Paired supervised learning and unsupervised pretraining of CNN-architecture for violence detection in videos*

Abel Díaz Berenguer[1], Meshia Cédric Oveneke[1], Mitchel Alioscha-Perez[1], and Hichem Sahli[1,2]

[1] Vrije Universiteit Brussel (VUB), Department of Electronics and Informatics (ETRO), VUB-NPU Joint Audio-Visual Signal Processing (AVSP) Research Lab. Pleinlaan 2, 1050 Brussels, Belgium

[2] Interuniversity Microelectronics Centre (IMEC), Kapeldreef 75, 3001 Heverlee, Belgium

{aberengu,mcovenek,maperezg,hsahli}@etrovub.be

Abstract. Recognizing violence in crowded scenes is a major challenge for automatic video surveillance. Indeed, there is a growing need of intelligent surveillance systems to strengthen public safety. In this paper we propose an effective approach to recognize violence in crowded videos based on a shallow Convolutional Neural Network (CNN) that is pretrained using an unsupervised layer-wise learning strategy. Afterwards, the pretrained hyper-parameters are fine-tuned to extract intermediate frame representations, which are subsequently aggregated via NetVLAD to obtain video representations to recognize violence in footage. Through experimental evaluation we validated that our proposal yields very competitive outcomes compared to results reported in the state-of-the-art.

Keywords: Convolutional Neural Networks · feature representations · spatio-temporal aggregation · violence recognition.

1 Introduction

Automatic video analysis for violence recognition has a broad array of applications in different domains. Over the years, several methods have been proposed to recognize violence in videos. However, current methods depend on hand-crafted techniques that hinder features generalization or data-driven approaches based on deep models pretrained using non crowded images, which limits their performance when dealing with crowded scenes. Our core contribution is a training strategy to pair supervised learning and unsupervised pretraining of a CNN-architecture, along with temporal pooling aggregation to attain a synthesized sequence representation for violence detection in crowded video footage.

* This work has been partly supported by the INNOVIRIS project "ADVISE – Anomaly Detection in Video Security Footage", and the VUB-IRMO Joint PhD grant.

2 Proposed approach

We propose a data-driven approach to extract CNN features with a shallow architecture from each frame, which are locally aggregated to encode video-level discriminative representations that are used by a classifier to recognize violence in videos. Our framework consists of four components: the Intermediate representation, the Spatio-temporal aggregation, the Context Gating and the Recognition. The video sequence is the input to the Intermediate representation which is employed to capture frame-level descriptors using a CNN. This CNN model is initially pretrained with an unsupervised representation learning strategy [2] to discover nuances directly from crowded scenes. Afterwards, we adopt a generalization of NetVLAD [3] to aggregate the frame-level CNN information while neglecting redundant features. The Spatio-temporal aggregation yields a synthesized representation of the video footage, whose components are subsequently recalibrated with the Context Gating. Finally, the sequence representation is fed into a classifier to recognize would the input video has violent content.

3 Experimental results

To evaluate the performance of our proposal we carried out experimental evaluation on the Violent Flows [1] dataset. The proposed approach, using unsupervised pretrained CNN and NetVLAD, overcomes the results reported in state-of-the-art studies and achieves an average accuracy score of $95.94\% \pm 2.43$ in the challenging dataset.

4 Conclusions

In this paper, we have presented an approach to recognize violent crowded scenes. Our proposal is based on frame-level CNN features to attain an intermediate spatial representation followed by an aggregation technique to obtain a representation of the footage content. Experimentation on a publicly available dataset proved that automatic crowded scenes analysis can benefits from data-driven domain specific representations while not such deep CNN structures are required to achieve state-of-the-art performance.

References

1. Hassner, T., Itcher, Y., Kliper-Gross, O.: Violent flows: Real-time detection of violent crowd behavior. In: 2012 IEEE Computer Society Conference on Computer Vision and Pattern Recognition Workshops. pp. 1–6 (June 2012). https://doi.org/10.1109/CVPRW.2012.6239348
2. Oveneke, M.C., Aliosha-Perez, M., Zhao, Y., Jiang, D., Sahli, H.: Efficient convolutional auto-encoding via random convexification and frequency-domain minimization. In: NIPS 2016 International Workshop on Efficient Methods for Deep Neural Networks (EMDNN) (2016)
3. Zhong, Y., Arandjelovic, R., Zisserman, A.: Ghostvlad for set-based face recognition. CoRR **abs/1810.09951** (2018), http://arxiv.org/abs/1810.09951