Classification of Human Actions in Videos with a Large-Scale Photonic Reservoir Computer

Piotr Antonik[1,2(✉)], Nicolas Marsal[1,2], Daniel Brunner[3], and Damien Rontani[1,2]

[1] LMOPS EA 4423, CentraleSupélec, 57070 Metz, France
{piotr.antonik,damien.rontani}@centralesupelec.fr
[2] Université de Lorraine, CentraleSupélec, LMOPS, 57000 Metz, France
[3] FEMTO-ST Institute/Optics Department,
CNRS & University Bourgogne Franche-Comté, 25030 Besançon, France

Abstract. The identification of different types of human actions in videos is a major computer vision task, with capital applications in e.g. surveillance, control, and analysis. Deep learning achieved remarkable results, but remains hard to train in practice. Here, we propose a photonic reservoir computer for recognition of video-based human actions. Our experiment comprises off-the-shelf components and implements an easy-to-train neural network, scalable up to 16,384 nodes, and performing with a near state-of-the-art accuracy. Our findings pave the way towards photonic information processing systems for real-time video processing.

Keywords: Photonic reservoir computing · Computer vision · Human action classification

1 Introduction

The recognition of human actions has recently become one of the most popular research areas in the field of computer vision [11], driven by the potential applications in various areas such as surveillance, control, and analysis [7]. The complexity of the task stems from the numerous problems, such as background clutter, partial occlusion, different scales, viewpoints, lighting, and appearance [10]. Deep learning has achieved remarkable results in this field [11], but presents several complications, such as the need for (very) large training datasets, the non-trivial tuning of the hyperparameters, and time- and energy-consuming training process, which commonly requires dedicated high-end hardware (GPU).

In this work, we propose a photonic neural network for video processing. Optical computing promise a high level of parallelism in e.g. optical communications. Therefore, neural networks could heavily benefit from parallel signal

P. Antonik—Supported by AFOSR (grants No. FA-9550-15-1-0279 and FA-9550-17-1-0072), Région Grand-Est, and the Volkswagen Foundation via the NeuroQNet.

I. V. Tetko et al. (Eds.): ICANN 2019, LNCS 11731, pp. 156–160, 2019.
https://doi.org/10.1007/978-3-030-30493-5_15

transmission, which is one of the strong suits of photonics. Our optical approach could thus allow one to build high-speed and energy-efficient photonic computing devices.

Our experimental system implements the reservoir computing paradigm – a set of methods for designing and training artificial recurrent neural networks [4,6]. A typical reservoir consists of randomly connected fixed network with random input coupling coefficients. Only the output weights are optimised, which reduces the training process to solving a system of linear equations [5]. The RC algorithm has been successfully applied to channel equalisation, chaotic series forecasting, and phoneme recognition. Its simplicity makes it well suited for electronic, opto-electronic, and all-optical analogue implementations (see [8] for a recent review).

2 Results

The system is tested on the well-known KTH database [9], consisting of video recordings of 25 subjects performing 6 different motions (walking, jogging, running, boxing, hand waving, and hand clapping). At the preprocessing stage, we use the histograms of oriented gradients (HOG) algorithm [2] to preprocess individual video frames and extract relevant spatial and shape information. The photonic reservoir computer classifies the 6 motions given the resulting HOG features, as illustrated in Fig. 1.

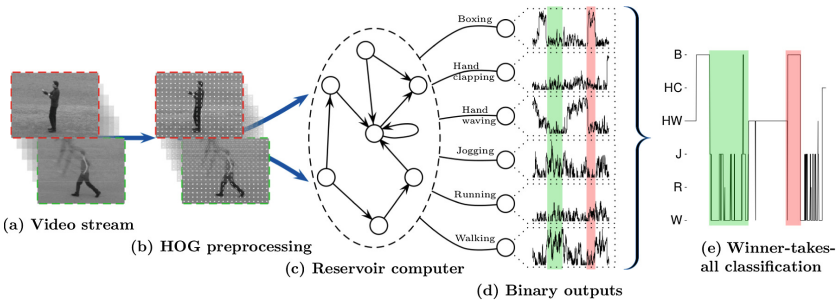

Fig. 1. Scheme of principle of the proposed video processing system. The video input stream (a) undergoes a preprocessing stage (b), where the HOG algorithm is applied to each individual frame. The resulting features are fed into the reservoir computer (c), trained to classify each individual frame. This is achieved by defining 6 binary output nodes (d), one for each action class, that are trained to output 1 for a frame of the corresponding class and 0 for the others. Target outputs are shown in blue. The frame-wise classification (e) is obtained by selecting the node with the maximum output, i.e. the winner-takes-all approach. The final decision for a video sequence is given by the class attributed to the most frames of the sequence. (Color figure online)

The proposed experimental setup, inspired by [1,3] and schematised in Fig. 2, is based on the phase modulation of a spatially extended planar wave by means of a spatial light modulator (SLM). The scheme's notable parallelisation potential allows to implement large neural networks, which is vital for successfully solving challenging tasks in computer vision. The proposed experimental setup can realise a reservoir of 16,384 nodes and could, in principle, be scaled up to as high as 262,144 neurons. The input, the recurrence of the network, and the output layer are implemented digitally to increase the flexibility of the setup. The digital part also dictates the processing speed and allows to classify 2 video frames per second with large reservoirs (16,384 nodes) and up to 7 frames per second with small reservoirs (1,024 nodes).

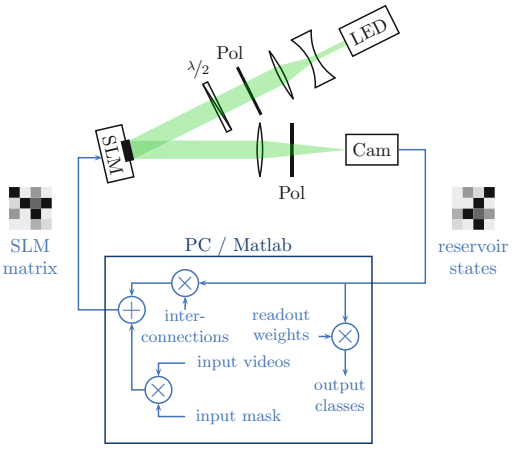

Fig. 2. Scheme of the experimental setup. The output of a green LED (532 nm) is collimated, expanded, polarised (Pol), and used to illuminate the surface of the spatial light modulator (SLM). The latter is imaged by a high-speed camera (Cam.) through a second polariser (Pol) and an imaging lens. Both the camera and the SLM are controlled by a computer, running a Matlab script. The latter generates the inputs from the input videos, and computes the values of pixels to be loaded on the SLM (SLM matrix). The computer uses the data from the camera to extract the reservoir states, compute the outputs and generate the output classes. (Color figure online)

We investigate different network sizes from 1,024 to 16,384 nodes, both numerically and experimentally, and obtain a classification accuracy as high as 92% (see Fig. 3), comparable to the state-of-the-art rates 90.7%–95.6% achieved with far more complex and demanding architectures implemented on noiseless digital processors [11]. This work thus shows that a challenging computer vision task can be efficiently solved with a photonic reservoir computer.

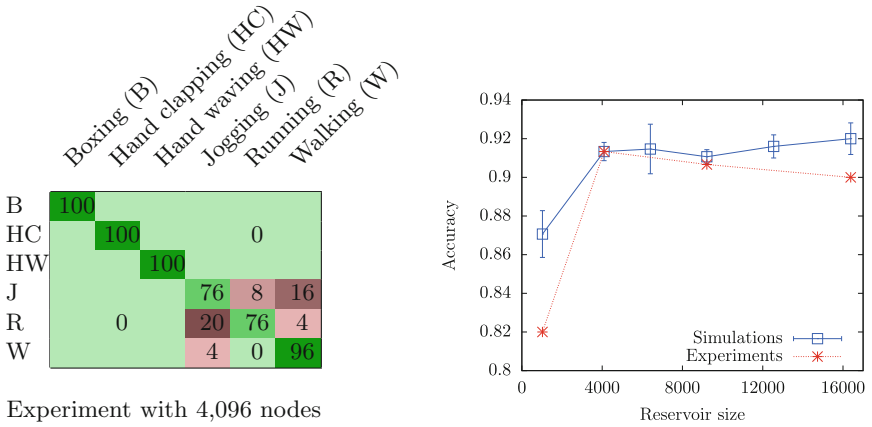

Fig. 3. (a) Experimental confusion matrix, giving the percentage of actions of class i classified into the class j. Hand gestures (B, HC, HW) are perfectly recognised, while fast spatial motions (J, R, W) are more challenging to differentiate. **(b)** Performance of our photonic reservoir computer on the human action classification task. The error bars on the numerical results show the variability of the accuracy (standard deviation) with 5 different input masks. Experimental variability could not be measured because of the long experimental runtime.

References

1. Bueno, J., et al.: Reinforcement learning in a large-scale photonic recurrent neural network. Optica **5**(6), 756 (2018). https://doi.org/10.1364/optica.5.000756
2. Dalal, N., Triggs, B.: Histograms of oriented gradients for human detection. In: 2005 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR). IEEE (2005). https://doi.org/10.1109/cvpr.2005.177
3. Hagerstrom, A.M., Murphy, T.E., Roy, R., Hövel, P., Omelchenko, I., Schöll, E.: Experimental observation of chimeras in coupled-map lattices. Nat. Phys. **8**(9), 658–661 (2012). https://doi.org/10.1038/nphys2372
4. Jaeger, H.: Harnessing nonlinearity: predicting chaotic systems and saving energy in wireless communication. Science **304**(5667), 78–80 (2004). https://doi.org/10.1126/science.1091277
5. Lukoševičius, M., Jaeger, H.: Reservoir computing approaches to recurrent neural network training. Comput. Sci. Rev. **3**(3), 127–149 (2009). https://doi.org/10.1016/j.cosrev.2009.03.005
6. Maass, W., Natschläger, T., Markram, H.: Real-time computing without stable states: a new framework for neural computation based on perturbations. Neural Comput. **14**(11), 2531–2560 (2002). https://doi.org/10.1162/089976602760407955
7. Moeslund, T.B., Granum, E.: A survey of computer vision-based human motion capture. Comput. Vis. Image Underst. **81**(3), 231–268 (2001). https://doi.org/10.1006/cviu.2000.0897
8. der Sande, G.V., Brunner, D., Soriano, M.C.: Advances in photonic reservoir computing. Nanophotonics **6**, 3 (2017). https://doi.org/10.1515/nanoph-2016-0132

9. Schuldt, C., Laptev, I., Caputo, B.: Recognizing human actions: a local SVM approach. In: Proceedings of the 17th International Conference on Pattern Recognition 2004, ICPR 2004. IEEE (2004). https://doi.org/10.1109/icpr.2004.1334462
10. Vrigkas, M., Nikou, C., Kakadiaris, I.A.: A review of human activity recognition methods. Front. Robot. AI **2** (2015). https://doi.org/10.3389/frobt.2015.00028
11. Wu, D., Sharma, N., Blumenstein, M.: Recent advances in video-based human action recognition using deep learning: a review. In: 2017 International Joint Conference on Neural Networks (IJCNN). IEEE, May 2017. https://doi.org/10.1109/ijcnn.2017.7966210