IOSUD – Universitatea Politehnica Timişoara
Şcoala Doctorală de Studii Inginereşti

"BEHAVIOUR CLASSIFICATION IN URBAN AREA USING VIDEO BASED SENSOR NETWORKS"

PhD Thesis – Abstract

for obtaining the Scientific Title of PhD in Engineering from
Politehnica University Timisoara
in the Field of Computers and Information Technology

author ing. Marius BABA
thesis supervisor Prof.univ.dr.ing. Ionel JIAN

Contents

The purpose of this thesis is to develop algorithms for analyzing video sequences to allow automatic detection of violent behavior in urban areas using a network of video sensors. The proposed algorithms use techniques for analyzing video sequences that do not require complex calculations and thus allow each camera in the network to process the collected video data individually. This approach makes it possible to analyze a large volume of data and at the same time improves the detection rate of events of interest by the surveillance system.

1. Introduction

This chapter describes the motivation for choosing the PhD research topic. We are all witnessing cities becoming overcrowded. For this reason, we often encounter conflicts between people on the streets. Most of these conflicts are aggressive behavior of traffic drivers and street fights between various criminal groups. In addition to interpersonal conflicts, a number of other violent events occur in the urban environment, such as stray dog attacks on pedestrians.

In order to prevent such situations specific to urban areas, a network of video sensors can be used together with a manual method for analyzing video sequences. The idea behind this approach is that, once detected, the suspicious event is transmitted as soon as possible to the competent authorities in order to take swift action to combat these undesirable situations. The main problem of this approach is the manual analysis of video sequences. Usually this task is performed by security officers, who, with the help of several monitors connected to various video sensors, visually inspect the footages captured by the sensors.

Visual inspection of video sequences is not a very effective analysis method. Human eyes get tired very quickly, especially if they constantly change their focus of attention. For this reason, many events of interest remain undetected. According to the experiment presented in the paper [1] a man fails to detect about 80% of the actions present in a single video with a duration of 30 minutes. This experiment is a clear indicator that processing videos manually is not efficient at all. Moreover, visual inspection of large volume of videos in real time is an expensive operation because it involves a lot of human work.

An alternative to this method is the automatic analysis of video streams using computer vision algorithms. In this process the videos are analyzed by the computer and only the events of interest are transmitted to the security officers. This approach allows the detection of events of interest with high accuracy and makes it possible to analyze in real time a large volume of video data at much lower costs.

However, video analysis algorithms are very rarely used in video surveillance systems because they face a number of problems. Indeed, event detection involves recognizing human behavior in video, which is not an easy task. This operation involves complex video analysis techniques which in turn use a lot of hardware resources. For this reason, most existing algorithms in the literature cannot be used for real-time analysis of a very large volume of video sequences.

Moreover, in many cases the behaviors are ambiguous, which impedes the proper functioning of the analysis algorithms. The same set of actions can represent totally different behaviors. For example, the act of running may be associated with normal behavior if the analyzed individual does sports or may be associated with abnormal behavior if the individual has committed a theft and runs so as not to be caught by the victim.

In order to allow surveillance networks to analyze video content in real time, this thesis proposes three algorithms designed for the analysis and detection of suspicious behavior in urban areas. The rapid detection of dangerous events is of particular importance because it makes possible the rapid intervention of police officers, which means an improvement of public security in the urban environment.

2. Why human behavior understanding is important

Chapter 2 contains a brief description of the importance of recognizing human behaviors. Also, here are studied the taxonomies used by algorithms for detecting human behavior in video, and some surveillance systems that use video analysis algorithms to detect either human action or human behavior in video.

3. Thesis structure

The thesis is organized as follows:

- Chapter 1 presents an introduction to the topic addressed in the thesis.
- Chapter 2 discusses the importance of automatic recognition of human behavior in video sequences.
- Chapter 3 details the structure of the thesis.
- Chapter 4 presents the current state of the art in video analysis algorithms, the field to which this thesis is addressed.
- Chapter 5 presents the first video analysis algorithm designed in this thesis. It is intended for traffic analysis of a congested intersection. Also in this chapter are presented the video data sets created in order to test and validate the efficiency of the algorithm. The results obtained by the algorithm are presented at the end of the chapter.
- Chapter 6 presents the second algorithm designed in this thesis that uses the properties of moving objects in the video, such as speed, acceleration, and trajectory, to recognize the attack of stray dogs on a person. The results obtained by the algorithm are also presented at the end of the chapter.
- Chapter 7 discusses the third algorithm developed in this thesis. Using motion descriptors extracted in the process of decoding MPEG encoded video sequences as well as artificial intelligence methods based on deep learning techniques, this algorithm is able to recognize complex human behaviors in urban environments, such as street fighting. The results obtained by this algorithm are presented at the end of the chapter.
- Chapter 8 presents the final conclusions as well as the list of personal contributions.

4. State-of-the-art

Chapter four is dedicated to the study of existing algorithms in the literature. This chapter examines several approaches to recognizing both human actions and behaviors in video.

The algorithms studied in this chapter are divided into two categories:

- handcrafted algorithms
- algorithms based on deep learning techniques

Handcrafted algorithms use video and image analysis techniques to detect human actions and behaviors in video. These algorithms are composed of several auxiliary algorithms. Each auxiliary algorithm analyzes only certain aspects of the action in the video. The results of analyzes are then combined to obtain the final result. The advantage of this method is that it allows designers to choose the right techniques to obtain a high-performance algorithm for detecting human behavior in video. In the section dedicated to these algorithms, several representative algorithms of this type were studied, such as algorithms proposed in papers [2], [3] and [4].

In the section dedicated to algorithms based on deep learning techniques, concepts used by this technique were presented, as well as some algorithms of this type such as those in [5], [6] and [7]. Algorithms based on deep learning are composed of an artificial neural network and a post-processing phase. The recognition of actions is performed by the neural network and the detection of behavior is performed by the rule based expert system in the post processing phase. Algorithms based on deep learning do not use auxiliary algorithms. Instead, they use training data to learn the features of objects in the process of classifying human behavior in video.

5. Basic behavior classification in low computational environments. Traffic surveillance application

This chapter presents a handcrafted algorithm for detecting and tracking vehicles in a crowded intersection. The designed algorithm is published in the article [8] and is a first step towards the development of advanced algorithms capable of recognizing complex human behaviors in video.

At the beginning of the research, simple scenarios were analyzed, such as the one regarding the traffic of vehicles in a crowded intersection. It is known that the directions of movement of vehicles are imposed by the road on which they travel. Therefore, vehicles can only move in certain directions and respect certain physical laws, which involve generally simple (basic) behaviors. Random, unpredictable trajectories are excluded in this scenario.

The algorithm proposed in this chapter is composed of the following functional parts:

- filtering phase

- foreground extraction
- correction of the foreground mask
- detection of the vehicles features
- post processing phase
- classification and counting phase

The part of the algorithm responsible for filtering the zones aims to eliminate irrelevant objects from the analysis process. At this stage, the part of the road on which the vehicles travel is extracted. Only this part is subject to the analysis process. Thus, irrelevant objects, such as pedestrians walking on the sidewalk, blocks near the intersection, or trees near the road are ignored by the algorithm.

The next step is extraction of the foreground objects which, as the name suggests, aim to separate the foreground objects from the background objects. This task is performed by the MOG (Mixture of Gaussians) algorithm proposed in the article [9]. It is a very common algorithm used in video processing techniques because it gets excellent results. Periodic updating of the background model allows it to adapt very easily to the changing lighting conditions of the analyzed scene. Thus, the foreground objects are separated with great precision from the background objects.

Despite its high accuracy, the MOG algorithm does not always manage to classify all pixels correctly. Very often, the foreground mask generated by the MOG algorithm also contains a small number of misclassified pixels that negatively affect the vehicle detection process. To remedy this problem, the foreground mask is corrected with the help of morphological operators. Morphological operators are very useful because they do not require complex calculations and manage to significantly reduce the number of incorrectly classified pixels.

The foreground mask is then processed by the next phase of the algorithm which is responsible for extracting the features of vehicles. In this phase the algorithm analyzes the foreground objects and associates to each detected object a bounding rectangle. Bounding rectangles are calculated using the contours of objects that are extracted from the foreground mask using the algorithm proposed in article [10].

The features of vehicles are then filtered. This operation is performed in the post-processing phase and aims to remove objects that are not vehicles from the analysis process of the algorithm. Also in this phase is the method of correcting the wrong extracted contours.

The last part of the algorithm is responsible for classifying and counting vehicles. At this stage, the vehicles are classified into either large or small vehicles and then counted. Small vehicles represent the class of cars, while large vehicles represent the class of large commercial vehicles (trucks).

The results obtained by the algorithm as well as the details related to the two video data sets created in this thesis are presented at the end of this chapter.

6. A framework for behavior classification based on motion understanding

Chapter 6 describes the second algorithm conceived in this thesis that was published in the article [11]. It is developed to recognize complex behaviors in video such as stray dog attacks on humans.

Similar to the algorithm introduced in the previous chapter, the algorithm proposed in this chapter is also divided into several functional parts. The operations were grouped by type resulting in:

- group one - responsible for low level processing
    - foreground extraction
    - shape classification

- group two - responsible for high level processing
    - trajectory feature extraction
    - event detection

The operation of extracting the foreground objects is performed using the MOG algorithm. The mask generated by this algorithm is then corrected using morphological operators. This technique was taken from the algorithm presented in the previous chapter because it generated good results.

The foreground objects are then classified. At this stage, the contours of the objects are extracted using the algorithm described in article [10] and then classified. The classification is performed using a robust technique, developed in this thesis, which does not use complex calculations and provides a satisfactory classification of objects.

Trajectory feature extraction is the next phase of the algorithm. At this stage, in addition to extracting the trajectories, the algorithm extracts the speeds and accelerations of objects. According to the analysis performed in this thesis, the trajectories are not descriptive enough to recognize complex behaviors. For this reason, in addition to the trajectories, additional features such as speed and acceleration of objects were used.

Thus, for each point of the trajectory of an object, the algorithm generates a composite vector that contains the position, velocity and acceleration of the object at that given moment. The resulting vectors are then provided to a support vector machines (SVM) classifier that is responsible for detecting the attack event.

The last section of this chapter contains the description of the data sets used in the experiments and discusses the results obtained by the algorithm.

7. Proposed hybrid Deep learning/VA features solution for complex behavior classification in sensors environments

In order to recognize complex behaviors in video, a third algorithm has been designed which is described in detail in this chapter. The proposed algorithm, published in article [12], is a hybrid one. It combines the methods used by handcrafted algorithms with the methods used by deep learning algorithms. This approach allows the algorithm to quickly learn the classification of video frames and makes the inference process unpretentious in terms of the use of hardware resources.

The algorithm analysis process consists of the following stages:

- motion feature extraction
- classification of video frames
- results filtering

The first stage of the algorithm uses the MPEG video codec to extract the motion features. In the decoding process, the codec provides a set of motion vectors that successfully capture the movements of objects in the video. This property of the codec is very beneficial because it provides information about the movement of objects for free (does not require complex processing). The motion features extracted in this way are called MPEG flow.

The next stage is dedicated to the classification of video frames. For this operation the algorithm uses a deep convolutional neural network. Networks of this type are able to learn the features and the classification process and thus manage to classify video frames with great precision. Despite this beneficial property, neural networks also have a disadvantage. For proper operation, the network must be trained with a very large volume of data. This process is laborious because it involves a lot of work. However, the learning process can be simplified by using manually extracted motion features. Thus, in order to reduce the volume of data required for training, the network used by the algorithm uses manually extracted motion features. That is, it uses the MPEG flow motion descriptor.

A fighting action is usually composed of violent movements followed by moments of pause. For this reason, the predictions of the network oscillate in the conditions of the presence of the fight in the video. This behavior of the algorithm is not exactly appropriate because such an algorithm is expected to provide stable predictions. To remedy this problem, the time domain filter has been designed, which successfully stabilizes the network predictions.

In addition to the description of the algorithm, chapter 7 also contains a section dedicated to the comparison of the MPEG flow descriptor with the motion descriptor generated by the optical flow which is also very frequently used for the detection of human behaviors in video. The last part of this chapter contains the description of the data sets used in the experiments, the network training procedure and the results obtained by the algorithm.

8. Conclusions

Chapter 8 contains the conclusions of the thesis. Also here is a section where personal contributions are presented, such as:

- A method for detecting violent behavior in urban areas using machine learning algorithms and a network of video sensors with limited computing resources has been proposed.
- Various algorithms for tracking moving entities in video have been studied. Case study: Urban traffic surveillance application.
- The detection of violent behaviors in the urban environment was investigated. Case study: Attacks conducted by stray dogs in urban environment.
- The performance of the motion descriptor generated by the optical flow was compared with the performance of the MPEG flow motion descriptor.
- An augmentation method has been proposed for enlarging the video dataset.
- A method for counting vehicles in video independent of driver behavior has been proposed.
- A correction method has been proposed for solving the contour splitting issue. This problem is frequent in traffic surveillance and affects negatively the performance of the video analysis algorithms.
- A method for extracting and classifying entities (eg humans vs. dogs) in video has been proposed.
- Various background subtraction algorithms were compared.
- Two video databases have been created that can be used to test traffic surveillance applications.

SELECTIVE BIBLIOGRAPHY

[1] D. Elliott, “Intelligent video solution: a definition”, Security Magazine, 47(6), pp.46–48, June 2010.

[2] Serhan Cosar, Giuseppe Donatiello, Vania Bogorny, Carolina Garate, Luis Alvares, François Bremond. "Toward abnormal trajectory and event detection in video surveillance." IEEE Transactions on Circuits and Systems for Video Technology 27.3 (2016): 683-695.

[3] Gorelick, Lena, et al. "Actions as space-time shapes." IEEE transactions on pattern analysis and machine intelligence 29.12 (2007): 2247-2253.

[4] Huangkai Cai, He Jiang, Xiaolin Huang, Jie Yang, “Violence Detection based on Spatio-Temporal Feature and Fisher Vector”, Chinese Conference on Pattern Recognition and Computer Vision (PRCV) 2018.

[5] K. Simonyan and A. Zisserman. Two-stream convolutional networks for action recognition in videos. In Advances in Neural Information Processing Systems, pages 568–576, 2014.

[6] Yue-Hei Ng, Joe, et al. "Beyond short snippets: Deep networks for video classification." Proceedings of the IEEE conference on computer vision and pattern recognition. 2015.

[7] Zhang, Bowen, et al. "Real-time action recognition with deeply transferred motion vector cnns." IEEE Transactions on Image Processing 27.5 (2018).

[8] Áron Virginás-Tar, **Marius Baba**, Vasile Gui, Dan Pescaru, Ionel Jian, “Vehicle Counting and Classification for Traffic Surveillance using Wireless Video Sensor Networks”, Conference: 22nd Telecommunications Forum (TELFOR 2014), 25-27 November 2014, SAVA Center, Belgrade, Serbia; Published in IEEE XPLORE Digital Library; Indexed in the ISI Web of Science.

[9] Z. Zivkovic. “Improved adaptive Gaussian mixture model for background subtraction”, In the proceedings of the 17th International Conference on Pattern Recognition ICPR’04, 2004.

[10] Suzuki, Satoshi. "Topological structural analysis of digitized binary images by border following." Computer Vision, Graphics, and Image Processing 30.1 (1985): 32-46.

[11] **Marius Baba**, Dan Pescaru, Vasile Gui, Ionel Jian, “Stray Dogs Behaviour Detection in Urban Area Video Surveillance Streams”, Conference: 12th IEEE International Symposium on Electronics and Telecommunications (ISETC 2016), 27-28 October 2016, Timisoara, Romania; Published in IEEE XPLORE Digital Library; Indexed in the ISI Web of Science.

[12] **Marius Baba**, Vasile Gui, Cosmin Cernazanu, Dan Pescaru “A Sensor Network Approach for Violence Detection in Smart Cities Using Deep Learning”, Journal: Sensors, Volume 19, Issue 7; Published by MDPI Switzerland, 8 April 2019. ISI journal Q1 (Instruments and instrumentation). Journal impact factor 3.735.