Hindawi
Complexity
Volume 2021, Article ID 6634156, 5 pages
https://doi.org/10.1155/2021/6634156

Research Article

Deep ChaosNet for Action Recognition in Videos

Huafeng Chen,[1] **Maosheng Zhang**,[2] **Zhengming Gao**,[1] **and Yunhong Zhao**[1]

[1]*School of Computer Engineering, Jingchu University of Technology, Jingmen, China*
[2]*School of Mathematics and Statistics, Yulin Normal University, Yulin, China*

Correspondence should be addressed to Huafeng Chen; chenhuafeng@jcut.edu.cn

Received 2 October 2020; Revised 30 January 2021; Accepted 2 February 2021; Published 13 February 2021

Academic Editor: Zhouchao Wei

Current methods of chaos-based action recognition in videos are limited to the artificial feature causing the low recognition accuracy. In this paper, we improve ChaosNet to the deep neural network and apply it to action recognition. First, we extend ChaosNet to deep ChaosNet for extracting action features. Then, we send the features to the low-level LSTM encoder and high-level LSTM encoder for obtaining low-level coding output and high-level coding results, respectively. The agent is a behavior recognizer for producing recognition results. The manager is a hidden layer, responsible for giving behavioral segmentation targets at the high level. Our experiments are executed on two standard action datasets: UCF101 and HMDB51. The experimental results show that the proposed algorithm outperforms the state of the art.

1. Introduction

Human action recognition in videos is an important area in computer vision, receiving sustained attention from the researchers due to its potential applications such as video supervision, entertainment, user interface, sports, video understanding, and patient monitoring. Current action recognition methods can be classified into three categories by action feature: chaos-based feature [1], manual feature [2], and deep learned feature. Inspired by the chaos-based feature and deep learned feature, we propose deep ChaosNet for action recognition to autonomously learn the nonlinear dynamical feature in video action.

2. Related Works

In this section, we briefly review the literature of action recognition from the chaos-based feature, manual feature, and deep learned feature.

2.1. Chaos-Based Feature. Ali et al. [3] introduced a human action recognition architecture by using the theory of chaotic systems to model and analyze nonlinear dynamics of human actions. Trajectories of reference joints are used as the representation of the nonlinear dynamical system that is generating the action. Xia et al. [4] proposed a human behavior recognition method based on chaotic invariant features and relevance vector machine (RVM). The trajectory generated by the motion of the human joint points is extracted to represent the nonlinear system of human action behavior, and the time delay is estimated by the C-C method. The chaotic invariants representing human behavior are extracted, and the RVM algorithm is used to identify human behavior. Venkataraman and Turaga [1] proposed to use the descriptor of the shape of the dynamical attractor as the feature representation of the nature of dynamics to solve the drawbacks of traditional approaches.

2.2. Manual Feature. Since human behavior is composed of body movements, general human behavior characteristics are based on the underlying visual movement characteristics. The underlying visual features are easy to extract and represent, and the underlying visual motion features of the same action have a certain degree of robustness under different cameras, so they are widely used in early human behavior recognition. There are two categories on human behavior characteristics: local feature representation and global feature representation. Existing global feature descriptions represent the formation of global

spatiotemporal cues through single-frame global features and video frame sequences from aspects of human body contours, posture joint points, and saliency segmentation such as the motion history image algorithm (motion history image, MHI) proposed by Bobick and Davis [5], the adaptation of the shape context algorithm (adaptation of the shape context) proposed by Zhang et al. [6], and the kinematic feature proposed by Ali and Shah [7]. Local feature description of underlying action features is still a hotspot in human behavior recognition research in recent years. Researchers considered the changes in the motion field between frames and proposed various local spatiotemporal feature descriptions, such as STIP [8], MoSIFT [9, 10], and dense trajectories [2, 11].

2.3. Deep Learned Feature. It includes two aspects of deep learning: action convolution features and action timing features. The former uses convolutional neural networks (CNNs) to learn the local depth features of human behavior from different modal data such as RGB image frames and optical flow of behavior videos [12]. On the basis of behavioral convolutional features, it uses methods such as recurrent neural network (RNN), time-series segment network, or linear coding to learn time-series features in multiple stages of behavior development [13]. Due to limited memory capacity of the GPU/CPU and different lengths of behavior duration (shown as different video frames), it is difficult to send all behavior video frames into the deep learning framework for feature learning. Therefore, it is necessary to perform key frame sampling on the behavior video in the behavior recognition process. Most of the existing behavior recognition algorithms use equal sampling [13] or sequential sampling [14–16], ignoring the differences in the development process of human behavior, and the key frames obtained are less representative.

3. Deep ChaosNet Framework

Inspired by Wang et al. and Balakrishnan et al. [15, 17], we propose deep ChaosNet framework for action recognition. The framework is illustrated in Figure 1. Deep ChaosNet features are extracted from video frames. And then, the features are sent to the low-level LSTM encoder and high-level LSTM encoder for obtaining low-level coding output and high-level coding results, respectively. The agent is a behavior recognizer for producing recognition results. The agent, based on hierarchical reinforcement learning, is mainly composed of manager and worker. Manager is a hidden layer, responsible for giving behavioral segmentation targets at the high level. Worker determines the spatiotemporal area of the video subsegment that best characterizes the segmentation target according to the segmentation target and outputs the segmentation recognition result.

3.1. Structure of the Network. The network system structure is shown in Figure 2. The manager LSTM unit obtains environmental status information $[h_t^M]$ according to the input $[c_{t-1}^M, h_{t-1}^W]$ and derives meaningful behavioral stage goals, which are used as the worker LSTM input to guide the worker to select the spatiotemporal region of the next behavioral video subsegment; the formula is as follows:

$$\begin{aligned} h_t^M &= S^M\left(h_{t-1}^M, \left[c_t^M, h_{t-1}^W\right]\right), \\ g_t &= u_M\left(h_t^M\right). \end{aligned} \tag{1}$$

S^M is the manager LSTM nonlinear function, and u_M is responsible for mapping the environmental state information h_t^M to the behavioral stage target g_t. The worker LSTM unit obtains context information h_t^W according to the input $[c_{t-1}^W, g_t]$. Based on h_t^W, we predict the next key frame position d_t, sampling area l_t, and behavior category p_t.

For manager and worker, this project uses a visual attention mechanism to explore areas of salient behavior. The manager attention model mainly explores the significant segment information of the behavior, and the worker attention model assists in searching the behavior key frames and significant areas within the frame. The parameters C_t^M and C_t^W are calculated as follows:

$$\begin{aligned} C_t^M &= \sum \alpha_{t,i}^M h_i^{E_m}, \\ \alpha_{t,i}^M &= \frac{\exp\left(e_{t,i}^M\right)}{\sum_{k=1}^n \exp\left(e_{t,k}^M\right)}, \\ e_{t,i}^M &= m^T \tanh\left(W_a^M h_i^{E_m} + U_a^M h_{t-1}^M + b_a^M\right), \\ C_t^W &= \sum \alpha_{t,i}^W h_i^{E_w}, \\ \alpha_{t,i}^W &= \frac{\exp\left(e_{t,i}^W\right)}{\sum_{k=1}^n \exp\left(e_{t,k}^W\right)}, \\ e_{t,i}^W &= w^T \tanh\left(W_a^W h_i^{E_m} + U_a^W h_{t-1}^W + b_a^W\right). \end{aligned} \tag{2}$$

3.2. Deep Learning Process. The worker strategy learning process is a standard reinforcement learning process. At each step t of the worker, the worker will give a classification prediction result P_t, and then the environment will give a reward R_t, so the goal of worker strategy learning is to minimize the negative value of the reward function. The loss function is

$$L\left(\theta_w\right) = -E_{p_t \sim \pi_{\theta_w}}\left[R\left(p_t\right)\right]. \tag{3}$$

Manager does not directly interact with the environment, and its strategy learning process cannot copy the worker. Compared with manager's time t, the worker strategy $\pi_{\theta_w}(p_t; g_t)$ is relatively stable, and this strategy directly affects the worker's behavior classification output results $p_{t,c}$ at time c. At this point, although the manager is a hidden layer, its strategic goal should be to minimize the negative value of the current reward. The loss function is

$$L\left(\theta_M\right) = -E_{g_t}\left[R\left(g_t\right)\pi\left(p_{t,c}; g_t\right)\right]. \tag{4}$$

4. Experiments and Results

We verify the proposed deep ChaosNet on two standard action datasets: UCF101 [18] and HMDB51 [19]. UCF101 is

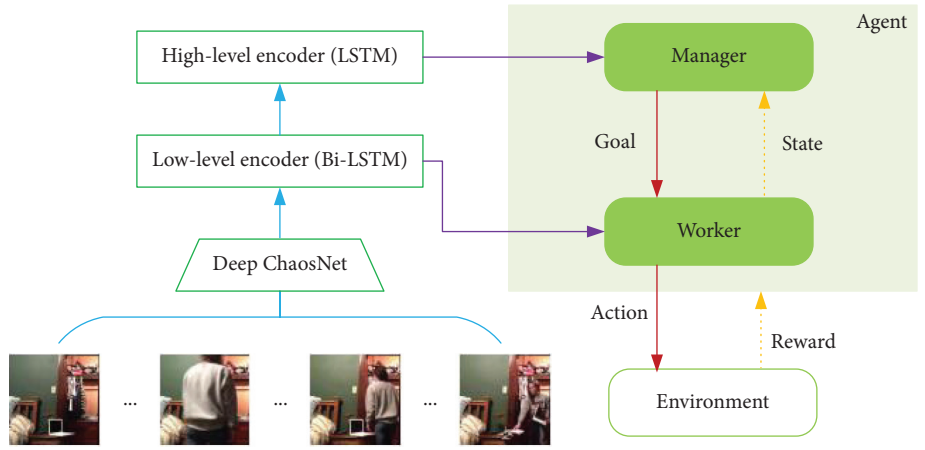

Figure 1: Illustration of the pipeline of the proposed deep ChaosNet feature extraction.

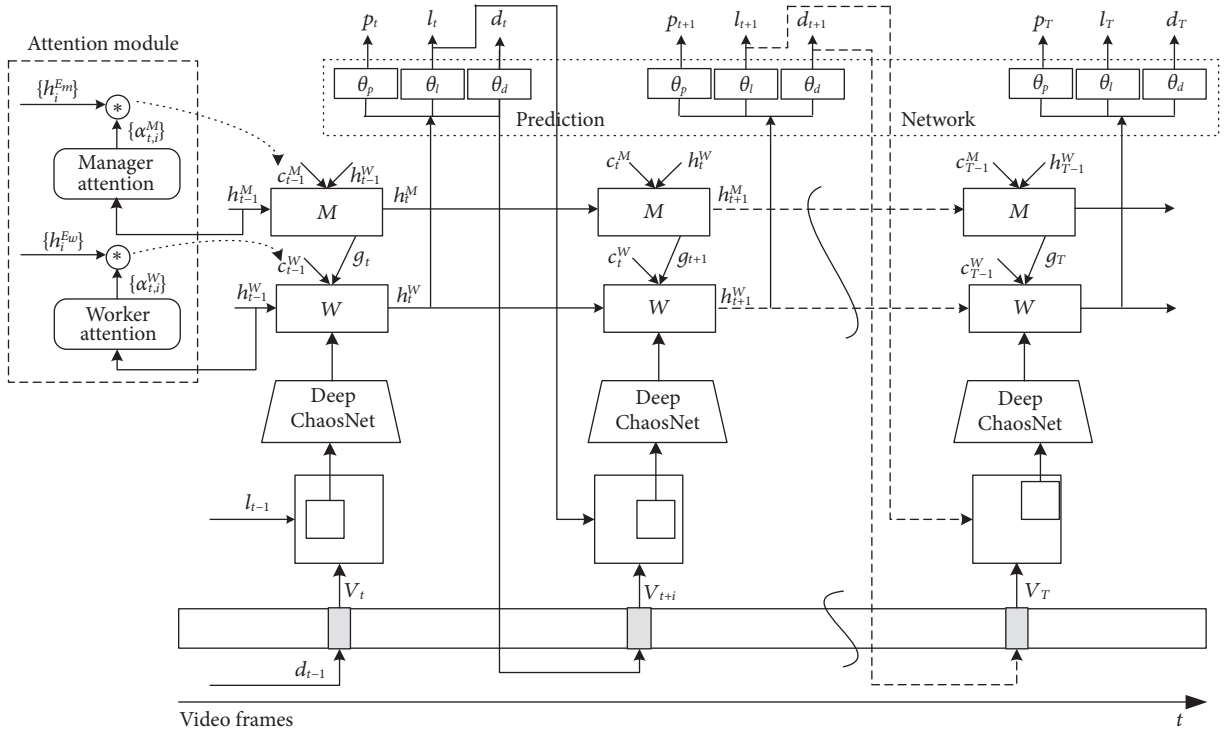

Figure 2: Illustration of the multilayer feature fusing.

an action recognition dataset of realistic action videos with 101 action categories collected from YouTube. Videos of the 101 action categories are divided into 25 groups, and each group can contain 4~7 action videos. Videos from the same group may share some common features, such as similar backgrounds and similar viewpoints. HMDB51 contains 51 types of actions, a total of 6849 videos which are collected from YouTube, Google Video, etc. Each action contains at least 51 videos with a resolution of $320 * 240$.

In the experiments, we construct 7-layer deep ChaosNet for both action datasets. The outputs of the deep ChaosNet are 2048-dim frame features, which are then projected to 512-dim. We use Bi-LSTM with hidden size 512 as the low-level encoder and LSTM with hidden size 256 as the high-level encoder [20]. The worker network consisted of worker LSTM with hidden size 1024. The manager network was composed of manager LSTM with hidden size 256, an attention module, and a linear layer that projected the output of the LSTM into the latent goal space. The environment internal critic was also an RNN, which contained a GRU, a built-in word embedding, a linear layer, and a sigmoid function.

We compare deep ChaosNet with the state-of-the-art deep learning methods [2, 12–16]. The comparison results are listed in Table 1. As shown in the table, the proposed deep ChaosNet exceeds the manual features [2] by 7.3% on UCF101 and by 5.8% on HMDB51. Our method is beyond the action convolution features [12] by 0.2% on UCF101 and

Table 1: Comparison of the proposed method with the state-of-the-art approaches.

Method	UCF101 (%)	HMDB51 (%)
Wang et al. [2]	86.0	60.1
Wu et al. [13]	88.0	59.4
Wang et al. [15]	89.1	65.2
Donahue et al. [14]	90.3	63.2
Chen et al. [16]	92.4	62.0
Sun et al. [12]	93.1	63.3
Deep ChaosNet	93.3	65.9

by 2.6% on HMDB51. The proposed method also outperforms the action timing features [13–16] by 1.1% on UCF101 and 0.7% on HMDB51 at least. Overall, our deep ChaosNet method surpasses all the state-of-the-art methods and becomes the new state of the art.

5. Conclusions

We extend ChaosNet to the deep neural network and apply it to action recognition. We deepen the hidden layers of ChaosNet, and then we separately input still frames and motions among frames into the deep network to extract spatial and temporal action features. The features act as the input for the attention-based action recognition framework. We verify our method on two standard action datasets: UCF101 and HMDB51, and the experimental results indicate that the proposed algorithm is competitive compared with the state of the art.

Data Availability

The data used to support the findings of this study are available from UCF101 (https://www.crcv.ucf.edu/research/data-sets/ucf101/), K. Soomro, A. R. Zamir, and M. Shah, UCF101: a dataset of 101 human action classes from videos in the Wild, CRCV-TR-12-01, November 2012, and HMDB51 (https://serre-lab.clps.brown.edu/resource/hmdb-a-large-human-motion-database/), H. Kuehne, H. Jhuang, E. Garrote, T. Poggio, and T. Serre, HMDB: a large video database for human motion recognition, ICCV, 2011.

Conflicts of Interest

The authors declare that they have no conflicts of interest.

Acknowledgments

This work was supported by the Natural Science Foundation of Hubei Province (Grant no. 2019CFC850), the Outstanding Youth Science and Technology Innovation Team Project of Colleges and Universities in Hubei Province (Grant no. T201923), the National Natural Science Foundation of China (Grant no. 61761044), and the Cultivation Project of Jingchu University of Technology (Grant no. PY201904).

References

[1] V. Venkataraman and P. Turaga, "Shape distributions of nonlinear dynamical systems for video-based inference," *Institute of Electrical and Electronics Engineers Transactions on Pattern Analysis and Machine Intelligence*, vol. 38, no. 12, pp. 2531–2543, 2016.

[2] H. Wang, A. Kläser, C. Schmid, and C.-L. Liu, "Dense trajectories and motion boundary descriptors for action recognition," *International Journal of Computer Vision*, vol. 103, no. 1, pp. 60–79, 2013.

[3] S. Ali, A. Basharat, and M. Shah, "Chaotic invariants for human action recognition," in *Proceedings of the 2007 IEEE 11th International Conference on Computer Vision*, pp. 1–8, IEEE, Rio De Janeiro, Brazi, October 2007.

[4] L.-m. Xia, J.-x. Huang, and L.-z. Tan, "Human action recognition based on chaotic invariants," *Journal of Central South University*, vol. 20, no. 11, pp. 3171–3179, 2013.

[5] A. F. Bobick and J. W. Davis, "The recognition of human movement using temporal templates," *Institute of Electrical and Electronics Engineers Transactions on Pattern Analysis and Machine Intelligence*, vol. 23, no. 3, pp. 257–267, 2001.

[6] Z. Zhang, Y. Hu, S. Chan, and L.-T. Chia, "Motion context: a new representation for human action recognition," in *European Conference on Computer Vision*, pp. 817–829, Springer, Berlin, Germany, 2008.

[7] S. Ali and M. Shah, "Human action recognition in videos using kinematic features and multiple instance learning," *Institute of Electrical and Electronics Engineers Transactions on Pattern Analysis and Machine Intelligence*, vol. 32, pp. 288–303, 2008.

[8] I. Laptev, M. Marszalek, C. Schmid, and B. Rozenfeld, "Learning realistic human actions from movies," in *Proceedings of the 2008 IEEE Conference on Computer Vision and Pattern Recognition*, pp. 1–8, IEEE, Anchorage, AK, USA, June 2008.

[9] M.-y. Chen and A. Hauptmann, *Mosift: Recognizing Human Actions in Surveillance Videos*, Springer, Berlin, Germany, 2009.

[10] H. Chen, J. Chen, H. Li, Z. Xu, and R. Hu, "Compressed-domain based camera motion estimation for realtime action recognition," in *Proceedings of the Pacific Rim Conference on Multimedia*, pp. 85–94, Springer, Hangzhou, China, November 2006.

[11] H. Wang, A. Kläser, C. Schmid, and C.-L. Liu, "Action recognition by dense trajectories," in *Proceedings of the CVPR 2011*, pp. 3169–3176, IEEE, Colorado Springs, CO, USA, June 2011.

[12] S. Sun, Z. Kuang, L. Sheng, W. Ouyang, and W. Zhang, "Optical flow guided feature: a fast and robust motion representation for video action recognition," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 1390–1399, Seattle, WA, USA, June 1994.

[13] Z. Wu, X. Wang, Y.-G. Jiang, H. Ye, and X. Xue, "Modeling spatial-temporal clues in a hybrid deep learning framework for video classification," in *Proceedings of the 23rd ACM*

International Conference on Multimedia, pp. 461–470, Brisbane, Australia, July 2017.

[14] J. Donahue, L. Anne Hendricks, S. Guadarrama et al., "Long-term recurrent convolutional networks for visual recognition and description," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 2625–2634, Seattle, WA, USA, June 1994.

[15] L. Wang, Y. Qiao, X. Tang, and L. Van Gool, "Actionness estimation using hybrid fully convolutional networks," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 2708–2717, Seattle, WA, USA, June 1994.

[16] H. Chen, J. Chen, R. Hu, C. Chen, and Z. Wang, "Action recognition with temporal scale-invariant deep learning framework," *China Communications*, vol. 14, no. 2, pp. 163–172, 2017.

[17] H. N. Balakrishnan, A. Kathpalia, S. Saha, and N. Nagaraj, "ChaosNet: a chaos based artificial neural network architecture for classification," *Chaos: An Interdisciplinary Journal of Nonlinear Science*, vol. 29, no. 11, p. 113125, 2019.

[18] K. Soomro, A. R. Zamir, and M. Shah, "Ucf101: a dataset of 101 human actions classes from videos in the wild," 2012, https://arxiv.org/abs/1212.0402.

[19] H. Kuehne, H. Jhuang, E. Garrote, T. Poggio, and T. Serre, "Hmdb: a large video database for human motion recognition," in *Proceedings of the 2011 International Conference on Computer Vision*, pp. 2556–2563, IEEE, Barcelona, Spain, November 2011.

[20] X. Wang, W. Chen, J. Wu, Y. F. Wang, and W. Y. Wang, "Video captioning via hierarchical reinforcement learning," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, Las Vegas, NV, USA, June 2016.