Dual many-to-one-encoder-based transfer learning for cross-dataset human action recognition

Tiantian Xu, Fan Zhu, Edward K. Wong, Yi Fang

PII: S0262-8856(16)00002-0
DOI: doi: 10.1016/j.imavis.2016.01.001
Reference: IMAVIS 3452

To appear in: *Image and Vision Computing*

Received date: 16 September 2015
Revised date: 10 December 2015
Accepted date: 2 January 2016

Please cite this article as: Tiantian Xu, Fan Zhu, Edward K. Wong, Yi Fang, Dual many-to-one-encoder-based transfer learning for cross-dataset human action recognition, *Image and Vision Computing* (2016), doi: 10.1016/j.imavis.2016.01.001

Dual Many-to-One-Encoder-Based Transfer Learning for Cross-Dataset Human Action Recognition

Tiantian Xu[a,c] , Fan Zhu[a,b], Edward K. Wong[a,c] and Yi Fang[a,b*]
[a]NYU Multimedia and Visual Computing Lab
[b]Department of Electrical and Computer Engineering, New York University Abu Dhabi
[c]Department of Computer Science and Engineering, Tandon School of Engineering, New York University[1]

Abstract

The emergence of large-scale human action datasets poses a challenge to efficient action labeling. Hand labeling large-scale datasets is tedious and time consuming; thus a more efficient labeling method would be beneficial. One possible solution is to make use of the knowledge of a known dataset to aid the labeling of a new dataset. To this end, we propose a new transfer learning method for cross-dataset human action recognition. Our method aims at learning generalized feature representation for effective cross-dataset classification. We propose a novel dual many-to-one encoder architecture to extract generalized features by mapping raw features from source and target datasets to the same feature space. Benefiting from the favorable property of the proposed many-to-one encoder, cross-dataset action data are encouraged to possess identical encoded features if the actions share the same class labels. Experiments on pairs of benchmark human action datasets achieved state-of-the-art accuracy, proving the efficacy of the proposed method.

Keywords: cross-dataset, action recognition, neural network, transfer learning, domain adaptation

1. Introduction

Human action recognition has drawn immense interests over the years, with its applications in a wide range of fields, including video labeling, video content retrieval, video surveillance, and sports video analysis. With the growing convenience of capturing and sharing videos, the computer vision community has seen a growing variety of human action datasets with substantial amount of videos. While the majority of these video data do not have annotations on them and hand labelling large datasets requires considerable amount of human efforts, researchers are interested in developing mechanisms to automatically generate annotations to these video data. Considering the fact that large-scale datasets always exhibit high intra-class variations, the requirement for a rational number of training data can easily go beyond the number of existing labelled data. Thus, researchers are thinking about the possibility of employing previously annotated datasets to facilitate automatic labeling of new datasets. For the human action recognition problem, different datasets share common actions. If it is possible to transfer knowledge between source and target action datasets, the annotated source dataset can serve as an augmentation to the training data, based on which effective labeling of the target dataset can be carried out. However, the auxiliary domain data may suffer from the serious domain-shift problem. For example, the action 'run' in one dataset may consist of videos of athletes running on field tracks while the same action from another dataset may contain videos of people running on the streets. In order to alleviate such problems, an algorithm that can reduce cross-domain variance is required. Algorithms of this type belong to transfer learning, which is a particular branch of machine learning that aims to utilize knowledge from one source or task to assist the same or a different task on another source.

In this paper, we tackle the problem of action recognition across four benchmark datasets. The major challenge comes from the significant cross-dataset variations. For example, *Diving* sequences in UCF Sports dataset [1] consists of TV broadcast videos with steady

*Corresponding author. E-mail:yfang@nyu.edu. Address: New York University Abu Dhabi, C1-156, P.O. Box 129188, Abu Dhabi, UAE