Temporal and Spatial Information Fusion for Human Action Recognition

This NSF-sponsored research (CMMI-1830295) investigates a bi-stream convolutional neural network (CNN)-based method for human action recognition, which is a prerequisite for human action prediction. The bi-stream CNN structure simultaneously analyzes the spatial and temporal information of human action embedded in images. We integrated *optical flow* with the CNN to extract temporal information in the form of distribution of apparent velocities of action-related movement. The distribution is computed by solving equations subject to the brightness constancy constraint. The temporal information is then analyzed by the temporal path and fused with the spatial path by the bi-stream CNN.

To address data imbalance in manufacturing applications, transfer learning is investigated by leveraging pre-trained networks from large-scale public datasets of human action in non-manufacturing scenarios. We hypothesize that human actions share similar local motifs in images and only differ in application-specific global patterns. This allows us to first train the network using large-scale public dataset to learn local motifs for general human action, before fine-tuning global patterns for manufacturing application, as shown in **Fig. 1**.

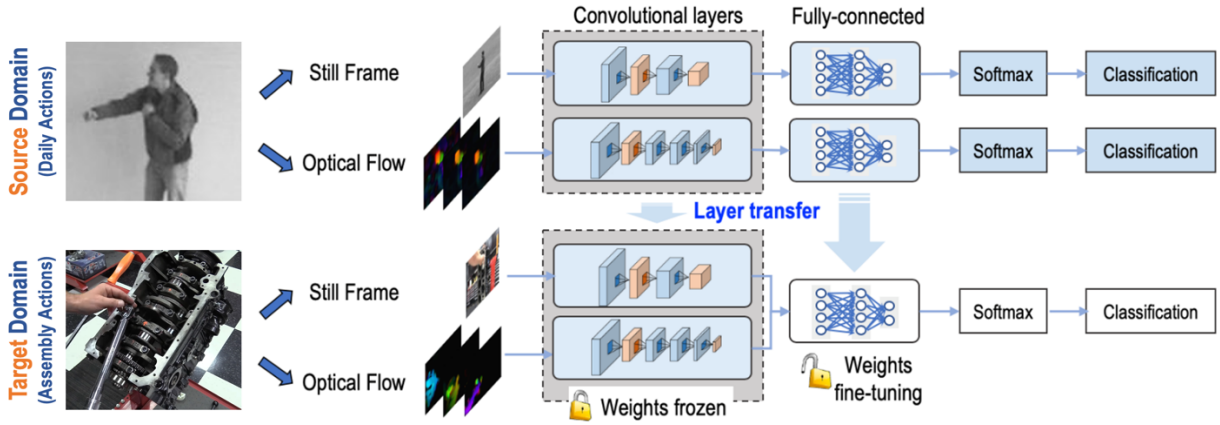

Fig. 1 Bi-stream CNN and transfer learning for human action recognition

We evaluated our approach on an engine assembly scenario consisting of 7 different human actions. We first pre-trained the bi-stream CNN using two non-manufacturing public datasets: KTH and UCF 101, which have more than 11,000 human action images. During transfer learning, we froze the CNN weights associated with local motif analysis and using a small amount of engine assembly scenario data to update the weights associated with global pattern analysis. We have shown that bi-stream CNN and transfer learning can achieve 100% human action recognition accuracy for engine assembly scenario.

Representative Publications

[1] P. Wang, H. Liu, L. Wang and R. Gao, "Deep learning-based human motion recognition for predictive context-aware human-robot collaboration," *CIRP Annals – Manufacturing Technology*, vol. 67, no. 1, pp. 17-20, 2018. https://doi.org/10.1016/j.cirp.2018.04.066

[2] Q. Xiong, J. Zhang, P. Wang, D. Liu and R. Gao, "Transferable two-stream convolutional neural network for human action recognition," *Journal of Manufacturing Systems*, vol. 56, pp. 605-614, 2020. https://doi.org/10.1016/j.jmsy.2020.04.007