ISSN 1870-4069

Automatic Recognition of Abnormal Human Actions with Semi-supervised Training: A Literature Review

Diana-Karina Guevara-Flores, Josefina Guerrero-García,
David-Eduardo Pinto-Avendaño

Benemérita Universidad Autónoma de Puebla,
Facultad de Ciencias de la Computación, Puebla, Mexico

diana.guevaraf@alumno.buap.mx

Abstract. In this paper is presented the literature review and the first tests performed for the development of a method consisting of differentiating typical actions in a given environment from those that can be categorized as abnormal or atypical.The objective of this method will be to determine which are the typical situations taking into account temporal and spatial information of a given environment to generate an alarm when a potentially undesirable situation arises. The main difference between this system and those existing in the literature is that it does not seek recognition of pre-established actions, such as running or sitting, and that the system can adapt to different environments. For the development of this research the use of Deep Learning is proposed and due to the complexity of the attributes required by the classifier, the use of a semi-supervised method is proposed.

Keywords. Deep learning, atypical action recognition, semi-supervised training.

1 Introduction

This research work aims to present the theoretical basis for the development and implementation of a model that allows the extraction of features in spatial and temporal dimensions for recognition of atypical situations in uncontrolled environments.

Video analysis with machine learning is a challenge task and there are few methods to capture the movement information of adjacent frames in a long term scheme.

The recognition of human actions in uncontrolled environments has multiple applications, among which are intelligent video surveillance for the analysis of behavior in shopping center customers, students in an educational facility or passengers in an airport. In public spaces it can be very useful for the timely detection of risk situations and in private companies it can allow the optimization of internal processes for the increase of profits.

Simple systems, such as monitoring a home, can be restricted to a single camera, however, more complex tasks require a multi-camera transmission environment.

Classical methods extract the characteristics of fixed images to train classifiers, however, in real applications, even the simplest ones, it is difficult to determine which characteristics are important for a given task.

The task becomes much more complex for video analysis. Most of the proposed methods treat each frame as a fixed image, losing the temporal information. Other methods perform short-term analyzes taking into account only the temporal information of a few frames, which makes the analysis very limited. Finally, the methods that take into account information to long-term are too costly computationally.

Additionally, the classification of atypical actions or situations can be complex even for humans. Actions that should be classified within the same class may seem dramatically different in terms of appearance and movement patterns. In the same way, it is possible that risky actions seem normal either because the individual who perform them pretends that this is the case or because of inherent faults in the system, such as the location of the cameras.

Another problem arises when analyzing highly concurred sites. To perform the analysis, the system must track each of them, for which it will be necessary to identify each individual in the scene. An analysis that includes the authentication of each individual for the analysis of their behavior over extended periods of time, would require limiting the system to a context of fixed users, for example, a company where workers are always the same or a university where students, teachers and administrative staff have relatively fixed routines.

This document is organized as follows: section 2 presents the state of the art; the characteristics of some of the methods used for video processing are presented, as well as the advantages and disadvantages of using each of them. Section 2.1 presents the methods used to evaluate the results obtained by some of the works analyzed in this document. Section 2.2 exposes some of the commercially available systems for video processing. Finally in section 3 the results of this research work are discussed.

2 Literature Review

For the development of the proposed method, an extensive literature review of the existing systems and methods is necessary, as well as the applications for which they have been required and the level of success achieved in their implementation. Also, a search of commercially available systems has been made, some of them dedicated to simple applications, such as monitoring of homes. It have also analyzed complex systems that involve a large number of modules and are capable of monitoring in real time large companies or public sites with a higher level of intelligence.

Models based on deep learning can be trained using supervised, semi-supervised or unsupervised approaches. The supervised approach is the most widely used

due to the accuracy of the results obtained but it requires a large amount of tagged data for the training.

There are extensive databases well labeled for supervised training of still images, however, for video tasks there are few data bases and none of them is as widely labeled as images databases. For this reason, some researchers have opted for the realization of synthetic labeling video databases, made from image databases.

Some tasks in which good results are presented with the implementation of Deep Learning techniques are visual recognition of objects, recognition of human actions, natural language processing, audio classification, tracking, image restoration, noise elimination and segmentation, among others. Even some of the results presented in the literature point to a performance superior to that obtained by humans in the same task.

One of the most widely described methods in the literature for the recognition of human actions in video is through convolutional neural networks. Traditionally, this approach allows to treat the video frames as fixed images and perform the recognition of the actions at the level of individual frames, however, in this way the temporal information is not processed and it is not possible to perform movement analysis. This approach is useful for recognizing actions such as sitting or talking on the phone, but does not provide information about how long someone sat or how many different calls an individual made.

To take advantage of temporal information, a proposed approach in the literature is based on 3D convolution to obtain characteristics not only in the spatial dimension [1]. With this method multiple information channels are generated from adjacent video frames, with which it is possible to perform the convolution and sub-sampling in an individual way and then combine the information of all channels.

For the evaluation of the method, a database provided by TRECVID [2] was used, which consists of a 49-hour surveillance video of the London airport using five different cameras with a resolution of 720x576 at 25 fps. This work focused on the recognition of three kinds of actions: talking by cell phone, leaving an object and pointing. In addition, a large number of samples of actions that are not within these classes were generated.

The authors compare the method with a version of the 2D model and four variants of spatial pyramid matching. The results of the cross validation allow to evaluate the performance of each method using three measures: precision, recall and area under the ROC curve. The results of the performance of the proposed method outperform the rest in two of the evaluated actions: talking by cell phone and leaving an object, but are slightly exceeded in the third action: pointing.

In that investigation, because the videos were recorded in a real environment, each frame contains several people, so it was necessary to apply a human detector and tracking. Finally, the developed model was trained using a supervised algorithm, so a large number of labeled samples was required. The number of samples needed can be significantly reduced if a semi-supervised method is used.

Fig. 1. Results for human detection and tracking using TRECVID database [1].

Another method used in the literature for video analysis is deep belief networks, which are probabilistic generative models composed of multiple layers of latent stochastic variables [3]. These latent variables typically have binary values and are called hidden units or feature detectors. This type of networks has generated a lot of interest because they are based on an unsupervised method of learning hierarchical generative models, however the processing of high-dimensional images is limited, so modifications have been proposed for the treatment of images in its real size by means of hybrid networks between deep beliefs and convolutional networks.

A common problem in the treatment of video by means of Deep Learning is the difficulty in handling videos of variable length. This generates degradation in the efficiency of the classification, so that approaches have emerged that allow the representation of video content of arbitrary length, as for example, by means of 3D motion maps [4]. For the classification of actions, this approach is based on the implementation of a generative network to learn the 3D movement map and a discrimination network, based on the learned movements, to classify the actions. The map proposed by this model is a compact and discriminatory representation that eliminates a large amount of redundant information and is capable of capturing distinguishable trajectories around the human body. This approach has been used successfully in action videos, but the quality of the results decreases when the movement is subtle.

Among the free access tools related to this research topic, TensorFlow stands out, which is an open source library capable of offering 93% accuracy, based on neural networks [5]. TensorFlow has been used in projects such as *Deep Dream* [6], which is an image processing algorithm by means of which it is possible to create an alternative dream scene from any image or scenario. Another project of interest created from TensorFlow is *Show and Tell* [7], which is capable of automatically generating precise and highly descriptive captions for any image presented. To carry out the training of this system, millions of manually labeled images were necessary.

2.1 Metrics for Evaluation

The effectiveness of the methods proposed in the literature was evaluated with different metrics. In the case of 3D convolution, the comparison of the results obtained with the results of the analysis with fixed images in a 2D method, spatial pyramid matching (SPM) and spatiotemporal interest points (STIPs) was performed. Other standard assessment metrics based on temporal and spatiotemporal localization are:

– Probability of detection failures,
– False alarm rate,
– Accuracy (average),
– Speed.

2.2 Commercially Available Systems

Below are three commercial systems representative of the technology currently available in the market. The first, developed by the company Panasonic, is a highly sophisticated system capable of analyzing and optimizing the logistics of busy sites such as airports and companies. On the other hand, the so-called Netatmo and Butterfleye systems are low cost and oriented towards home security.

Face recognition Panasonic-FacePRO [8] is one of the most sophisticated commercial systems launched in July 2018. The system is aimed at detecting criminals and thieves in stores or companies and performs recognition and authentication by means of photographs in format jpeg. It has an alarm that notifies the guards of the presence of someone considered dangerous. Suspicious people can be tracked in the sales floor and with the information, it can be create a timeline of their route. The system is robust before:

– Aging,
– Makeup,
– Various facial expressions,
– Image quality.

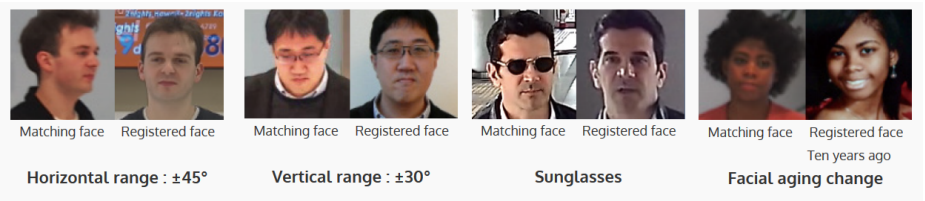

Fig. 2. Facial recognition function of Panasonic-FacePRO.

This system allows authentication even in difficult conditions for conventional technologies, such as faces at an angle greater than 45 degrees to the sides or 30 degrees up or down and partially covered with lenses. In terms of performance, the system achieved the highest level of facial recognition in the world in a comparison test (IJB-A Face Verification Challenge Performance Report/IJBA Face Identification Challenge Performance Report) of NIST. In addition, it is expected to improve performance to achieve face recognition partially covered with face masks.

This technology focuses on obtaining the best possible photograph to make the comparison with their database. The system takes several photographs and sends the best for facial recognition, reducing server and network costs. In a system with 10 or more cameras, costs can be reduced from 40% to 50% (with 10 or more cameras). Conventional systems require a large bandwidth and must be large-scale due to the large amount of storage required for the storage of all captured images. Additionally, the system allows up to 20 cameras in the network connected to a single server. The system dramatically decreases network traffic, as well as the transmission and construction costs of the network.

The algorithm combines Deep learning, machine learning and a similarity calculation method with error suppression. Deep learning technology was developed with the National University of Singapore. The camera adjusts automatically to focus on moving objects, at high speeds and at different light intensities (day and night). For its part, the facial recognition software collects the data of each face it detects and stores it in an easily accessible database. The system allows registering from 10,000 to 30,000 faces. Users can select a face and program an alarm for subsequent detections of this individual or track their movements in chronological order through all the cameras in the system. Additionally it is possible to obtain the count of people and graphic statistics by gender and age.

This system has been successfully tested at Tokyo international airport, in the area of immigration control. The system has been able to compare the photography embedded in the passport chips with a photo taken at the door of facial recognition to verify identity without the need to register travelers or take biometric data.

Netatmo Welcome [9] is an indoor camera that sends alerts to the cell phone when it detects unknown faces. Send photograph of the intruder and record video. For familiar faces it does not record video to protect privacy. The alarm does not activate with pets unless this functionality is activated to monitor them. The video can be stored on a micro SD card, in the personal Dropbox cloud or on the personal FTP server. It has a viewing angle of 130 degrees and full HD resolution. It cost is $200.00 usd. It also has an outdoor version that allows the detection and notification of the presence of people, animals and cars. Its cost is $300 usd.

Butterfleye [10] is a wireless camera with face recognition with 1080p resolution. It has thermal, motion and sound sensors. Some features are available

only with monthly plan rent. It is able to recognize family, friends, strangers and pets. It works even if the light or the internet goes out. Stores video from 5 seconds before an event occurs. The system is armed or disarmed according to GPS location. The viewing angle is 120 degrees and the battery life is 2 weeks. Its cost is $200 usd.

3 Discussion

The state of the art was reviewed and an overview of the future work that motivated this investigation was presented. Different methodologies were analyzed for video processing with Deep learning, based on different types of networks and the advantages and disadvantages of each were presented. Additionally, the different evaluation metrics and strategies for training that have been used in similar work are exposed.

In terms of commercial systems, three of the technologies with characteristics more representative of those available in the market were analyzed. Versatility, functionality and cost vary widely depending on whether the application is for home security or for companies and public sites.

References

1. Ji, S., Xu, W., Yang, M., Yu, K.: 3D convolutional neural networks for human action recognition. IEEE transactions on pattern analysis and machine intelligence, 35(1), 221–231 (2013)
2. TREC Video Retrieval Evaluation, https://trecvid.nist.gov/ (2018)
3. Lee, H., Grosse, R., Ranganath, R., Ng, A. Y.: Convolutional deep belief networks for scalable unsupervised learning of hierarchical representations. In Proceedings of the 26th annual international conference on machine learning, 609–616 (2009)
4. Sun, Y., Wu, X., Yu, W., Yu, F.: Action recognition with motion map 3D network. Neurocomputing, 297, 33–39 (2018)
5. Home TensorFlow, https://www.tensorflow.org/?hl=es (2018)
6. Deep Dream Generator, https://deepdreamgenerator.com/ (2018)
7. Deep Dream Generator, https://github.com/tensorflow/models/tree/master/... ...research/im2txt (2018)
8. Panasonic Security, https://security.panasonic.com/Face_Recognition (2018)
9. Netatmo oficial, https://www.netatmo.com/es-ES/site/ (2018)
10. Ooma Butterfleye, https://getbutterfleye.com/ (2018)