Report

Deep Neural Networks for Video Analysis of Human Pose Estimation

By Eralda Nishani, Betim Cico

Introduction

Since CNNs are the most outstanding deep neural network, artificial intelligence has been enhanced in the computer vision field. Because of today's availability of high-end GPUs, in addition to the possibility to have access to sufficient amount of data, training of deep learning models is feasible and number of challenging problems like human pose estimation in frames of videos is solvable. This report explores a systematic mapping study focused on the application of deep learning and CNNs in human pose estimation from videos, highlighting three research directions: It also applies the improvement of the CNN architectures, utilization of the RNNs, and usage of the unsupervised learning schemes.

Summary

Said work presented a meta-review of the electronic literature and classified 119 papers based on the area of research, the type of network, the learning paradigm, and the findings. The authors propose three research questions aimed at enhancing human pose estimation: concentrating on enhancements of the current CNN configurations, the use of RNNs, and utilizing unsupervised learning to address the issue that about the number of low-quality videos which are still unlabeled. The discussion of the existing study demonstrates the growing interest of people in the subject area which has the prospects for research and development of the video analysis in other applications including gesture recognition for sign language.

Critical Analysis

The directions of the further researches pointed by the author in this work refer to the problematic of human pose estimation and are justified. Further enhancement of the existing CNN models may provide solutions for the development of more superior models and in improving the speed and higher accuracies especially in the aspects of joint coordinate and heatmaps with multiple loss functions. The current investigations of RNNs are inspiring because the demonstrated temporal dependences which are significant for video analysis can be within the reach of this subject. However, the training process together with the implementations of RNNs that has been provided here is likely to be challenging. The form of learning is new which may imply reduced reliance on labeled data however the applicability of this mode of learning is going to rely on the formulation of superior algorithms which can discern patterns from streaming video without reference.

Conclusion

This work shows the possibility of applying deep machine learning for the improvement of human pose estimation from the frames of a video. The identified mapping also helps in providing a broad perspective on the state of the current research for developers and to establish what is left to be done and what needs to be done next. The combination of CNNs with RNNs and the application of unsupervised learning methods are on the number of opportunities for improvement. Some of such affordances can lead to enhanced developments as well as the other continuous concerns such as the sign language gesture recognition that will foster interaction between people and computers.