Computer Vision and Image Understanding 144 (2016) 62–72

Contents lists available at ScienceDirect

Computer Vision and Image Understanding

journal homepage: www.elsevier.com/locate/cviu

Hierarchical transfer learning for online recognition of compound actions

Victoria Bloom[a,b,*], Vasileios Argyriou[a], Dimitrios Makris[a]

[a] Digital Imaging Research Centre, Kingston University, United Kingdom
[b] Coventry University, United Kingdom

ARTICLE INFO

Article history:
Received 21 December 2014
Accepted 3 December 2015
Available online 11 December 2015

Keywords:
Online action recognition
Online interaction recognition
Hierarchical
Transfer learning

ABSTRACT

Recognising human actions in real-time can provide users with a natural user interface (NUI) enabling a range of innovative and immersive applications. A NUI application should not restrict users' movements; it should allow users to transition between actions in quick succession, which we term as compound actions. However, the majority of action recognition researchers have focused on individual actions, so their approaches are limited to recognising single actions or multiple actions that are temporally separated.

This paper proposes a novel online action recognition method for fast detection of compound actions. A key contribution is our hierarchical body model that can be automatically configured to detect actions based on the low level body parts that are the most discriminative for a particular action. Another key contribution is a transfer learning strategy to allow the tasks of action segmentation and whole body modelling to be performed on a related but simpler dataset, combined with automatic hierarchical body model adaption on a more complex target dataset.

Experimental results on a challenging and realistic dataset show an improvement in action recognition performance of 16% due to the introduction of our hierarchical transfer learning. The proposed algorithm is fast with an average latency of just 2 frames (66 ms) and outperforms state of the art action recognition algorithms that are capable of fast online action recognition.

1. Introduction

The research field of human action recognition has rapidly expanded in recent years with many innovative applications in a range of sectors including healthcare, education and entertainment. In healthcare, action recognition enables touch-free browsing of medical images in operating rooms, physical therapy at home and in clinics and for patient monitoring. In education, action recognition can increase the engagement of users by providing realistic and immersive training simulations. In entertainment, action recognition enables touch-free interaction with Smart TVs and games consoles for more intuitive and natural interaction. A key requirement of these interactive applications is the ability to robustly detect actions in real-time so the system can provide an appropriate response to the user with no apparent delay.

Historically, action recognition research has focused on increasing accuracy on datasets in highly controlled environments. These datasets normally contained a single person that was instructed to perform a single action clearly (see Fig. 1). Recognition was performed offline after viewing a complete sequence and algorithms were evaluated by the number of correctly classified sequences. A recent survey [1] showed perfect or near perfect action recognition accuracy on simple datasets with a small number of actions.

The traditional offline approach led to simplification of the problem, overinflated accuracy and lack of applicability to real world situations. Recent research toward more realistic action recognition has changed to online action recognition where different actions are detected in real-time whilst they are being observed. However, the focus has been on recognising actions which are temporally well separated and easy to segment. In contrast, this work considers multiple actions performed in quick succession, which are critical for robust action detection in natural user interface (NUI) applications. When multiple actions are performed in quick succession movements from different actions may temporally overlap resulting in complex poses, which we term as compound actions. For example, in a full body fighting game a player may throw punches in quick succession, one arm may still be finishing the previous punch whilst the other arm is

* Corresponding author at: Digital Imaging Research Centre, Kingston University, United Kingdom.
E-mail addresses: Victoria.Bloom@kingston.ac.uk, Victoria.Bloom@coventry.ac.uk, vrbloom@gmail.com (V. Bloom), Vasileios.Argyriou@kingston.ac.uk (V. Argyriou), D.Makris@kingston.ac.uk (D. Makris).

http://dx.doi.org/10.1016/j.cviu.2015.12.001
1077-3142/

Fig. 1. Simple boxing sequence with a single person performing a punch (KTH) [2].

Fig. 2. Complex fighting sequences between multiple players, performing multiple actions in quick succession so that the movements temporally overlap (G3Di) [3]. Each row represents a different sequence with visual examples taken every 3 frames.

performing the next punch or a player may leave one arm in the defend position and punch with the other arm (as shown in Fig. 2). Detecting multiple actions in quick succession is a more complex problem than recognising actions which are temporally well separated.

Existing work on recognising more complex actions has to date only been researched in an offline context. To evaluate the performance of action recognition algorithms on more realistic actions several datasets have been extracted from TV and film (YouTube Action Dataset [4], Hollywood Human Actions Dataset [5], UCF sports action dataset [6]). In these datasets the actions are performed in real-world scenarios with diverse and cluttered backgrounds as well as significant changes in viewpoint. The individual actions are realistic but the major limitation of these datasets is that they have been segmented into sequences containing a single action suitable for offline action recognition. The diversity and complexity of real-world datasets makes accurate labelling difficult and time consuming. To overcome this problem Ma et al. [7] employed transfer learning to transfer knowledge from a simpler domain (e.g. KTH [2]) to a more complex target domain (e.g. YouTube Action Dataset) but their approach was limited to offline action recognition. An area that has not