A deep unified framework for suspicious action recognition

Amine Ilidrissi[1] and Joo Kooi Tan[1]

[1]Kyushu Institute of Technology, Japan

ilidrissiamine@gmail.com, etheltan@cntl.kyutech.ac.jp

Abstract: As action recognition undergoes change as a field under influence of the recent deep learning trend, and while research in areas such as background subtraction, object segmentation and action classification is steadily progressing, experiments devoted to evaluate a combination of the aforementioned fields, be it from a speed or a performance perspective, are far and few between. In this paper, we propose a deep, unified framework targeted towards suspicious action recognition that takes advantage of recent discoveries, fully leverages the power of convolutional neural networks and strikes a balance between speed and accuracy not accounted for in most research. We carry out performance evaluation on the KTH dataset and attain a 95.4 percent accuracy in 200 milliseconds computational time, which compares favorably to other state-of-the art methods. We also apply our framework to a video surveillance dataset and obtain 91.9 percent accuracy for suspicious actions in 205 milliseconds computational time.

Keywords: suspicious action recognition, deep learning, convolutional neural networks, background subtraction, optical flow estimation, action classification

1 INTRODUCTION

The past few years have witnessed significant progress in the field of action recognition, with new research in areas such as background subtraction, object segmentation and action classification coming out at an increasing pace, and more and more applications to automatic video classification and video surveillance being found. Moreover, the recent deep learning trend has brought about a data-based change in how these areas are approached, with a focus on large datasets and generalization to all kinds of scenes.

However, while deep learning has been and is being applied to an increasing number of fields, not many attempts have been made at combining these efforts into a unified framework, and practically no evaluations have been recently carried out to measure both speed and performance of commonly-used techniques. In this paper, we explore the current state-of-the-art computer vision techniques, focusing on optical flow estimation, background subtraction and action classification, propose a unified action recognition method that takes advantage of all these advances and apply it to abnormal scenes.

The paper is structured as follows. In Section 2, we describe prior research on related fields and outline methods relevant to our research. In Section 3, we delve into the framework's implementation and provide a step-by-step explanation from video input to final action classification. Finally, Section 4 is devoted to experimenting, complete with background regarding choice of training and test datasets, as well as interpretation of results.

2 RELATED WORK

2.1 Background subtraction

Depending on performance needs, one might simply use frame differencing, or a dynamic method [1] [2] to estimate the background of a given scene. These methods are usually used on a whole frame or two frames at a time and may require prior knowledge of the processed scene. One technique which does not follow that paradigm is deep background subtraction [3], which operates by dividing a frame into small pixel-centered patches and classifying these as background or foreground patches. It should be noted that focusing on these pixel-centered patches instead of the whole scene does not hamper performance. Also, contrary to expectations, using only a small number of frames (25 to 50) suffice to obtain very good to excellent results and allows for camera stabilization and shadow removal among other benefits.

2.2 Optical flow

Optical flow estimation has traditionally been done using differential methods, be they sparse like the Lucas-Kanade method [4], or dense like the Farnebäck method [5] and the TV-L1 method [6]. While speed and accuracy vary wildly across these kinds of methods, a major drawback common to all of them is lack of generalization to large data. One breakthrough in this domain is FlowNet [7], an optical flow estimation method that uses convolutional neural networks to take advantage of large datasets. FlowNet 2.0 [8], an improved iteration standing as the current state-of-the-art, has also been released, as well as a dataset aimed at stereo

optical flow estimation [9]. It is worth pointing out that while datasets used for training purposes are artificial (with some created with 3D modeling software), the resulting models generalize surprisingly well to real-world data.

2.3 Action classification

Action classification relies mostly on extracting features. These can either be handcrafted like Motion History Images [10], Motion History Volumes [11] or Directional Motion History Images [12] [13] [14] and processed using SVMs for example, or automatically computed using neural networks such as two-stream convolutional networks [15], which combine both static features and optical-flow-powered dynamic features to achieve great video recognition performance. Temporal segment networks [16] expand on that concept in an original way, by operating on small snippets instead of a large, possibly unrepresentative part of the video. This technique leads to state-of-the-art results on various datasets and changes the way we are thinking about how to handle video inputs to a convolutional neural network.

3 THE FRAMEWORK

The present research will be focusing on action recognition using deep background subtraction, deep optical flow estimation, and two-stream convolutional networks.

3.1 Deep background subtraction

We use the method described in [3] to achieve efficient background subtraction with the use of convolutional neural networks. More specifically (assuming we are inputting a grayscale video):

1. We construct a simple background model of the video input by computing a temporal average of each pixel.
2. We generate for each input frame a 3-channel frame, where the first channel is the untouched input, the second channel is the background model, and the third channel is left empty.
3. We extract for each pixel of the generated frame a square patch centered around that pixel, and we feed it to the neural network.
4. The neural network, if training, learns from the input patches, given pixel-precise ground truth; if predicting, it classifies each patch as either background or foreground.
5. We generate the foreground video based on the above classification results and pass it on to the optical flow estimation part of the framework.

Results of the above steps on sample frames can be seen in **Fig. 1.** The background model, while retaining faint traces of motion, is sufficiently accurate for an uncluttered scene. The 3-channel frame shows the static background as brown, and motion (including removed shadows) as red.

3.2 Optical flow estimation

We rely on FlowNet 2.0 [8] for deep optical flow estimation. A variety of pre-trained models were made available by the University of Freiburg (https://github.com/lmb-freiburg/flownet2), all differing in speed and accuracy. We aim for maximum accuracy and choose the “FlowNet2” model. The resulting output is usable as-is. However, the raw optical flow for a single frame pair is over 100 kilobytes, which is inconvenient when training the action classifier since GPU memory is quite limited.

To work around this, we use post-estimation optical flow compression inspired by dense_flow (https://github.com/yjxiong/dense_flow) to push file size down to less than 2 kilobytes. Compression here means saving optical flow information to two grayscale images, each representing an axis, and computed using Equation (1), where x is the estimated optical flow value for the working axis, $f(x)$ the computed image, and α a parameter acting as an optical flow bound.

$$f(x) = \begin{cases} 0 & x < -\alpha \\ 255 * \frac{x+\alpha}{2\alpha} & |x| < \alpha \\ 255 & x > \alpha \end{cases} \quad (1)$$

Example results can be seen in **Fig. 2.**, with the images to the left representing the horizontal axis, and the images to the right the vertical axis. We can observe significant improvements when optical flow estimation is preceded by deep background subtraction.

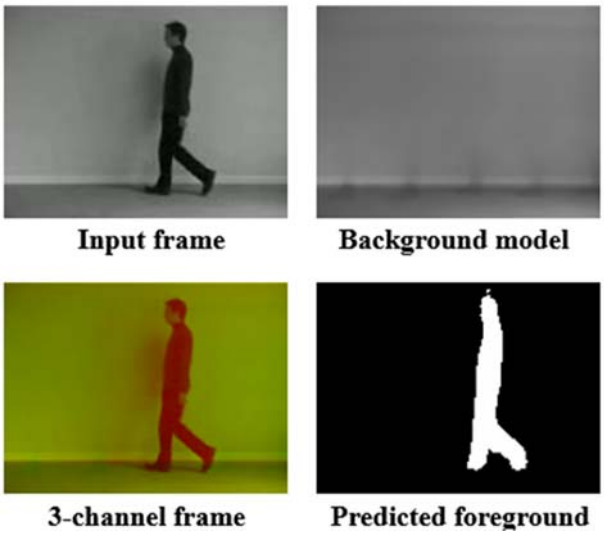

Fig. 1. Deep background subtraction on a sample frame

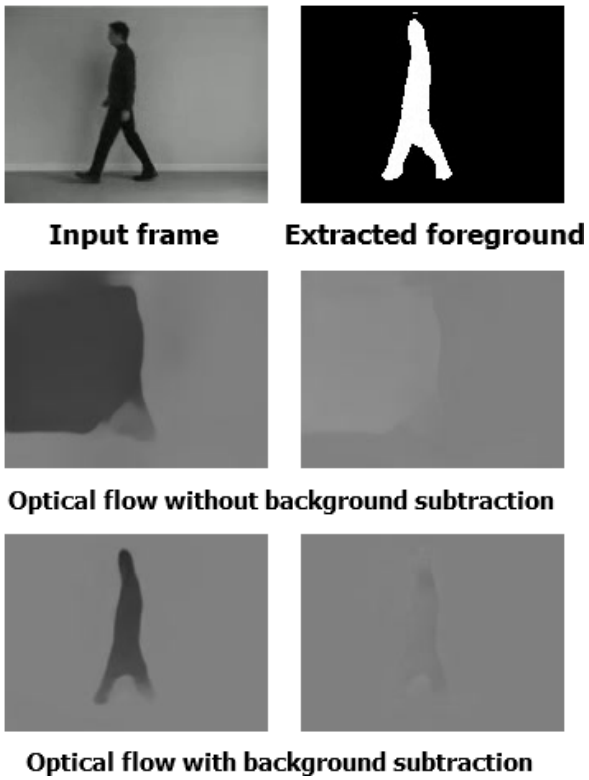

Fig. 2. Sample optical flow estimation (left optical flow images represent the horizontal axis, right optical flow images represent the vertical axis)

3.3 Action classification

While action classification can be performed using a variety of more or less deep methods, in this paper, we adopt two-stream convolutional networks as a simple yet modular way to accurately recognize human actions. To summarize the method described in [15] and [17]:

1. We alternately stack a number of optical flow frame pairs (representing both axes) and feed them into a convolutional network for training or classification.
2. We augment the optical flow input before the first layer for improved training and classification.
3. We begin from step 1 again for a number of times (and choosing different optical flow frames) and average the obtained training losses or predictions.

4 EXPERIMENTS

This section is devoted to testing our action recognition framework using multiple sets of networks and parameters. Subsection 1 will describe the hardware and software environment we worked under. Subsection 2 will be about the datasets used as well as training and test details. Subsection 3 is the results' section, with computational time required and comparisons to state-of-the-art methods are included.

4.1 Working environment

Our hardware working environment is as follows:

- Processor: Intel Core i7-6950 @ 3GHz, 10 cores, 20 logical processors
- Physical memory: 128 GB
- GPU: NVIDIA GeForce GTX 1070, 1920 CUDA cores, 8 GB memory

As for software, we use Ubuntu 16.04 as our operating system, as well as stock and custom versions of Caffe [18].

4.2 Datasets, training and test details

For training and predicting in the case of deep background subtraction, we use the same network as in [3], which is LeNet-5 [19] slightly modified to get better results and trained on the 2014 motion detection database available at ChangeDetection.net [20]. More specifically, we use the shadow detection dataset and focus on shadow removal as it is most relevant to our human action dataset. We use 48 frames for training and 6 frames for testing. Other motions, such as camera jitter and zooming, will be the object of further research. We let the training run for 100,000 iterations, with a batch size of 20,000 and an initial learning rate of 0.01 that decreases asymptotically.

As for action classification, we use the VGG-19 network [21] and the ResNet-18/ResNet-101 networks [22]. We let training run for 90,000 iterations (differing from [17] as we only use one GPU), with an initial learning rate of 0.005 that is divided by 10 each 30,000 iterations. We use a batch size of 50 for VGG-19, 88 for ResNet-18 and 15 for ResNet-101.

We first apply the framework to the KTH dataset [23] and choose the VGG-19 and ResNet-18 networks. We split the dataset into a training set comprising two-thirds of the videos, and a test set comprising one-third of the videos. It is worth noting that while comparisons to state-of-the-art methods will be made later in the paper, the way the KTH dataset is split can lead to up to a 10% difference in accuracy [24], which can make comparisons less reliable that they seem.

We then carry out a second experiment on a video surveillance dataset used in [25], where a camera installed at the Grand Central Station of New York captured an hour-long video at 1 frame per second that has been subsequently annotated with trajectories of 12,684 pedestrians. Example trajectories can be found in **Fig. 3** and **Fig. 4**.

Fig. 3. Example of a suspicious trajectory

Fig. 4. Example of an ordinary trajectory

We manually annotated each trajectory as either suspicious (940 trajectories) when the pedestrian seemed to follow a non-linear or premeditated path, or ordinary (11,744 trajectories) when the pedestrian did not seem to follow any kind of structured path. To extract the region of interest for each trajectory, we first applied deep background subtraction to each relevant frame, then cropped the frame to a 20x60 patch centered around the spatial coordinates of the person of interest. We choose the ResNet-101 network for this experiment to allow for better generalization.

Finally, the output of the optical flow estimation method we used depends on the value of the bound parameter. For the KTH dataset, experimenting led us to choose a value of 5, striking a balance between small and large motions, a value we kept for the video surveillance dataset as the movements were uniform enough that the parameter value did not affect results.

4.3 Results

Results of the experiment can be found in **Table 1** and **Table 2**. We also include computational time required per frame in **Table 3**, as well as accuracy comparisons to state-of-the-art methods for the KTH dataset in **Table 4**.

While the method we propose is not suitable for real-time action recognition (5 frames per second against typically 30 or 60 frames per second required), we believe it could be used for purposes such as video surveillance and suspicious action recognition, where a few seconds' delay is usually acceptable.

Table 1. Results of the experiment on the KTH dataset

Action	**VGG-19 network**	**ResNet-18 network**
Walking	97.9%	99.3%
Boxing	100%	96.5%
Running	86.7%	86.1%
Jogging	95.5%	91%
Handclapping	100%	99.3%
Handwaving	92.4%	93.7%
Total	**95.4%**	**94.3%**

Table 2. Results of the experiment for the video surveillance dataset (confusion matrix)

Real/predicted behavior	**Ordinary**	**Suspicious**
Ordinary	80.3%	19.7%
Suspicious	8.1%	91.9%

Table 3. Framework computational time

Operation	**Time required**
Background subtraction	75 milliseconds
Optical flow estimation	30 milliseconds
Action classification (VGG-19)	95 milliseconds
Action classification (ResNet-18)	75 milliseconds
Action classification (ResNet-101)	100 milliseconds
Total (VGG-19)	**200 milliseconds**
Total (ResNet-18)	**180 milliseconds**
Total (ResNet-101)	**205 milliseconds**

Table 4. Comparison to state-of-the-art methods (KTH)

Method	**Accuracy**
Schuldt et al. [23]	71.7%
Han et al. [26]	93.1%
Kim et al. [27]	95.3%
Ahsan et al. [12]	95.6%
Our method (VGG-19)	**95.4%**
Our method (ResNet-18)	**94.3%**

5 CONCLUSION

In this paper, we proposed a unified action recognition framework combining background subtraction, optical flow estimation and action classification all in a single ready-to-use solution. We tested our method using the KTH and ChangeDetection.net datasets, and the LeNet-5, VGG-19 and ResNet-18 networks. We attained a 95.4 percent accuracy in 200 milliseconds computational time, which compares favorably to other state-of-the art methods. We also carried out a second experiment on the video surveillance dataset in [25] and obtained 91.9 percent accuracy for suspicious actions.

Various topics are open to future research. Using a larger dataset for deep background subtraction would help in detecting larger and wider motion changes. Relying on different models for optical flow estimation would enable one to choose more efficiently between maximizing either speed or accuracy. Finally, using larger human action datasets such as UCF101 [28], HMDB-51 [29] and Youtube-8M [30], as well as experimenting with temporal segment networks [16] and other convolutional networks may lead to better action classification accuracy.

Acknowledgements This research was supported by JSPS Kakenhi, Grant number 16K01554.

REFERENCES

[1] Baranwal M, Khan MT, De Silva CW (2011), Abnormal motion detection in real time using video surveillance and body sensors. International Journal of Information Acquisition, pp. 103-116

[2] Setyawan FXA, Tan JK, Kim H, et al (2017), Moving objects detection employing iterative update of the background. Artificial Life and Robotics, Vol. 22, No. 2, pp. 168-174

[3] Braham M, Van Droogenbroeck M (2016), Deep background subtraction with scene-specific convolutional neural networks. International Conference on Systems, Signals and Image Processing, pp. 1-4

[4] Lucas BD, Kanade T (1981), An iterative image regis tration technique with an application to stereo vision. Proce edings of Imaging Understanding Workshop, pp. 121-130

[5] Farnebäck G (2003), Two-frame motion estimation based on polynomial expansion. Image analysis, pp. 363-370

[6] Zach C, Pock T, Bischof H (2007), A duality based approach for realtime TV-L1 optical flow. Joint Pattern Recognition Symposium, pp.214-223

[7] Dosovitskiy A, Fischer P, Ilg E, et al (2015), Flownet: Learning optical flow with convolutional networks. Proceedings of the IEEE International Conference on Computer Vision, pp. 2758-2766

[8] Ilg E, Mayer N, Saikia T, et al (2017), FlowNet 2.0: Evolution of Optical Flow Estimation with Deep Networks. IEEE Conference on Computer Vision and Pattern Recognition, pp. 2462-2470

[9] Mayer N, Ilg E, Hausser P, et al (2016), A large dataset to train convolutional networks for disparity, optical flow, and scene flow estimation. Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 4040-4048

[10] Bobick AF, Davis JW (2001), The recognition of human movement using temporal templates. IEEE Transactions on pattern analysis and machine intelligence, Vol. 23, No.3, pp. 257-267

[11] Canton-Ferrer C, Casas JR, Pardas M (2006), Human model and motion based 3D action recognition in multiple view scenarios. 14th European Signal Processing Conference, pp. 1-5

[12] Ahsan SMM, Tan JK, Kim H, et al (2015), Human action representation and recognition: An approach to a histogram of spatiotemporal templates. International Journal of Innovative Computing, Information and Control, Vol. 11, No. 6, pp. 1855-1868

[13] Ahad MAR, Ogata T, Tan JK, et al (2008), A complex motion recognition technique employing directional motion templates. International Journal of Innovative Computing, Information and Control, Vol. 4, No.8, pp. 1943-1954

[14] Ahsan SMM, Tan JK, Kim H, et al (2016), Spatiotemporal LBP and shape feature for human activity representation and recognition. International Journal of Innovative Computing, Information and Control, Vol. 12, No. 1, pp. 1-13

[15] Simonyan K, Zisserman A (2014), Two-stream convolutional networks for action recognition in videos. Advances in neural information processing systems, pp. 568-576

[16] Wang L, Xiong Y, Wang Z, et al (2016), Temporal segment networks: Towards good practices for deep action recognition. European Conference on Computer Vision, pp. 20-36

[17] Wang L, Xiong Y, Wang Z, et al (2015), Towards Good Practices for Very Deep Two-Stream ConvNets. (arXiv)

[18] Jia Y, Shelhamer E, Donahue J, et al (2014), Caffe: Convolutional architecture for fast feature embedding. Proceedings of the 22nd ACM international conference on Multimedia, pp.675-678

[19] LeCun Y, Bottou L, Bengio Y, et al (1998), Gradient-based learning applied to document recognition. Proceedings of the IEEE, vol. 86, pp. 2278–2324

[20] Wang Y, Jodoin PM, Porikli F, et al (2014), CDnet 2014: An Expanded Change Detection Benchmark Dataset. Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops, pp. 387-394

[21] Simonyan K, Zisserman A (2014), Very deep convolutional networks for large-scale image recognition. (arXiv)

[22] He K, Zhang X, Ren S, et al (2016), Deep residual learning for image recognition. Proceedings of the IEEE conference on computer vision and pattern recognition, pp. 770-778

[23] Schuldt C, Laptev I, Caputo B (2004), Recognizing human actions: a local SVM approach. Proceedings of the 17th International Conference on Pattern Recognition, pp. 32-36

[24] Gao Z, Chen MY, Hauptmann A, et al (2010), Comparing evaluation protocols on the KTH dataset. Human Behavior Understanding, pp. 88-100

[25] Yi S, Li H, Wang X (2015), Understanding pedestrian behaviors from stationary crowd groups. IEEE Conference on Computer Vision and Pattern Recognition, pp. 3488-3496

[26] Han Y, Zhang P, Zhuo T, et al (2017), Going deeper with two-stream ConvNets for action recognition in video surveillance. Pattern Recognition Letters

[27] Kim TK, Wong SF, Cipolla R (2007), Tensor canonical correlation analysis for action classification. Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 1-8

[28] Soomro K, Zamir AR, Shah M (2012), UCF101: A Dataset of 101 Human Action Classes From Videos in The Wild.

[29] Kuehne H, Jhuang H, Garrote E, et al (2011), HMDB: A Large Video Database for Human Motion Recognition. IEEE International Conference on Computer Vision, pp. 2556-2563

[30] Abu-El-Haija S, Kothari N, Lee J, et al (2016), Youtube-8M: A Large-Scale Video Classification Benchmark. (arXiv)