Information Sciences 181 (2011) 1874–1887

Contents lists available at ScienceDirect

Information Sciences

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

A lattice-based neuro-computing methodology for real-time human action recognition

Vassilis Syrris *, Vassilios Petridis

Department of Electrical and Computer Engineering, Aristotle University of Thessaloniki, Thessaloniki 541 24, Greece

ARTICLE INFO

Article history:
Available online 4 June 2010

Keywords:
Human action recognition
Real-time prediction
Lattice theory
Time-series approximation
PREMONN algorithm
Neural networks

ABSTRACT

This work describes a computational approach for a typical machine-vision application, that of human action recognition from video streams. We present a method that has the following advantages: (a) no human intervention in pre-processing stages, (b) a reduced feature set, (c) modularity of the recognition system and (d) control of the model's complexity in acceptable for real-time operation levels. Representation of each video frame and feature extraction procedure are formulated in the lattice theory context. The recognition system consists of two components: an ensemble of neural network predictors which correspond to the training video sequences and one classifier, based on the PREMONN approach, capable of deciding at each time instant which known video source has potentially generated a new sequence of frames. Extensive experimental study on three well known benchmarks validates the flexibility and robustness of the proposed approach.

1. Introduction

Video streams are considered a rich information source exhibiting a wide-range of applications: autonomous robots/vehicles, surveillance and security systems, video/image retrieval from multimedia databases and Internet, human–computer interaction, medical diagnosis, etc. Until recently, image sequences remained unexploited due to demanding computation and storage requirements. The combination of sophisticated computational techniques with advanced hardware support has led now the researchers to propose high performance and very promising approaches.

The problem domain of this work is the real-time recognition of single person basic actions such as walking, bending, hand-waving, etc. from an ordered set of images (video sequence). The real-time character of the algorithm means that the system has to be of relatively low complexity. The video shots are considered to be captured by a non-moving monocular camera.

In this context, a typical human action recognition system must deal with two pivotal tasks:

(a) *Low-level processing:* Here, informative features are sought, that can adequately represent each video frame or object of interest expressively and economically. These features can be groups of pixels (connected or segregated) such as corners, blobs, edges, curves, ridges, or elementary geometrical objects like lines, rectangles or circles. This type of local descriptors can be converted (by means of transformation functions) into invariant features unaffected by illumination, rotation, displacement discrepancies, etc. Probabilistic methods or statistical measures can also represent the

* Corresponding author. Tel.: +30 2310996399; fax: +30 2310996367.
E-mail addresses: vsyrris@auth.gr (V. Syrris), petridis@eng.auth.gr (V. Petridis).

0020-0255/$ - see front matter © 2010 Elsevier Inc. All rights reserved.
doi:10.1016/j.ins.2010.05.038

objects or image regions of interest in a more abstract way. In addition, a more advanced feature extraction process can be based on color, shape or texture information. When the time dimension is incorporated either explicitly or implicitly this leads to the mining of motion patterns. This first task constitutes the main focus of the presented work.

(b) *High-level processing:* Here, mapping between extracted features and semantics is pursued. That is, how objects or points of interest are related to living beings or objects such as humans, animals, cars, bikes, trees, etc. Capturing all the possible object appearances, structuring and maintaining a compact and informative knowledge base, conflict handling when new information is entering the database are some of the issues that are addressed at this stage. The classification of descriptors sequences into different classes of human actions is considered to be of great importance as well; this is the second issue addressed in this paper.

Several approaches have been presented for the extraction of suitable descriptors such as learned geometrical models of human body parts [13]; motion/optical flow patterns [8]; characteristic spatial–temporal volumes [11]; feature points [23]; feed-forward hierarchical template matching architectures [14,25]; hierarchical model of shape and appearance [27]; motion history images [4]; dense form (shape) and motion (flow) features [39]; region features [5,42]; spatial–temporal interest points [19,20]; spatio-temporal salient points that represent activity peaks [30,31]; motion estimation [2]; non-linear dimensionality reduced stacks of silhouettes [43]; silhouette histogram of oriented rectangle features [12]; estimated optical flows [22], etc.

Human actions are events affected strongly by the parameter of time. In order to exploit the valuable information which is encapsulated in video sequences, researchers apply a model capable of capturing and modeling the temporal relationships among the frames. Several techniques have been proposed for temporal modeling; the selection of the appropriate technique depends highly on the type of representation one chooses for the object of interest. Some indicative references are: nearest-neighbor matching [18], Support Vector Machines [40], recognition based on spatio-temporally windowed data [7], stochastic dynamic modeling [42], Hidden Markov Modeling [6], exemplar-based nearest-neighbor classification [11] and k-NN classification [1]. In the concrete context of real-time human action recognition, we mention indicatively the cases of:

- Gilbert et al. [10] where a global classifier works as a voting scheme by accumulating the occurrences of the mined compound features (spatio-temporal corners).
- Rigoll et al. [37] where a discrete statistical model is used consisting of a vector quantizer and a special probabilistic neural network or a discrete Hidden Markov Model for image sequences classification.
- Yeo et al. [45] where use is made of compressed domain features such as motion vectors and Discrete Cosine Transform coefficients and a frame-to-frame motion correlation (similarity) measure for action classification.
- Ragheb et al. [36] where features are extracted in the Fourier domain and they are converted to space–time volumes which are handled by a distance-based classifier (like Euclidean distance).
- Peursum et al. [34] where a stick-figure skeleton is estimated and its features are used with discrete Hidden Markov Models.
- Natarajan and Nevatia [26] where a Hierarchical Variable Transition Hidden Markov Model is employed.
- Meng et al. [24] where motion features are used combined with a linear support vector machine classifier.
- Li et al. [21] where luminance field trajectory and its geometric features that contain discriminating information are considered with a classification scheme based on a simple Gaussian Mixture Model classifier.

This paper constitutes an integrated extension of a preliminary work that appeared in [41]. It presents two novel approaches: one for motion detection and one for real-time human action recognition. The latter is formulated as a real-time recognition problem of lattice elements time-series: Firstly, the feature extraction process is based on lattice valuations, first order statistics and second order statistical moments of each video frame. Secondly, real-time recognition is implemented by the PREMONN algorithm [32]. It is a classifier based on the exponential distance function, which evolves a number of credits that are assigned to a respective number of predictors. These predictors or regression models are considered potential generators of a new unseen video sequence; the model with the highest credit at a specific moment is chosen as the best predictor.

The remaining part of this paper is organized as follows: Sections 2 and 3 present the elements of lattice theory and the feature extraction process in the lattice theory context. Section 4 describes the action recognition carried out by the PREMONN algorithm. Section 5 refers to the experimental analysis that is based on three different sets of videos. Section 6 discusses the motion localization capability of the proposed approach and demonstrates its performance on some extra video sequences created for validation purposes. Section 7 refers to the characteristics of the model, possible improvements, future extensions and candidate applications. Finally, the paper closes with Section 8 which summarizes the scope and the significant points of this contribution.

2. Concepts of lattice theory

The structural element of lattice theory [3,38] is the concept of lattice defined as a partially ordered set $(\mathbb{L}, \leqslant) : \forall u, w \in \mathbb{L}, u \wedge w$ and $u \vee w$ exist, where $u \wedge w$ is called *meet* (the greatest lower bound of $\{u,w\}$: $\inf\{u,w\}$) and $u \vee w$