Neurocomputing 100 (2013) 98–106

Contents lists available at SciVerse ScienceDirect

Neurocomputing

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

A hardware friendly algorithm for action recognition using spatio-temporal motion-field patches

Ruihan Bao *, Tadashi Shibata

Department of Electrical Engineering and Information System, The University of Tokyo, 7-3-1 Hongo, Bunkyo-ku, Tokyo 113-8656, Japan

ARTICLE INFO

Available online 7 May 2012

Keywords:
Video analysis
Action recognition
Gesture perception
Patches
Motion-fields

ABSTRACT

A VLSI-hardware-friendly action recognition algorithm using spatio-temporal motion-field patches has been developed. The system employs a hierarchical structure so that the robust recognition can be achieved gradually. At the lower level, motion fields representing local features such as speed and direction are directly calculated from video sequences and further blurred by max filters. At the higher level, a collection of so-called prototype patches are used to recognize query actions by comparing local features in the query videos with those prototypes. In addition, in order to design a system for real-time performance, we intentionally simplify all the calculations into summation operations or boolean operations so that the algorithm can be directly implemented on ultra high speed VLSI chips without much effort. Finally, We tested our system on a gesture perception database as well as widely used action recognition database, and promising recognition performance has been demonstrated.

1. Introduction

Gesture perception or action recognition is receiving growing attention due to its applications in smart surveillance [1], sign language interpretation [2], advanced user interface [3,4] and intelligent robot control. As compared to static image recognition, action recognition usually requires handling overwhelmingly large amount of data because a whole set of video sequences must be analyzed. Moreover, if action recognition is subject to cluttered background, results are often degraded significantly. Therefore, it sometimes requires taking additional measures, i.e. tracking windows or background estimation on the frame-basis. In some cases, it is desirable to build the recognition system directly in the VLSI hardware such as ASICs (application specific integrated circuits) or FPGAs in order to achieve real-time performance. Therefore several constraints need to be further imposed on the algorithms. One important requirement is that the background elimination should be incorporated to the system so that video sequences can be taken as direct input. Another constraint is that computation in the system should be simple enough to be implemented on VLSI circuits either by analog or digital technology.

The process of action recognition usually contains two stages: feature extraction stage and the template matching stage. In the first stage, feature vectors are generated to represent actions in videos. In the second stage, feature vectors are classified using some sorts of classifiers such as Hidden Markov Models (HMMs) or Support Vector Machines (SVMs) [?]. Most of the researches in action recognition, nevertheless, devote their efforts to the first stage, namely, how to generate good features to represent actions. Those algorithms generally fall into three categories. In the first category [5,6], particular parts of human bodies are identified at the beginning, and feature vectors are generated by tracking those parts in spatial and time coordinates. Recognition rates therefore highly depend on the accuracy of distinguishing these specific parts. In the second set of algorithms [7], optical flow estimation is applied in low resolution video samples. Tracking objects from videos is, nevertheless, prerequisite for such systems. In the third group, feature vectors are extracted by using patches (or so-called bag-of-words) [8–10], inspired by latest development in visual cortex and image recognition [11]. Patches in image recognition can be seen as small portions of images that capture local features (see Fig. 1). Algorithms within this category generally include two essential processes: generating prototypes and finding matches between the prototypes and inputs by calculating the similarities. In [8], a system built in a hierarchical way based on the study of vision models has been proposed, and patches in their context are defined as the 2-D spatial form within each frame while matching is being done on the frame-basis. Feature vectors of actions are then calculated by finding best matches across video sequences. High recognition rate for action recognition has been reported. However, the system contains six layers due to additional process in time sequences. Pre-processing to eliminate background is also required for the applications. In contrast, [9] extends the

* Corresponding author.
E-mail address: ray@if.t.u-tokyo.ac.jp (R. Bao).

0925-2312/$ - see front matter © 2012 Elsevier B.V. All rights reserved.
http://dx.doi.org/10.1016/j.neucom.2011.12.041

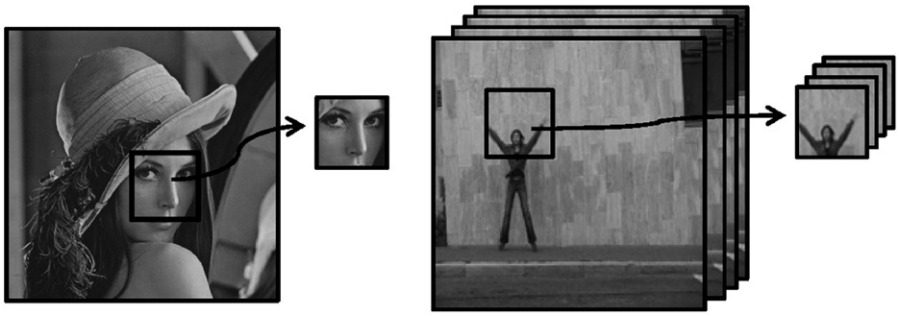

Fig. 1. Two examples of patches in image and video sequences, as one may expect that patches can preserve local features while losing absolute position information.

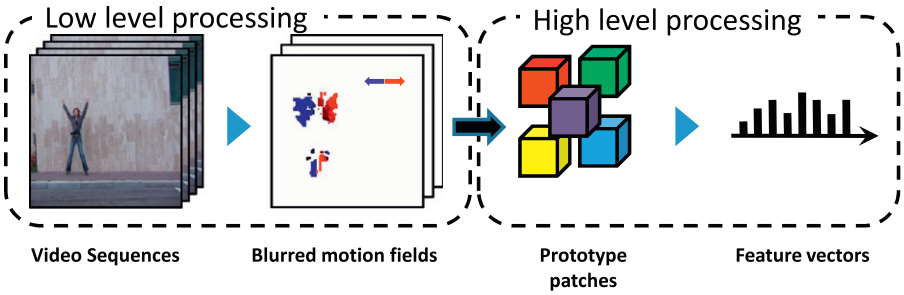

Fig. 2. System overview: at the lower level global features that can capture directions and speeds are calculated, and at the higher level feature vector is generated using prototypes extracted from learning samples.

definition of patches into a spatio-temporal form, in which a small period of time sequences are also included. In the beginning, interesting points must be detected using Harris corner detector. Once interest points are detected, matching is conducted between training samples and testing samples around interest points. In order to represent spatio-temporal patches, PCA is applied in advance to reduce the size of the patches.

Our previous works focus on building hardware compatible systems which may be expected to achieve real-time response performance. In [12], a concept called Directional Edge Displace (DED) maps was introduced, in which motions in videos are highlighted by taking the difference between the time-integrated edge maps and initial edge maps. Then motion fields are calculated using a block matching algorithm, followed by integrating those values along different directions in each frame to represent the characteristics of motions in compact forms. Finally, feature vectors generated from each frame are input to HMMs as a time sequence to carry out recognition. An alternative to generating such a time sequence of vectors is called Projected Directional-Motion Histogram (PDMH) which is formed by integrating motion field maps in both spatial and time domains. Several VLSI chips have already been developed to accelerate the processing using digital [13] as well as analog CMOS technologies [14].

In this paper, we propose a hierarchical model using spatio-temporal patches for action recognition with an effort to make the algorithms compatible to the VLSI architecture developed previously [15]. In order to limit the complexity of the system, we proposed a two level structure. At the lower level, we introduced the concept of Essential Directional Edge Displacement (EDED) maps to eliminate most of the background noises. At the higher level, when feature vectors of actions are to be generated, matching is performed by calculating the similarities between input video sequences and prototype patches, which are extracted from learning video samples. Since patches can capture the local features while losing absolute position information in space and time domains, recognition based on patches is robust to several kinds of variations in positions and time sequences. Our approach is different from traditional spatio-temporal patch methods [9,10] in several aspects. Firstly, interest points detected by Harris corner detector or such kinds are of no necessity for our processing. Instead, we only select patches that contain enough non-zero motion field values. Since background is effectively eliminated (see Fig. 3(c)) at lower levels and only moving parts from video sequences are captured, patches extracted by the simple criterion are well tuned to local features. Secondly, in order to reduce the size of each patch, integration along space and time are adopted, rather than computationally demanding methods like PCA [9]. Thirdly, being inspired by the latest researches from visual cortex, we calculate best matches of prototypes with the whole video sequences, while in previous methods matching among interesting points must be defined at certain locations in videos. Experiment was conducted over a database for gestures with cluttered background [12]. Furthermore, we also tested the algorithm on widely used action databases such as Weizmann human database [9] and KTH human database. Results show that our system can achieve robust recognition performance despite its simplicity in calculations.

2. Robust features for actions

The proposed architecture employs a hierarchical structure (Fig. 2). At the lower level, global features such as speed and directions are extracted. While at the higher level, invariance to spatio-temporal positions is achieved by applying matching between prototype patches and input video sequences.

2.1. Processing at lower level

2.1.1. Essential Directional Edge Displacement (EDED) maps

The lower level processing includes calculating motion field maps and blurring them by max operations. In order to reduce the computational cost and repress background for clear motion field representation, we introduce the idea of Essential Directional Edge Displacement maps at the beginning, which can effectively