Engineering Applications of Artificial Intelligence 31 (2014) 116–125

Contents lists available at ScienceDirect

Engineering Applications of Artificial Intelligence

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

Optimizing human action recognition based on a cooperative coevolutionary algorithm

Alexandros Andre Chaaraoui [a,*], Francisco Flórez-Revuelta [b]

[a] Department of Computer Technology, University of Alicante, P.O. Box 99, E-03080 Alicante, Spain
[b] Faculty of Science, Engineering and Computing, Kingston University, Penrhyn Road, KT1 2EE Kingston upon Thames, United Kingdom

ARTICLE INFO

Available online 30 October 2013

Keywords:
Human action recognition
Evolutionary computation
Instance selection
Feature subset selection
Coevolution

ABSTRACT

Vision-based human action recognition is an essential part of human behavior analysis, which is currently in great demand due to its wide area of possible applications. In this paper, an optimization of a human action recognition method based on a cooperative coevolutionary algorithm is proposed. By means of coevolution, three different populations are evolved to obtain the best performing individuals with respect to instance, feature and parameter selection. The fitness function is based on the result of the human action recognition method. Using a multi-view silhouette-based pose representation and a weighted feature fusion scheme, an efficient feature is obtained, which takes into account the multiple views and their relevance. Classification is performed by means of a bag of key poses, which represents the most characteristic pose representations, and matching of sequences of key poses. The performed experimentation indicates that not only a considerable performance gain is obtained outperforming the success rates of other state-of-the-art methods, but also the temporal and spatial performance of the algorithm is improved.

1. Introduction

Nowadays, human behavior analysis (HBA) is gaining more and more interest in the field of artificial intelligence and machine learning. Motivated by the wide possible application areas as gaming, natural user interfaces or assistive technologies, just to name a few, great advances have been made in the learning and recognition of human behavior, especially by means of computer vision techniques (Moeslund et al., 2006). For instance, in the case of video surveillance, initially the main goal was to handle person detection, identification (and re-identification) and tracking, whereas activity recognition and scene analysis currently spark the greatest interest, not only of the researchers but also of the industry (Wang, 2013). Specifically, human action recognition (HAR) deals with the lowest level of semantic interpretations of basic human behaviors. For example, motion-based actions as walking, jumping or falling fit into this category. This essential part allows to process further recognition stages. In combination with scene analysis and event detection techniques, complex human activities and long-term behaviors or routines can be recognized (Jiang et al., 2013), for instance, in smart home environments (Rho et al., 2012; Silva et al., 2012; Chaaraoui et al., 2012b).

In this paper, a state-of-the-art vision-based human action recognition method is used addressing multiple optimization targets. First, we seek the best possible set of instances. Due to different kinds of recording errors, noise and instance- or subject-related peculiarities (as clothes, body build, etc.), not all the available instances of a training set are equally useful. Whereas having more samples is usually valuable to learn the intra-class variance of an action class, this is not the case for random noise appearances and outlier values, which tend to spoil and overfit the learning model (Cano et al., 2005). Furthermore, filtering out the redundant instances and optimizing the set of instances to the smallest one which maintains or improves the initial recognition rate lead to a significant spatial and temporal improvement. Second, feature subset selection is applied in order to obtain the optimal selection of elements out of a feature vector. Also in this case, redundant or noisy feature elements can be discarded, which benefits the work of the classifier (Cantú-Paz, 2004). In our case, a human silhouette-based feature is employed whose spatial organization follows a radial fashion. This leads to the natural relation between feature elements and body parts. Certainly, depending on the action to recognize, some body parts may be more relevant than others, and some can be discarded completely. Finally, the optimal values of the algorithm's parameters are determined in order to achieve the best empirical configuration,

* Corresponding author. Tel.: +34 965903681; fax: +34 965909643.
E-mail addresses: alexandros@dtic.ua.es (A.A. Chaaraoui), F.Florez@kingston.ac.uk (F. Flórez-Revuelta).
URL: http://www.dtic.ua.es (A.A. Chaaraoui).

0952-1976/$ - see front matter © 2013 Elsevier Ltd. All rights reserved.
http://dx.doi.org/10.1016/j.engappai.2013.10.003

i.e. the one that leads to the highest recognition rate. The use of evolutionary algorithms for parameter selection is the most common, since it has been applied for decades (De Jong, 1975).

So as to compute these optimizations in an acceptable amount of time, a cooperative coevolutionary algorithm is proposed. Unsupervised selection of instances, features and parameter values is performed simultaneously by using a coevolutionary algorithm with a cooperative behavior. This choice is motivated by the fact that coevolutionary algorithms make it possible to split the domain of the problem relying on the *divide and conquer* strategy, and tackle each part of the optimization problem with respect to their different solution spaces and data types. Furthermore, the cooperative coevolution allows us to consider the intrinsic dependencies which may exist between optimization goals by using a global fitness function and evaluating the cooperation between populations (Derrac et al., 2012). As Section 5 shows, a significant increase in recognition accuracy and a considerable decrease in spatial and temporal complexity are achieved with this proposal, leading to outstanding results on publicly available datasets.

The remainder of this paper is organized as follows. Section 2 summarizes recent and related work in human action recognition and data reduction techniques. A brief definition of a cooperative coevolutionary algorithm is also included. In Section 3, the human action recognition method which is our object of optimization is outlined. Section 4 details the coevolutionary algorithm that is proposed for simultaneous instance, feature and parameter selection. Experimental results and a comparison with the state of the art are specified in Section 5. Finally, we present conclusions and discussion in Section 6.

2. Related work

In this section, recent and related works in human action recognition and data reduction techniques are summarized. The necessary background on coevolutionary algorithms is also included.

2.1. Human action recognition

Existing human action recognition methods based on vision can be categorized by the visual features they employ in order to classify a specific image or a sequence of frames. These are either *local* (also known as *sparse*) descriptors which describe characteristics of multiple relevant points or areas in the image, or global (also known es *dense*, or *holistic*) representations which encode the image information as a whole. Whereas the former mostly rely on color- and gradient-based information in order to detect and describe the points of interest, global features can rely on shape, motion and/or temporal data (Poppe, 2010).

Regarding global representations, several research works rely on human silhouettes (e.g. Bobick and Davis, 2001; Blank et al., 2005; Tran and Sorokin, 2008; Weinland et al., 2006; Thurau and Hlaváč, 2007 and İkizler and Duygulu, 2007). Human silhouettes can be obtained based on image processing techniques as background subtraction, human body detection, or using infra-red, laser or depth cameras. Commonly background subtraction is applied to remove the static background from an image and extract the foreground. Then, a blob detector can identify the part of the foreground that corresponds to the human silhouette. This reduces the problem to a single shape-based region of interest. Bobick and Davis (2001) proposed motion history and motion energy images (MHI, MEI), which respectively encode the age and the location of the motion at pixel-level over a sequence of frames. Weinland et al. (2006) extended this technique to a multi-view and viewpoint-independent motion history volume (MHV) by means of invariant motion descriptors in Fourier space. Classification has been performed combining principal component analysis (PCA) and linear discriminant analysis (LDA) for dimensionality reduction, and Mahalanobis distance for feature matching. Radial histograms of the human silhouette and the optical flow of the *X*- and *Y*-axis are employed in Tran and Sorokin (2008). This visual descriptor has successfully been used by other authors as, for example, in Li and Zickler (2012) for cross-view action recognition. İkizler and Duygulu (2007) describe the human silhouette based on histograms of oriented rectangular patches extracted over the whole body. Then, different ways of considering the temporal domain are tested. Although best results have been achieved using dynamic time warping, frame-by-frame voting and global histogramming achieved similar results, suggesting that dynamics are not indispensable.

Looking at related optimizations of HAR methods, we find that feature subset selection has been applied previously with success. In Jhuang et al. (2007), feature subset selection by means of a support vector machine (SVM) is applied to position-invariant spatio-temporal features, resulting in a reduction of 24 times of the number of features. Spatio-temporal interest points are also used by Bregonzio et al. (2010), where the global distribution information of interest points is exploited. Since the feature space dimension is very high, redundant features are eliminated. Feature selection is applied based on the relevance of each feature, i.e. the proportion of inter-class variation with respect to the intra-class variation. Kovashka and Grauman (2010) target to learn the most discriminative shapes of space–time feature neighborhoods. Multiple kernel learning is employed in order to determine the appropriate distance metrics between interest points. Entropy is used as a measure of importance in Ikizler et al. (2008), so as to choose the region of the human body where most of the motion occurs. In this way, the feature size of a histogram of orientations of border lines could be reduced by a factor of three.

2.2. Data reduction based on evolutionary algorithms

As has been briefly introduced in Section 1, two of our optimization targets address data reduction. These are to find the best performing selection of instances and feature subset. As stated in Liu and Motoda (2002), this can be seen as selecting rows (training instances) and columns (features) out of the training data. In this sense, a two-fold objective is pursued. First, the recognition rate can be improved by filtering noisy and outlier data (which could lead to overfitting, Wilson and Martinez, 2000), obtaining a more consistent learning model. Second, execution time can be reduced without compromising the success rate if the redundant training data is ignored. Note at this point that obtaining suboptimal selections in acceptable execution times is pursued, since to assure condition of optimality would require exhaustive search algorithms.

Whereas a solid state of the art exists regarding instance selection (Wilson and Martinez, 2000; Jankowski and Grochowski, 2004; Grochowski and Jankowski, 2004), evolutionary algorithms (EA) for this purpose are still sparingly being used. Cano et al. elaborated a comparison between evolutionary and non-evolutionary instance and feature selection methods, and concluded that the former consistently performed better in both terms of recognition accuracy and spatial and temporal performance. A generational genetic algorithm (GA), a steady-state GA, a heterogeneous recombination and cataclysmic mutation (CHC) adaptive search algorithm, and a population-based incremental leaning specific EA have been included in the comparison (Cano et al., 2003). In García et al. (2008), a memetic algorithm is proposed for instance selection, tackling the problem of selection in large scale databases. A cooperative coevolutionary algorithm is used for instance selection in García-Pedrajas et al. (2010), where the obtained results compared favorably with standard and also recently