to detect audio sources and remove noisy segments, spatial-temporal segmentation and feature selection.

5. ACKNOWLEDGMENTS

This work was supported in part by two grants from CAPES and CNPq.

6. REFERENCES

[1] C.-H. Demarty, C. Penet, M. Schedl, B. Ionescu, V. L. Quang, and Y.-G. Jiang. The mediaeval 2013 affect task: Violent scenes detection. In *MediaEval*, MediaEval Workshop, 2013.

[2] D. R. Hardoon, S. R. Szedmak, and J. R. Shawe-taylor. Canonical correlation analysis: An overview with application to learning methods. *Neural Comput.*, 16(12):2639–2664, Dec. 2004.

[3] H. Hotelling. Relations Between Two Sets of Variates. *Biometrika*, 28(3/4):321–377, 1936.

[4] H. Izadinia, I. Saleemi, and M. Shah. Multimodal analysis for identification and segmentation of moving-sounding objects. *Multimedia, IEEE Transactions on*, 15(2):378–390, 2013.

[5] K. Murphy. *Dynamic Bayesian Networks: Representation, Inference and Learning*. PhD thesis, UC Berkeley, Computer Science Division, July 2002.