Two Stream Convolutional Neural Networks for Anomaly Detection in Surveillance Videos

Adarsh Jamadandi, Sunidhi Kottursheettar and Uma Mudenagudi

Abstract In this paper we propose a deep learning framework to identify anomalous events in surveillance videos. Anomalous events are those which do not adhere to normal behaviour. We propose to use two discriminatively trained Convolutional Neural Networks, to capture the spatial and temporal features of videos, the classification scores obtained from the two streams are later fused to assign one final score. Since our approach is scenario-based, this eliminates the need for adopting a particular definition of anomaly. We show that the Two Stream CNNs perfectly capture the intricacies involved in modelling a video data by demonstrating the framework on airport and mall surveillance datasets respectively. We achieve a final test accuracy of 99.1% for spatial stream and 91% for temporal stream for airport scenario and an accuracy of 94.7% for spatial and 90.1% for the temporal stream for the mall scenario. Our framework can be easily implemented in real-time and is capable of detecting anomaly in each frame fed by a live surveillance system.

1 Introduction

Convolutional Neural Networks have been largely successful in modelling image data and is constantly being bettered to learn exhaustive learning representation techniques for videos. Some of the key areas where Deep Learning algorithms have found its applications include Crowded Scene Understanding, Video Classification, Action Recognition, Anomaly Detection [1–3] etc. Anomaly detection mainly refers

A. Jamadandi (✉) · S. Kottursheettar
B. V. Bhoomaraddi College of Engineering and Technology, Hubli, India
e-mail: adarsh.jam@gmail.com
URL: https://adarshmj.github.io

S. Kottursheettar
e-mail: sunidhikshettar@gmail.com

U. Mudenagudi
KLE Technological University, Hubli, India
e-mail: uma@kletech.ac.in

A. Elçi et al. (eds.), *Smart Computing Paradigms: New Progresses and Challenges*,
Advances in Intelligent Systems and Computing 766,
https://doi.org/10.1007/978-981-13-9683-0_5

to detection and modelling of unusual activities or activities that do not adhere to normalcy. Detection of anomalies in surveillance videos is a non-trivial task, because the definition of "anomaly" is subjective and it's difficult to generalise this definition to every scenario at hand. It forms a pre-requisite feature for any claimed intelligent surveillance system to detect anomalous events, with almost minimal human intervention. We propose a methodology to model anomalous events in specific scenarios, for example, consider airport surveillance videos, a person entering the airport with baggage but leaving it unattended for longer periods of time could be considered an anomaly or people engaged in an angry brawl in the airport premises can be detected as an anomalous event. Thus adopting a scenario-based anomaly detection framework eliminates the hassles of assuming a generalised definition of "anomaly". In this paper we propose a scenario-based anomaly detection framework, we model the anomalous events in surveillance videos using a Two Stream Convolutional Neural Network architecture (Two Stream CNN), the two streams trained discriminatively encode the spatial and temporal aspects of the query video, we reduce the problem of anomaly detection to a binary classification problem, The classification score given by the two separately trained CNNs are later fused by taking the average of both the scores and a final classification score is obtained. Our contributions can be summarised as follows:

1. We provide a scenario-based Two Stream CNN framework to detect anomalies in surveillance videos by,
   - training an image recognition CNN architecture to model spatial features of the video.
   - training a CNN architecture to model the temporal features of the video.
   - provides an averaging technique to fuse the classification scores of both the streams to assign a final score.
2. We provide an intelligence enabled anomaly detection framework which is capable of detecting anomalies either on pre-recorded videos or video-feed from the cameras can be directly fed into our system to get classification in real-time.
3. We demonstrate our framework on two scenarios—airport and mall surveillance videos.

2 Related Work

The standard approach for modelling video data is to use an architecture that is able to effectively capture the long-term dependencies of the video. A video is basically images that evolve dynamically with time, RNNs, particularly LSTMs have been extremely successful in capturing the long term dependencies that exist in the video-data [4]. Anomaly detection in videos is also usually tackled by using the LSTMs approach [5], the idea is to model normal behaviour and check for any deviations, the errors observed from the normal behaviour is calculated and parametrised by

a regularity score, the LSTMs are used to predict the possible future behaviour by observing the video at hand, any deviations from the predictions made by the model are treated as anomaly. For example, a pedestrian walking on pedestal if modelled using the LSTM approach, the prediction usually will be that in a given later time, the pedestrian continues to walk on the pedestal, any deviation from this behaviour is calculated as an error which is later quantified by a regularity score to indicate the anomalous event.

There have been more traditional approaches that hinge on using Hand-crafted features to detect anomalies such as Cong et al. in [6] and in works proposed in [7], while such methods might prove successful in some cases, they become difficult to scale to more complex situations, because of non-ubiquitous nature of the hand-crafted features. In this paper, we have tried to exploit the idea of Two Stream CNN first introduced by Simonyan et al. [3], the rationale behind using two separate streams to model the spatial and temporal information stems from the fact that human beings tend to perceive information like shape, colour, texture and motion information through two distinct channels. This approach is extended to train two separate neural networks, one neural network is tasked to learn spatial information while the other neural network learns temporal-information, this temporal information is usually fed into, in the form of motion information. The motion information can be either optical flow or trajectory based motion-information, as discussed in [3]. The classification score obtained from both the streams are combined using different fusion techniques [8].

3 Methodology

In this section of the paper we discuss in depth the dataset chosen, the CNN architecture employed for training and testing, the scenario considered for testing and training, the various preprocessing steps involved in the training, the implementation details and evaluation of the results.

3.1 Dataset

Since the framework developed by us hinges on contextual based anomaly detection, we considered two different scenarios—airport and mall, the airport and mall surveillance datasets were acquired from Youtube. The datasets used for training and testing were challenging because most of the videos available on Youtube are either captured from very low-end devices by the bystanders present in and around the situation, which have lot of jittering and adverse lighting conditions or the videos could be a result of the surveillance systems installed at that particular location. There were totally 22 videos of airport scenario which were used for training and

testing, each of varying length with events covering from passenger baggage theft, shootouts, brawls, bomb blasts etc. The mall surveillance dataset had 10 videos for training and testing with again a variety of situations.

3.2 *Two Stream Architecture for Anomaly Detection*

An overview of the Two Stream CNN architecture is provided in Fig. 1, a two stream architecture mainly involves training two different streams which capture complementary information. In case of video data, it is essential to learn the temporal dependencies so that the relationship between the consequent frames is established. Thus we have two streams—a spatial-stream and a temporal-stream that encode complementary information about the video. The spatial stream is basically a state-of-the-art CNN architecture which is trained for image-classification. We use the famous Inception V3 model which was used for the ImageNet Large Visual Recognition Challenge (2012). The architecture was trained for over a 1000 different classes, and reports a top-5 error percent of 3.46 [9]. Since our major motive is to deploy the solution in real-time, we decided to use a technique called Transfer-Learning described in the work [10], the idea is to use a well-trained model architecture like Inception v3, chop off the final layers and retrain them for new categories. Since the architecture already has been trained on vast amount of image data, this type of training provides us with quicker results and can be easily trained on systems that dont incorporate GPUs or have very limited training resources. The spatial stream was trained with RGB frames sampled at 30 fps from the video clips. The temporal stream was trained with the motion information across the video frames. To capture the motion information we performed the dense optical flow and saved them as grey optical images. Saving them as images, allowed us to once again train the Inception V3 architecture for new categories. Each stream gives a separate classification score for a given query

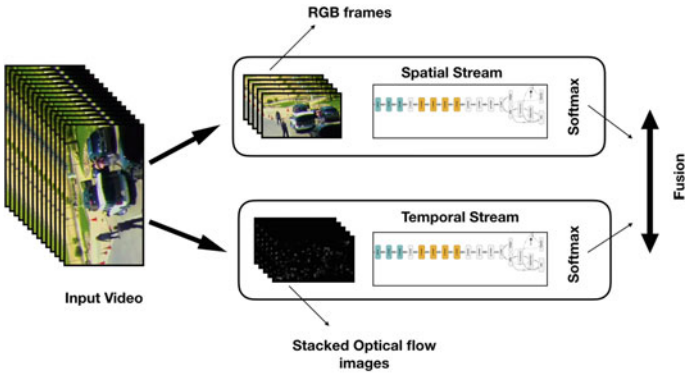

Fig. 1 An overview of the proposed two stream architecture for anomaly detection

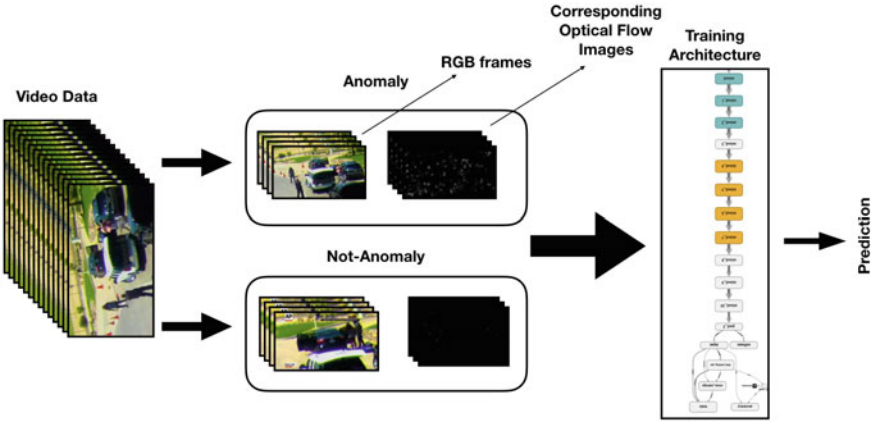

Fig. 2 The frames in the form of RGB images and their corresponding optical flow images are separately trained on a different CNN architecture to obtain two independent classification scores

frame which will later be combined using different fusion techniques [8]. In our case, we have considered simple averaging technique, wherein we fuse the classification scores from the two streams by taking a mean of those individual values to get a final classification score.

3.3 Implementation Details

The videos are trimmed to separate out the anomaly events and the preceding events which could be termed as normal behaviour. These video clips are converted to frames using the FFMPEG tool and are stored in two separate folders—anomaly and not-anomaly. The RGB frames will serve as input for the spatial stream for training. To train the temporal-stream, we consider the Farneback dense optical flow method [11], the video clips which contained the anomaly are subjected to optical flow, so that we have RGB frames and their corresponding motion information via the optical flow images. The optical flow images are saved as grey images with the white content indicating the motion and black, a lack of. The software used for training was TensorFlow, all the training and testing was done on a local machine, with configuration—8GB RAM and i5 processor with no dedicated GPU support. The implementation workflow is depicted in Fig. 2.

4 Results and Discussions

In this section we discuss the different scenarios considered for training and testing, their respective test-train accuracies. We discuss about the fusion technique used to combine the classification score of both the streams. Tables 1 and 2 shown below, summarises the train-test accuracy for the airport and mall surveillance videos.We have considered 80% of the images for main training, 10% of the images for validation and the final 10% of the images are kept for final prediction. We achieve a final test accuracy of 99.1 and 91% for airport scenario and a final testing accuracy of 94.7 and 90.1% for the mall scenario. From the tables we can infer that as the number of training images for the temporal stream has decreased, correspondingly we see a drop in the accuracy. This is compensated by the fusion technique, which helps us to arrive at a final score by combining the outputs of both the stream instead of basing our classification score on just one stream. Many different fusion strategies have been adopted to combine the scores of both the streams, literature [12] talks about different fusion techniques, in the context of action recognition, in our framework we have adopted a simple averaging technique that involves getting the final prediction score of both the streams for both the categories—anomaly and not-anomaly. We take the classification score of both the streams for both the categories and compute the mean, that is, the mean of the anomaly score from both the streams is computed, similarly the mean of the not-anomaly score is also computed, the final resultant score is assigned as a final classification score for the query frame. This method is demonstrated in Fig. 3, where we have a query frame with its corresponding RGB and Optical flow image, the Two stream CNN gives a classification score for the frame under consideration, later we compute the mean and assign a final classification score as to whether the given frame is anomalous or not-anomalous. In Fig. 3, the first image is the RGB image and the second image is its corresponding optical flow image. The classification score shown in the image, is the final score obtained after averaging the scores from both the streams.

Table 1 Scenario: airport surveillance videos

No. of training images	Train accuracy (%)	Cross-validation accuracy (%)	Final test accuracy (%)
3,44,128 (RGB)	98	98	99.1
3532 (Optical flow images)	91	89	91

Table 2 Scenario: mall surveillance videos

No. of training images	Train accuracy (%)	Cross-validation accuracy (%)	Final test accuracy (%)
61,946 (RGB)	96	93	94.7
3320 (Optical flow images)	94	87	90.1

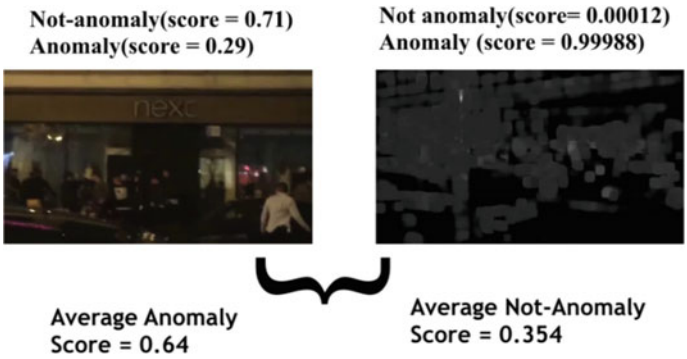

Fig. 3 The classification scores obtained for the RGB image and the optical flow image of a query frame are fused by taking averages, and a final score is assigned for the frame

5 Conclusion

In this paper, a Two Stream CNN for anomaly detection in surveillance videos is proposed. The proposed model exploits the spatial and temporal components of the video to provide effective classification of events as anomalous or not anomalous events. We have furthered this approach by creating a framework that is capable of working on live video feeds and classify each frame in real-time on the fly. This framework can further be improved by training an object classification architecture to localise anomalous events by annotating the frames, thus not only classifying a given frame as anomalous or not, but also localising where the anomaly is happening. The localising mechanism can also be based on a Reinforcement learning architecture, where the policy function helps in focusing on the relevant parts of an image.

References

1. Li, T., Chang, H., Wang, M., Ni, B., Hong, R., Yan, S.: Crowded scene analysis: a survey. arXiv:1502.01812v1 [cs.CV]. Last accessed 6 Feb 2015
2. Karpathy, A., Toderici, G., Shetty, S., Leung, T., Sukthankar, R., Fei-Fei, L.: Large-scale video classification with convolutional neural networks. In: CVPR (2014)
3. Simonyan, K., Zisserman, A.: Two-stream convolutional networks for action recognition in videos. In: Proceedings of the 27th International Conference on Neural Information Processing Systems, vol. 1 (2014)
4. Srivastava, N., Mansimov, E., Salakhudinov, R.: Unsupervised learning of video representations using LSTMs. arXiv:1502.04681v3 [cs.LG]. Last accessed 4 Jan 2016
5. Medel, J.R., Savakis, A.: Anomaly detection in video using predictive convolutional long short-term memory networks. arXiv:1612.00390; Cong, Y., Yuan, J., Liu, J.: Sparse reconstruction cost for abnormal event detection. In: CVPR, pp. 3449–3456 (2011)
6. Cong, Y., Yuan, J., Liu, J.: Sparse reconstruction cost for abnormal event detection. In: CVPR, pp. 3449–3456 (2011)

7. Mahadevan, V., Li, W., Bhalodia, V., Vasconcelos, N.: Anomaly detection in crowded scenes. In: Proceedings of IEEE Conference on Computer Vision and Pattern Recognition (2010)
8. Feichtenhofer, C., Pinz, A., Zisserman, A.: Convolutional two-stream network fusion for video action recognition. arXiv:1604.06573v2 [cs.CV]. Last accessed 26 Sep 2016
9. Szegedy, C., Vanhoucke, V., Ioffe, S., Shlens, J., Wojna, Z.: Rethinking the inception architecture for computer vision. arXiv:1512.00567 [cs.CV]
10. Donahue, J., Jia, Y., Vinyals, O., Hoffman, J., Zhang, N., Tzeng, E., Darrell, T.: DeCAF: a deep convolutional activation feature for generic visual recognition. arXiv:1310.1531v1 [cs.CV]. Last Accessed 6 Oct 2013
11. Farneback, G.: Two-frame motion estimation based on polynomial expansion. In: Proceedings of the 13th Scandinavian Conference on Image Analysis (2003); Liu, W., Anguelov, D., Erhan, D., Szegedy, C., Reed, S., Fu, C.Y., Berg, A.C.: SSD: single shot multibox detector. arXiv:1512.02325 [cs.CV]
12. Lin, K., Chen, S.-C., Chen, C.-S., Lin, D.-T., Hung, Y.-P.: Abandoned object detection via temporal consistency modelling and back-tracing verification for visual surveillance. IEEE Trans. Inf. Forensic Secur. (TIFS) (2015)