Contents lists available at ScienceDirect

Signal Processing: *Image Communication*

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

A human action recognition approach with a novel reduced feature set based on the natural domain knowledge of the human figure

Gloria Castro-Muñoz [a], Jorge Martínez-Carballido [a], Roberto Rosas-Romero [b,*]

[a] *Instituto Nacional de Astrofísica, Óptica y Electrónica, Tonantzintla, Puebla, Mexico*
[b] *Universidad de las Américas-Puebla, San Andrés Cholula, Puebla, Mexico*

ARTICLE INFO

Article history:
Received 7 March 2014
Received in revised form
1 October 2014
Accepted 6 October 2014
Available online 1 November 2014

Keywords:
Human action recognition
Machine learning
Multi-class classifiers
Feature extraction
Video signal processing

ABSTRACT

Current video surveillance systems are not designed to raise an automatic alert in case of situations that put people lives at risk such as accidents, assaults and terrorism among others. This is due to the fact that these systems are not able to analyze huge amounts of video signals at higher processing speed where these signals come from cameras installed in the worldwide network. Faced with this situation, scientific communities are combining efforts to design algorithms and hardware to accelerate the processing of video signals. However, most of the methods proposed to date are too complex to be implemented in hardware at the place where the video camera is installed. In this paper, we report a significantly reduced novel feature set to design an analysis algorithm of significant less complexity which recognizes human actions from video sequences. The proposed method is based on the natural domain knowledge of the human figure such as proportions of the human body and foot positions. The analysis is characterized by working on sub-sequences of the entire video signals, processing a small fragment of the whole image, estimating the location of the region of interest, using simple operations (sum, subtraction, multiplications, divisions), extracting a reduced number of features per frame (6 features), and using a combination of four linear classifiers (one perceptron and three support vector machines) with a hierarchical structure. The method is evaluated on two of the datasets cited in the human action recognition literature, the *Weizmann* and the *UIUC* datasets. Results show that for the case of the Weizmann dataset, the *correct classification rate* (CCR) is 99.95% when the LOOCV Protocol is used and 98.38% for the case of Protocol 60–40, which is comparable or even higher than that of current state-of-the-art methods. Confusion matrices were also obtained for the UIUC dataset, where the obtained CCR is 100% for the case of the LOOCV Protocol and 99.35% when Protocol 60–40 is used. The experimental results are promising with much fewer features (between 85 and 113 times less features), compared with other methods, and the possibility of processing more than 200 fps.

[*] Corresponding author.
E-mail addresses: cgloria@inaoep.mx (G. Castro-Muñoz), jmc@inaoep.mx (J. Martínez-Carballido), roberto.rosas@udlap.mx (R. Rosas-Romero).

http://dx.doi.org/10.1016/j.image.2014.10.002
0923-5965/© 2014 Elsevier B.V. All rights reserved.

1. Introduction

1.1. Motivation

Human action recognition (HAR) from video sequences is an area of research which has captured the interest of a large number of researchers since some years from now [1–5] because of its potential application areas among which we find ambient-assisted living (AAL) [6], automatic annotation of video [7], human–computer interfaces [8], and video surveillance [9,10]. In relation to the video surveillance application, although decrease of costs in electronics has allowed populating cities with video surveillance cameras, this technology has not been good enough to provide the security required by modern society in relation to issues such as accidents, crime prevention and terrorism concerns. This is mainly due to the fact that continuous and autonomous analysis of millions of video sequences from multiple video cameras is not an easy task. As a consequence, scientific communities have been joining efforts to design algorithms and hardware for implementation of distributed video surveillance systems at a higher speed. Despite different efforts, most of the proposed algorithms to date are too complex to be implemented in hardware at the place where the camera is located and it is too difficult to provide these systems with a distributed nature. For this reason, we seek to collaborate in the design of algorithms aimed at HAR which are characterized by simplicity. We propose a method based on the natural domain knowledge of the human figure such as proportions of the body and number of feet touching the ground, represented as standard features of corresponding image coordinates and geometrical magnitudes.

1.2. Contributions

Major contributions presented in this work are:

1. Recognition of human actions is carried out to a high level so that it is not necessary to track limbs of the human body on an individual basis.
2. The image does not have to be scaled to a fixed size and do not require perfectly segmented silhouettes for high performance.
3. There is an extraction of a reduced set of silhouette's features (six) per frame using only simple operations such as addition, subtraction, multiplication, and division operations.
4. The method uses a multi-class hierarchical classifier which is purely linear.
5. Based on the points mentioned above, the algorithm has low computational cost and can be readily implemented as an embedded system in a video camera.
6. We achieved comparative or even higher *correct classification rates* (CCR) compared to the other state-of-the-art methods in the literature.

1.3. Outline of the paper

The rest of this paper is organized as follows. In Section 2, there is a brief review of HAR related work. The proposed recognition method is presented in Section 3. Section 4 describes the design of the multi-class classifier based on a tree structure. Experiments and results are presented in Section 5. Finally, Section 6 concludes the paper.

2. Related work

2.1. Human representation

In the following, we will review the work on fundamental representation of the human figure, followed by a discussion.

Based on the modeling of human representation, previous methods can be organized into two large groups, those that use a *holistic representation* and those that use a *part-based representation*. Methods based on a holistic representation use the whole ROI information to characterize the action. ROIs, which contain the whole human silhouette, are used in the following papers. In [11] Blank et al. represent human actions as three-dimensional objects generated by grouping of silhouettes in volumes on the space-time domain. In [12] Guo et al. modeled an action as a temporal sequence of deformed centroid-centered silhouettes. In [13] Chen models the action as a sequence of parameters from star figures represented as Gaussian mixture models (GMM), where a star figure is bounded by the smallest convex polygon containing the human silhouette. Other works have also used whole ROIs, but on raw images instead of silhouettes. In [14] Schindler and Van Gool recognize the human action of short sequences of video (snippets) using shape information (local edges) and movement (optical flow) on raw images. Derpanis et al. [15] generate three-dimensional volumes by grouping whole ROIs on a space-time map and measuring energy through derivatives and widely tuned three-dimensional Gaussian filters. Instead of using the whole image contained by the ROI, other methods are based on rectangular image patches for extraction of information to characterize human action (part-based representation). In the following methods, the ROI is divided in an equal-size grid. Jimenez et al. [16] propose a multi-scale HAR descriptor based on a *Pyramid of Accumulated Histograms of Optical Flow* (PaHOF). Optical flow between two consecutive frames is represented as histograms of orientation vs. magnitude accumulated over time and computed for each cell on a grid of non-overlapping regions. In [17], Ikizler et al. proposed a pose descriptor for HAR, called *histogram of oriented rectangles* (HOR), where the ROI is divided on an equal-size grid and the HOR is computed within each cell. In [18], Baysal and Duygulu use a part-based representation for HAR, where information of speed and direction of movement are used along with a pose representation based on a collection of line-pairs adjusted to the contour of the human figure.

2.2. Discussion

For the case of methods based on the holistic representation there is an extraction of global features from the ROI where the whole human figure is analyzed and the length of the description vector is usually fixed. General