SCIENCE CHINA
Information Sciences

• LETTER •

November 2023, Vol. 66 219401:1–219401:2
https://doi.org/10.1007/s11432-022-3710-6

Human action recognition using a time-delayed photonic reservoir computing

Chao KAI[1], Pu LI[1,2*], Yi YANG[1], Bingjie WANG[1], K. Alan SHORE[3] & Yuncai WANG[2]

[1]*Key Laboratory of Advanced Transducers and Intelligent Control System, Ministry of Education, Taiyuan University of Technology, Taiyuan* 030024, *China;*
[2]*Guangdong Provincial Key Laboratory of Photonics Information Technology, School of Information Engineering, Guangzhou* 510006, *China;*
[3]*School of Computer Science and Electronic Engineering, Bangor University, Wales* LL57 1UT, *UK*

Received 24 October 2022/Revised 23 December 2022/Accepted 28 February 2023/Published online 20 October 2023

Citation Kai C, Li P, Yang Y, et al. Human action recognition using a time-delayed photonic reservoir computing. Sci China Inf Sci, 2023, 66(11): 219401, https://doi.org/10.1007/s11432-022-3710-6

Human action recognition (HAR) has important applications, including video retrieval, entertainment, autonomous navigation systems, and visual surveillance systems [1]. Unfortunately, accurate and efficient HAR remains a challenging task in the field of computer vision because its modeling and characteristic representation belong to three-dimensional space-time instead of common two-dimensional space [2].

Because of the development of artificial intelligence, artificial neural networks (ANNs), particularly deep neural networks (DNNs), are considered a type of very effective method for HAR [3]. However, these DNNs are computationally expensive because they require training various hidden layers, and their training usually is difficult because of the problem of exploding or vanishing gradients in back-propagation.

Time-delayed reservoir computing (TDRC) is a simple ANN that has recently gained considerable research interest [4]. Instead of a series of hidden layers, this type of ANN uses a nonlinear physical component with delayed feedback as a reservoir to construct a virtual network. Only the connection weights of the output layer must be trained, and the connection weights of other layers must be generated randomly. Thus, TDRC can greatly simplify neural network architecture, thereby reducing the training cost. In particular, we note that more recently, TDRC has been successfully translated to the photonic platforms with the potential of fast processing speeds and low power consumption [5]. In this regard, TDRC is an ideal hardware-friendly candidate for complex tasks.

Herein, we innovatively explore a time-delayed photonic reservoir computing (TDPRC)-based scheme for HAR. In particular, an optical feedback semiconductor laser is used as the optical reservoir to realize HAR. This feedback endows our TDRPC with a short-term memory that is crucial for temporal tasks [3]. Furthermore, we use outputs from several feedback delay times to represent the states of total virtual nodes so that the size of the optical reservoir can be greatly scaled within a limited length of the feedback cavity. After investigating the influence of critical hyperparameters on the recognition performance, we numerically demonstrate that using our method can achieve a high recognition accuracy rate of 98% for HAR. This study confirms that TDPRC is a promising candidate for applications in computer vision because of its simple structure and high performance. To our knowledge, this study is also the first introduction of RC to the field of HAR.

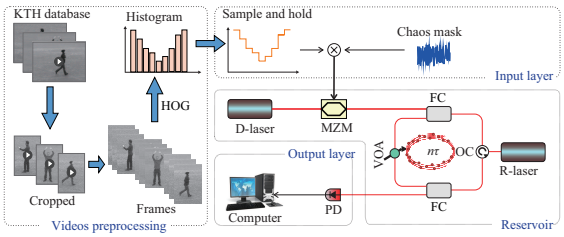

Figure 1 (Color online) Schematic diagram of TDPRC for HAR.

Theoretical model of TDPRC. Figure 1 schematizes our TDPRC for HAR. The human action database used is the famous KTH, one of the most popular and most used HAR datasets. The KTH dataset contains six types of human action behavior, including walking, running, jogging, waving, clapping, and boxing. Each action is performed by 25 volunteers and recorded using a relatively stationary common camera. Before being put into the input layer, the human action videos derived from the KTH database are simply preprocessed to extract associated human action features.

* Corresponding author (email: lipu8603@126.com)

Particularly, the original videos are first extracted in frames, and then the frames are cropped from the original 160 × 120 pixels to 80 × 120 pixels centered on the human body. Second, the histogram of the gradient (HOG) feature extraction technique is used to obtain the feature descriptor $h(n)$. In the input layer, we sample $h(n)$ and hold each sampling value of $h(n)$ for a time T to obtain the continuous signal $h(t)$. Then, $h(t)$ is used to compute the input vector $\boldsymbol{S}(t)$ = mask × $h(t)$. Here, we use a chaotic mask with a periodicity of T ($T = n\tau$), which is based on the white chaos. n represents the sum of multiple feedback delay times, and τ is the length of the feedback delay time. The details of the white chaos are provided in Appendix A.

In the reservoir layer, the response laser (R-laser) with a short delay loop is the solitary physical node of TDPRC. First, we use the input vector $\boldsymbol{S}(t)$ from the input layer to modulate the phase of the drive laser through a phase modulator, and then the modulated output light is injected into the R-laser. Second, the states of the virtual nodes are obtained by taking the transient response of the R-laser within each interval time θ (θ = 50 ps). Thus, the delay loop with a length of 5 ns induces up to 100 virtual nodes (5 ns/50 ps) using the conventional method described in [5]. To obtain more virtual nodes under a simple structure, a novel technique of multiple delay times $n\tau$ is used to define the sum of virtual nodes, where n is a positive integer. That is, the total virtual node states come from the sampled output of the R-laser within a time of $n\tau$. This method is equal to expanding the sum of virtual nodes by continuously running TDPRC for n delay times without increasing the length of the feedback delay time. Thus, the sum of the virtual nodes N can be computed as $N = n\tau/\theta$. Finally, we achieve an N-dimensional vector $\boldsymbol{x}_i(n)$ that reflects the total states of N virtual nodes in the time of $n\tau$. The simulation model of the optical reservoir is shown in Appendix A.

In the output layer, we can obtain the output vector $\boldsymbol{y}(n)$ as the output result by computing the product of $\boldsymbol{x}_i(n)$ and the output layer connection weights vector $\boldsymbol{W}_i$, which is shown in (1). Then, we employ the winner-take-all decision strategy and one-hot encoding to correlate the output $\boldsymbol{y}(n)$ to a certain action. During training, the output vector of the TDPRC $\boldsymbol{y}(n)$ should be as closely as feasible to the target vector $\boldsymbol{y}_{\text{target}}(n)$. Thus, we acquire the optimal value of $\boldsymbol{W}_i$ by using the Tikhonov regression algorithm, as shown in Appendix A.

$$\boldsymbol{y}(n) = \sum_{i=1}^{N} \boldsymbol{W}_i \ \boldsymbol{x}_i(n). \tag{1}$$

Hyperparameter optimization of TDPRC. The recognition performance of TDPRC and its system states are strongly correlated. Thus, we investigate the influence of seven hyperparameters on the system states and the performance of our TDPRC: the number of samples, the virtual node size, the mask standard deviation, the R-laser bias current, the injection strength, the feedback strength, and the frequency detuning. By analyzing their influence, we obtain a set of optimal hyperparameters that enable TDPRC to achieve a good recognition result. The specific optimum values of the parameters can be confirmed as follows: (i) The sample size is set to 15. (ii) The virtual node size N is chosen as 900. (iii) The mask standard deviation σ is chosen as 0.3. (iv) The bias current I_{RL} is set as 1.35 I_{th}. (v) The injection strength κ_{inj} is chosen as 0.5. (vi) The feedback strength κ is set as 0.1125. (vii) The frequency detuning Δf is set as −15 GHz. The detailed optimization process is shown in Appendix B.

Typical recognition results. We utilize optimized TDPRC to identify the actions of the other ten volunteers in the testing set. Overall, the total recognition accuracy is 98% for all human behavior classes using TDPRC. The detailed recognition results are found in Appendix C. Notably, the recognition performance can be further enhanced by increasing the sum of the training samples when the resource efficiency is ignored. Meanwhile, we supplement HAR results without using the reservoir layer or HOG, which are provided in Appendix C. Finally, we also consider the complexity and computational cost of our TDPRC through comparison with other typical DNNs for HAR, as shown in Appendix C.

Conclusion. We have demonstrated a HAR method based on TDPRC. A semiconductor laser running several feedback delay times is used as a single-physical-node reservoir. Through the use of R-laser output from several feedback delay times, our TDPRC can avoid the inherent limitation of relatively few virtual node numbers induced by the small scale of the feedback cavity. Our simulation results show that the proposed method can achieve a recognition accuracy rate of more than 98% with 900 virtual nodes for HAR. Considering its simple architecture and compact size, we believe that TDPRC may be a promising alternative for neural networks in computer vision applications.

Acknowledgements This work was supported in part by National Natural Science Foundation of China (Grant Nos. 61927811, 62175177, U19A2076), Program for Guangdong Introducing Innovative and Entrepreneurial Teams, and Natural Science Foundation of Shanxi Province (Grant Nos. 201901D211116, 201901D211077).

Supporting information Appendixes A–C. The supporting information is available online at info.scichina.com and link.springer.com. The supporting materials are published as submitted, without typesetting or editing. The responsibility for scientific accuracy and content remains entirely with the authors.

References

1 Lu M, Hu Y, Lu X. Driver action recognition using deformable and dilated faster R-CNN with optimized region proposals. Appl Intell, 2020, 50: 1100–1111
2 Poppe R. A survey on vision-based human action recognition. Image Vision Comput, 2010, 28: 976–990
3 Zhang P F, Lan C L, Zeng W J, et al. Semantics-guided neural networks for efficient skeleton-based human action recognition. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2020. 1112–1121
4 Appeltant L, Soriano M C, van der Sande G, et al. Information processing using a single dynamical node as complex system. Nat Commun, 2011, 2: 1–6
5 Song Z, Xiang S, Cao X, et al. Experimental demonstration of photonic spike-timing-dependent plasticity based on a VCSOA. Sci China Inf Sci, 2022, 65: 182401