Second Report: Literature review

Steven Le Moal[1*], Ahmed Azough[1], Dirss Yakoub[1], Nicolas Travers[1]

[1]Devinci Research Center [*]Research student responsible for this report

1 Research Topic

Title:

Detection and prevention of urban violence events in a distributed video surveillance system

Aims:

Conception of a method capable of tracking meaningful objects and detecting events (especially violent behavior) happening in real-time in a multi-camera video system. Video detection has received incremental interest in research, but a lot of existing techniques are not implemented for the problem (ex: Vision Transformer), so we'll discuss current challenges with the intention to use these algorithms in real-time on low-cost hardware (distributed surveillance system).

Objectives:

1. Extraction of objects and meaningful events (firstly violence than other human activity) from a video sequence using existing methods to create a semantic scene representation in a graph (with the PhD student)

2. Experiment with urban violence prediction and tracking for distributed video surveillance system (multiple views) using the constructed graphs.

3. Describe the result and propose a viable solution

2 Research questions

Questions:

- How can we learn and construct the relationship between high-level semantic representation of events and low-level (and mid-level) detections?
- How can we extract realistic descriptions and interactions from video footages?
- How do we narrow the semantic gap?[1]
- How can we link different scene semantic representation into a real-world description (meaningful interpretation)?
- How can we minimize the training time and the hardware requirement for an automatic distributed surveillance?

Keywords : *Artificial Intelligence, machine learning, neural networks, deep learning, violence detection, activity recognition, big data, video data, smart surveillance, distributed system*

[1]Semantic gap: "Understanding semantics is the most fundamental step in all kinds of computer vision problems as it paves the way for general artificial intelligence. The semantic gap is a general term widely used in content-based image retrieval (CBIR). It is defined in [134] as follows. "Humans tend to use high-level concepts in everyday life. However, what current computer vision techniques can automatically extract from image are mostly low-level features. In constrained applications, such as the human face and fingerprint, it is possible to link the low-level features to high-level concepts (faces or fingerprints). In a general setting, however, the low-level features do not have a direct link to the high-level concepts.

3 Relevant sources

What are the leading conferences, journals, and authors concerning your topic? Evaluate the selected sources.

Conferences: Conference on Computer Vision and Pattern Recognition (CVPR), International Conference on Computer Vision (ICCV), European Conference on Computer Vision (ECCV) and International Machine Learning Society (ICML).

Journals: IEEE Transactions on Pattern Analysis and Machine Intelligence, Pattern Recognition, International Journal of Computer Vision and Computer Vision and Image Understanding.

4 Context

Why is this question important to answer? How and why does it include in the current state of the world? What were the objectives of your research?

Violence is a part of the human experience, and its impact may be seen in several ways around the world. Individual (even self-inflicted) or social violence kills millions of people each year, with many more suffering non-fatal injuries. As a result, violence is one of the leading causes of mortality worldwide among persons between the age of 15 and 44 (5). Violence (both verbal and physical) can occur at any moment and in any location. Violence has expanded swiftly at universities, whether among teaching staff, administration personnel, students, or others. Even still, violence against healthcare professionals is on the rise throughout the world: the World Medical Association defined violence against healthcare employees as" an international emergency that undermines the very foundations of health systems and impacts critically on patients' health" (7). Violent behavior may occur in any crowded setting, including shopping centers, malls, sports stadiums, prisons etc... Because of the growth in security concerns throughout the world, the use of video cameras to monitor persons has become critical, and early discovery of these violent behaviors may considerably reduce risk.

Given that violence might strike at any time, relying exclusively on humans to undertake the task of monitoring and recognizing violent or dangerous situations is inadequate (ineffective). Typically, this role has required security personnel to constantly watch a large number of monitors. As a result of human weariness (exhaustion), poor attention, and inexperience, occlusions, violent behavior (including violence) events can be missed. Therefore, to assist security staff and ensure safety, automated surveillance systems that recognize specific patterns, such as abnormalities, in real time are critical.

Due to the repetitive nature of the task of video monitoring and the time necessary, real-time automatic detection of violence is essential to prevent such situations. However, occlusion, poor image resolution, visual noise, fast changes in events and/or interactions, data encoding, illumination changes, and other factors make the process difficult. A methodology and good practice for distributed video surveillance must be generalized (based on previous works).

An increasing number of surveillance cameras are placed across the world to monitor and assure safety. These surveillance systems are useful in numerous domains and produce a lot of data. Instead of having the produced data, it will be more useful to have some analysis result or a process data (ex: key frame, graph scene representation...) to maximize efficiency and reduce the necessary storage space. An intelligent monitoring system is required to solve this problem. With the huge amount of data, we have (even though a lot of the data is kept private), deep learning techniques can achieve the highest accuracy. There are several problems researchers face due to the subjective nature of violence detection (and prediction). One of the major problems is that tasks like crowd analysis need a lot of improvement (to be able to be performed in real-time, on long video, and on low-cost hardware).

5 Literature review

5.1 Mind map and/or concept map

Violence detection is among the video classification groups (Fig. 1) focused on the detection of violent event patterns from an input video. Violence detection is supported by the work done on video indexing, retrieval, and summarization, as well as activity recognition, but it also refers to abnormal patterns detected from a normal surveillance video that could occur with varying lengths of duration.

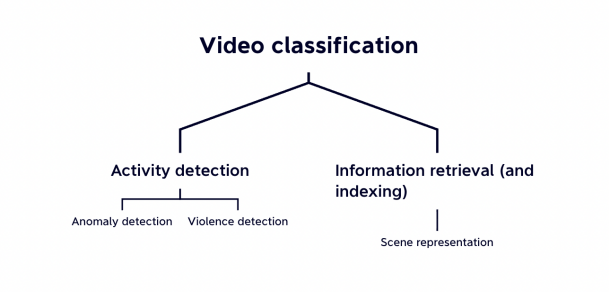

Figure 1: Map

Its early detection assists in either prevention or reducing the damage done by initiating protocols (ex: calling for rescue or the authority). Violence detection is a subdomain of the activity recognition domain that has not been explored as much as its potential real-world values. Violence detection is facing various problems, such as the limited datasets that are publicly available. Even though the main goal of a surveillance system is to detect and report any abnormal behavior, abnormal activity recognition is different from violent activity recognition.

**Described in detailed in the state of the art*
Violence detection can describe in two major ways (Fig. 2). The prior research violence detection methods followed the basic papeline of generic classification problem 'solve' using machine learning :

1. **data preprocessing**: data cleansing and video segmentation
2. **features extraction**: extractions of features such as motion, speed, and optical flow.
3. **classification** : classification of these features using methods such as support vector machine (SVM), K-means, AdaBoost.....

The advents of deep learning in computer vision domains and its advancement, showed tremendous improvement in accuracy on variation dataset. In most of the cases, they are marked by an absence of preprocessing and have an end-to-end architecture : input of a sequence of frames (video) or a single one and output their probability of being violent or not. We can distinguish between three type of model :

- **spatial** : CNN 2D, ...
- **temporal** : RNN, LSTM, ...
- **hybrid** : CNN - BiLSTM, CNN 3D ...

5.2 State of the art

This paper (2) provided an assessment of video violence detection problems described in the state-of-the-art research (qualitative and quantitative) in terms of methodology (procedure), including datasets, and performance metrics. Additionally, challenges and potential future directions are also discussed. Based on this paper, we can divide the state of the art following the pipeline described above.

5.2.1 Feature extraction

Features extraction is one of the core steps to obtain meaningful information from large-scale data for processing. There are two basic types: traditional feature collection and NN-based deep feature collection. This section discusses various types of feature extraction methods, the most popular of which are also the most useful.

5.2.1.1 Handcrafted features

There are a large number of features that are interesting : motion blobs, motion vector, speed direction

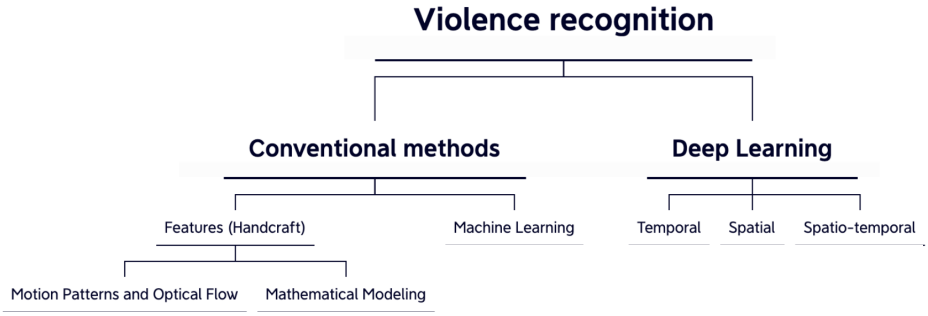

Figure 2: Map

and density, acceleration and movement (with direction), spatio-temporal and motion streams, optical flow and motion region.

Key points detection is largely used, because it finds the points in an image distinct of its surroundings, and ideally it considers two segments : robustness and computational speed. Some state-of-the-art feature extraction techniques include SIFT[2], SURF[3], MoSIFT[4], HOG[5], HOF[6] with the use of BoW[7]. There existed varient/ extension of the above feature such as STIP, Violence flow descriptor, Motion boundary histograms, MoBSIFT and many other (2: page 7-11). Even though these features can give detail on a specific aspect of an image, they are very costly to compute and give bad results with bad illumination, occlusion, or slit movement of the camera.

5.2.1.2 Deep Learning

CNN (or ConvNet)[8], which consist in most cases of three layers (convolutional layers, pooling layers, and fully connected layers), are the most popular subfield owing to a wide range of data representation(2: page 16-20). The major advantage of convolutions is the extraction of deep feature information from frames, computing the correlation between the pixels, with the aspect of location also included. Different configuration of layers (ex : softmax layer, morphological analysis, encode/decode...) can be applied to get the best convergence or get more in-depth information. The NN-model are mostly trained in a supervised manner, and computed the gradients through a back-propagation method. Deformable or dilated convolutions can also be interesting given the additional information we can get, even though they are most costly to compute.

5.2.2 Learning

Several strategies have been used to extract and learn features from violent video sequences. These features are derived from either a sequence of frames or a single one following a learning approach. There are two sorts of learning approaches: traditional machine learning approaches and deep learning-based approaches.

[2]Scale-invariant feature transform (SIFT): maxima and minima of the result of difference of Gaussians function applied in scale space to a series of images

[3]Speeded Up Robust Features (SURF)

[4]MoSIFT is a popular extension of SIFT

[5]Histogram of oriented gradients (HOG): corresponds to a gradient direction distribution

[6]Histogram of opticals flow (HOF) : optical flow characteristic that describes the series of events at each point in time.

[7]Bag of words (Bow): represents feature as vector, by building a vocabulary of the visual words obtained from the features, then we can use classification techniques

[8]There is also 2D CNN, 3D CNN (spatio-temporal),...

5.2.2.1 Machine learning

In this section, we discuss traditional machine learning techniques. The early VD methods (2: page 7-15), methods were largely considering of only image processing and pattern recognition methods to predict violence (or abnormal) events. Once features are extracted, they are fed into classifiers such as SVM [9], KNN, K-Means, Naïve Bayes, particle filters with an approximation inference scheme that computes the distributions of hidden Markov model, AdaBoost... In supervised learning-based approaches (for ML), the most popular approach used by researchers is the support vector machine (SVM) approach.

Machine learning is mostly limited to situations with instant normal motion, not complex real-world situations. Traditional image-based surveillance methods are designed to function for static data, where actions and events occur in a sequence of images. Most of these methods do not result in trustyworthy decisions due to their spatial image processing (only 1 dimension, even though there are works done on multi-modal data).

5.2.2.2 Deep Learning

Violent video classification (video classification by extension) is an intriguing subject with a primary focus on effective temporal content representation, which contributes greatly to exact video label prediction. Early research used basic CNNs, but modern methods use different types of temporal and spatio-temporal approaches (2: page 16-21). This paper (8) discusses the performance of training neural networks (CNN in this case) using transfer learning, which can be very helpful for deep learning training.

Spatial deep models use features to classify the events (activities). Spatial deep learning approaches are uncommon because by nature, a pattern emerges in a sequence of frames, hence, working with spatial data is not always successful for an activity recognition task.

Deep sequence learning models are the most popular research topic. RNNs[10] are used to comprehend these sequences. One of the problem is the lost of long-term sequential information, also known as vanishing gradient, and it may be solved using a certain sort of RNN using LSTM[11]: ; and use multilayer LSTM networks. We see an improvement in performance when adding additional network layers, for sequential data such as violent frame sequences, which are analyzed by multiple layers.

5.2.3 Results discussion

The **dataset** for violence detection are described in (2: page 26-27) and the most popular are : Violence in Movies, Violent Crowd/Violent Flows, Hockey Fight, RWF-2000 and UCF Crime. But because this is a new topic, there are too few publically accessible datasets for video violence detection (Furthermore, we can also talk about the data imbalance, illumination variation or motion blur in some dataset). Violence detection is in need of appropriate standards for dataset.

Various **performances metrics** are used such as accuracy, precision, recall, F1-score, AUC-ROC, false alarm and missing alarm. Other metrics should also be incorporate, which are not enough used, computational and time-consuming cost, to have a more realistic discussion of performance result.The use of deep sequential learning algorithms improved the performance of VD techniques significantly. Because of their long-term accurate and clear spatio-temporal information dependencies learning abilities, emerging DNN[12] (such as CNN-BiLSTM) are widely used in many video classification tasks as it outperforms other approaches.

The most challenging work that need to be is to continue to explore new approaches for better performances and reduce the cost and computational resources : processing time, number of parameters, and model size. The promising approaches are Federated Learning (for a distributed system especially), Reinforcement Learning and Vision Transformer(1), which have shown tremendous performance in general activity recognition which will be discussed in the next section (**not added).

[9] Support Vector Machine (SVM)
[10] Recurrent Neural Network
[11] Long-Short Term Memory
[12] Deep Neural Networks (DNN)

6 Notation

- Research topic → 1 point (clarity of the aims and objectives)
- Research question → 2 points (clarity, relevance, and defense of the questions)
- List of keywords → 1 point (relevance concerning your topic)
- Relevant sources → 1 point (relevance concerning your topic)
- Context → 5 points (clarity, relevance, and discussions)
- Literature review 1 (Mind Map) → 5 points (clarity, relevance of the scheme/table)
- Literature review 2 (State of the art) → 5 points (clarity, relevance, and discussions)

References

[1] Ganna Pogrebna Ajmal Mian Anwaar Ulhaq, Naveed Akhtar. Vision transformers for action recognition: A survey. *arXiv*, 2022. 5

[2] Zhandos Zhumanov Aidana Gumar Batyrkhan Omarov, Sergazi Narynov1 and Mariyam Khassanova. State-of-the-art violence detection techniques in video surveillance security systems: a systematic review. *PeerJ Computer Science*, 2022. 3, 4, 5

[3] Amin Ullah Khan Muhammad Mohammad Hijji Sung Wook Baik Fath U Min Ullah, Mohammad S. Obaidat. A comprehensive review on vision-based violence detection in surveillance videos. *ACM Compt. Surv.*, 2022.

[4] Shabana Habib Syed Muhammad Mohsin Prayag Tiwari Shahab S. Band Neeraj Kumar Nadia Mumtaz, Naveed Ejaz. An overview of violence detection techniques: Current challenges and future directions. *arXiv*, 2022.

[5] World Health Organization. World report on violence and health. 2002. 2

[6] World Health Organization. World report on violence and health. In *Advances in Neural Information Processing Systems*, 2015.

[7] Francesca Cainelli Sandro Vento and Alfredo Vallone. Violence against healthcare workers: A worldwide phenomenon with serious consequences. 2

[8] Wanli Ouyang Wenhao Wu, Zhun Sun. Transferring textual knowledge for visual recognition. *arXiv*, 2022. 5