Violence Detection in Video by Large Scale Multi-Scale Local Binary Patterns Dynamics

Martin V.[1,2],Glotin H.[1,2,3],Paris S.[2], Halkias S.[1,2],Prevot JM.[1]
{vincent.martin, glotin, halkias, jmp}@univ-tln.fr, sebastien.paris@lsis.org
(1) Université de Toulon, CNRS, LSIS, UMR 7296, 83957 La Garde, France
(2) Aix Marseille Université, CNRS, ENSAM, LSIS, UMR 7296, 13397 Marseille, France
(3) Institut Universitaire de France, 75005 Paris, France

ABSTRACT
We propose a violence detector based on the dynamics of new multi-scale local binary pattern histogram features (MSLBP), that generate high-dimensional space (20 480 dimensions), trained on linear SVM. The results show that MSLBP dynamics can represent violent scenes. Even if the current scores are medium, we think that this simple visual only model can be greatly improved with some futher optimization out of the challenge schedule.

Keywords
Violent event detection, Multiscale LPB, Large scale SVM

1. INTRODUCTION
The definition of *violence* is very broad and can be related to a specific image (for example where blood appears), action dynamics in a video (car chases) or even more complicated aspects like gloomy and oppressive atmospheres that can be difficult to define in a formal way. For the above mentioned reasons, we propose here to map violence representations into the time dynamics of a high dimensional representation of visual content. Thus, we decide not to use additional meta-data information given with this challenge [1]: neither visual concepts (blood, fights, fire, firearms, cold arm, car chases, nor gory scenes), nor audio concepts (presence of gunshots, explosions, nor screams). We then integrate visual information directly from the video stream downsampled at 5Hz. For each image, we generate a vector of 10240 elements that represents the Multi-Scale Local Binary Pattern (MSLBP) of the image. MSLBP is our proposed extension of the local binary pattern which provides texture descriptors at different scales, already used during the Mediaeval 2011 Affect Task in a unsupervised system and in others [2, 3]. The conclusion suggests straightforward improvement and the fact that this approach provides an original and complementary information to usual ones.

2. MULTI-SCALE LBP DYNAMICS
We define two multi-scale versions of the LBP operator for an image I $(n_y \times n_x)$, *i.e.* the HC operator and its *improved* variant HIC.

MediaEval 2012 Workshop, October 4-5, 2012, Pisa, Italy. Acknowledgements: work partly done during the master of V. Martin, and supported by COGNILEGO ANR 2010-CORD-013, Scene Analysis 2011-2016 of the Institut universitaire de France, and CNRS PEPS RUPTURE Scale Swarm Vision projects. We thank Technicolor for shot detections.

Basically operator HC encodes the relationship between a central block of $(s \times s)$ pixels located in (y_c, x_c) with its 8 neighboring blocks, whereas HIC adds a ninth bit encoding a term homogeneous to the differential excitation [4]. Both can be considered as a parametric local texture encoder for scale s. In order to capture information at different scales, the range analysis $s \in \mathcal{S}$, is typically set at $\mathcal{S} = [1, 2, 3, 4]$, where $S = Card(\mathcal{S})$:

$$\begin{cases} C(y_c, x_c, s) = \sum_{i=0}^{i=7} 2^i 1_{\{A_i \geq A_c\}} \\ IC(y_c, x_c, s) = \sum_{i=0}^{i=7} 2^i 1_{\{A_i \geq A_c\}} + 2^8 1_{\left\{\sum_{i=0}^{7} A_i \geq 8Ac\right\}}, \end{cases} \quad (1)$$

where Ac and $\{A_i\}_{i=0,\ldots,7}$ denotes the area of the central block and the areas of its 8 neighbors. The areas in eq.(1) are computed efficiently using the image integral. As in the BoF framework [5], efficient descriptors corresponding to the operator $op = HC$ or $op = HIC$, are obtained by counting occurrences of the j^{th} parametric visual LBP at scale s in a ROI $R \subseteq I$:

$$h_{op}(R, j, s) = \sum_{(x_c, y_c) \in R} 1_{\{op(y_c, x_c, s) = j\}}. \quad (2)$$

Then the full histogram HC for $op = HC$ (respectively HIC for $op = HIC$), with $b = \{256\}$ bins (512 respectively), is:

$$h_{op}(R, s) = [h_{op}(R, 0, s), \ldots, h_{op}(R, b-1, s)]. \quad (3)$$

In order to improve discrimination between classes we take into account additional local information: the entire image can be divided into several sub-windows (possibly overlapping) *via* a spatial pyramid $\underline{\Lambda}$ defined with $\underline{L}$ layers. For each layer $l = 0, \ldots, \underline{L} - 1$, I is divided in $\{\underline{R}_{l,v}\}$ ROI's, with $v = 0, \ldots, \underline{V}_l - 1$ where $\underline{V}_l$ is the total number of sub-windows for the l^{th} layer. A total of $\underline{V} = \sum_{l=0}^{\underline{L}-1} \underline{V}_l$ histograms $h_{op}(\underline{R}_{l,v}, s)$ (where $op = HC$ or $op = HIC$) are computed where $\underline{R}_{l,v}$ is the v^{th} sub-window of layer l. For each scale s, the feature vector $h_{op}(\underline{\Lambda}, s)$ is obtained by the weighted concatenation of all sub-window histograms. Then: $h_{op}(\underline{\Lambda}, s) \triangleq \{\underline{\lambda}_l h_{op}(\underline{R}_{l,v}, s)\}$, where $l = 0, \ldots, \underline{L} - 1$, $v = 0, \ldots, \underline{V}_l - 1$, and $\underline{\lambda}_l$ denotes the weight applied to all sub-windows of the l^{th} layer. In the case of S different scales are simultaneously used, the total dimension of the feature vector $h_{op}(\underline{\Lambda})$ is $\underline{d} = b\underline{V}S$. Then a ℓ_2 norm is applied if a linear hyperplane separator is the classifier [6]. It yielded to the best result in automatic ImageCLEF plants identification [7].

In the present dynamic modeling, we compute the time derivative of MSLBP consecutive images (5 frames per sec). Using $op =$

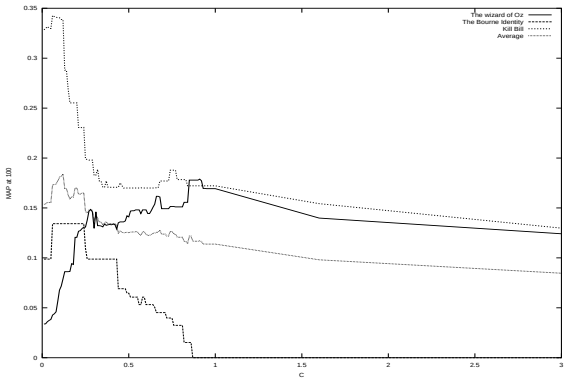

Figure 1: MAP at 100 as a function of C on dev. set.

Table 1: AP at 100 for test runs.

	Dead P. S.	Fight C.	MAP
RUN1 (C=0.2)	0	0.237	0.118
RUN2 (C=0.4)	0.133	0.21	0.171
RUN3 (C=0.8)	0.108	0.205	0.156
RUN4 (C=1.6)	0.108	0.183	0.145
RUN5 (C=10)	0.071	0.076	0.073

HC, we build a fixed-size matrix (5*20 x 10240) each 20 seconds (shifted of 20 sec.) by stacking each MSLBPs. We compute their 99 time derivatives, from which we mesure average and variance in time, resulting into a 2x10240 vector. Each of them is used as positive sample (resp. negative) if it is entirely within a violent (resp. a non-violent) shot. Then we train Linear Support Vector Machine (LSVM [6]), using BourneIdentity, KillBill, and WizardofOz as development set. We map each 20 sec. section estimates with non fixed size shot by setting the score of a shot to the maximum of the scores of the sections having a non-empty intersection with this shot. The Fig.1 shows the AP for various LSVMs (with cost C from 10^{-3} to 10^{6}).

3. RESULTS AND CONCLUSION

Figure 2: Consecutive keyframes from *Dead Poets Society* for non-violent shots (TOP), Prob=0.19 run2, near shot 1142 (resp. violent (BOTTOM), Prob=0.50, run2, near shot 981).

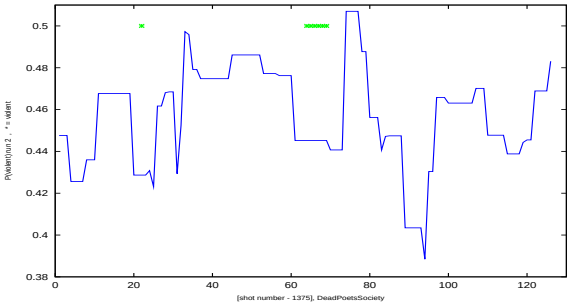

Figure 3: Estimations for each shot to be violent (run 2, DeadPoetsSociety, from shot 1375, *= is violent) showing coherent estimation variations, but suffering from desynchronization.

The MAP@100 (see Tab. 1) for each run on two test movies (other runs on the 3rd movie are misleading due to feature computation errors) demonstrate a good correlation with the MAP from the development set. The Fig. 2 gives keyframes of two continous shots that have been estimated as the lowest vs highest violent, and which are correctly estimated. This shows that even if the non violent shots depict people running, they are not labeled as violent: the visual dynamics model induced by the SVM is not directly related to MLSBP time difference, but also to their variance. The proposed visual dynamics modelization detects violence, but it will be improved in futher work by better synchronization (see desynchronization evidences Fig.3 resulting into a dramatic MAP decrease). As it does not integrate any visual concept neither audio concepts, it gives complementary information to the usual methods, and their fusion might outperform the original methods. Moreover, we will improve the feature extraction using the HIC operator, optimizing the pyramidal configuration, and better encoding scheme (Fisher vectors and/or pooling).

4. REFERENCES

[1] C. DEMARTY, C. PENET, G. GRAVIER, and M. SOLEYMANI, "The mediaeval 2012 affect task: Violent scenes detection in hollywood movies," in *MediaEval 2012 Workshop*.

[2] H. GLOTIN, J. RAZIK, S. PARIS, and J. PREVOT, "Real-time entropic unsupervised violent scenes detection in hollywood movies," in *MediaEval*, vol. 807, september 2011.

[3] S. PARIS and H. GLOTIN, "Pyramidalmulti-level features for the robotvision@icpr 2010 challenge," in *ICPR*, 2010.

[4] J. CHEN and al., "Wld: A robust local image descriptor," *IEEE Trans. PAMI, 32(9), 2012.*

[5] T. G. K. VAN de SANDE and C. SNOEK, "Evaluating color descriptors for object and scene recognition," *IEEE Trans. PAMI., 32, 2012.*

[6] VEDALDI and A. ZISSERMAN, "Efficient additive kernels via explicit feature maps," *IEEE Trans. PAMI, 2011.*

[7] S. PARIS, X. HALKIAS, and H. GLOTIN, "Dyni at imageclef 2012 plant images classification," in *CLEF*, p. 12, 2012.

[8] S. PARIS, X. HALKIAS, and H. GLOTIN, "Sparse coding for histograms of lbp applied for image categorization: Toward a bag-of-scenes analysis," in *ICPR*, 2012.

[9] B. DELEZOIDE and al., "Irim at trecvid 2011: Semantic indexing and instance search," in *Notebook*, march 2012.