Contents lists available at ScienceDirect

Neurocomputing

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

Learning action patterns in difference images for efficient action recognition

Guoliang Lu [a,b,*], Mineichi Kudo [a]

[a] Graduate School of Information Science and Technology, Hokkaido University, Sapporo 060-0814, Japan
[b] School of Mechanical Engineering, Shandong University, JiNan 250061, China

ARTICLE INFO

Article history:
Received 3 January 2013
Received in revised form
19 June 2013
Accepted 25 June 2013
Communicated by Liang Wang
Available online 22 August 2013

Keywords:
Action patterns
Efficient action recognition
Temporal self-similarities
Bag-of-words

ABSTRACT

A new framework is presented for single-person oriented action recognition. This framework does not require detection/location of bounding boxes of human body nor motion estimation in each frame. The novel descriptor/pattern for action representation is learned with local temporal self-similarities (LTSSs) derived directly from difference images. The *bag-of-words* framework is then employed for action classification taking advantages of these descriptors. We investigated the effectiveness of the framework on two public human action datasets: the Weizmann dataset and the KTH dataset. In the Weizmann dataset, the proposed framework achieves a performance of 95.6% in the recognition rate and that of 91.1% in the KTH dataset, both of which are competitive with those of *state-of-the-art* approaches, but it has a high potential to achieve a faster execution performance.

1. Introduction

Vision-based human activity/action analysis has been received more and more attention in recent years. There are many datasets that can be used as benchmarks. Some of them include a large variety of scenes where more than one performer, sometimes a crowd of people, are doing actions simultaneously, as seen in CMU action detection dataset, Hollywood realistic action dataset and INRIA pedestrian detection dataset. Other datasets include one performer in a scene as seen in KTH dataset, Weizmann dataset and Keck gesture dataset. In this study, we concentrate on the latter type of datasets, driven by natural demands of it in single-person oriented applications such as human–computer interaction (HCI), home the-elderly assistance system, motion retrieval and so on.

In this direction, significant research efforts have been made to extract effective action representation for recognition. The detailed surveys can be seen in the works [1–3]. In this study, they are divided into two groups by whether requiring bounding boxes of human body or not. In the former group, bounding boxes are segmented first in each frame by either background subtraction or human detection/tracking, and then, the features are extracted from the normalized bounding boxes based on, e.g., motion flow [8,10,11,28], geometrical modeling of body parts [4–7,39], color/binary appearance [11–13,31–33,37]. This kind of methods has shown good performances in recognition accuracy in some public datasets but also have two common disadvantages:

- It needs a prior procedure of modeling background/human body for background subtraction or human detection. In addition, detection of bounding boxes in real applications is often computationally expensive.
- Its performance heavily relies on the quality of the found bounding boxes. Unfortunately, it is not so easy to obtain high-quality results due to noise and variations between training and observation images.

The latter group of methods do not need pre-segmentation nor tracking individuals in a video. Instead they rely on space–time interested points (STIP) [14,15,35], *spatio-temporal* features (e.g., obtained by combining 3D gradient descriptor and optical flow descriptor [20] or by using 3D DT-CWT [21]), volumetric analysis of video frames [16,18,19,34,36,37], difference images [9,22,23] and so on. For the case of space–time points, the sparse interested points, e.g. detected by [14], are sometimes not sufficient to characterize the human actions [20,35]; the larger number of interested points, e.g. detected by [35], requires a larger computational cost due to the higher dimensionality of action video even for the sparse detection [24]. The *spatio-temporal* features relax the requirement of space–time interested points but need a relatively complex computation method. For volumetric analysis, a large number of space–time pitches may also result in higher computational load. On the other hand, the difference image has been proved the potential capacity/power of

[*] Corresponding author at: Graduate School of Information Science and Technology, Hokkaido University, Sapporo 060-0814, Japan. Tel.: +81 11 706 6854.
E-mail addresses: luguoliang@main.ist.hokudai.ac.jp, luguoliangsdu@163.com (G. Lu), mine@main.ist.hokudai.ac.jp (M. Kudo).

0925-2312/$ - see front matter © 2013 Elsevier B.V. All rights reserved.
http://dx.doi.org/10.1016/j.neucom.2013.06.042

action discrimination [22,23], which can be obtained very fast by frame substraction.

1.1. The previous work

Temporal self-similarity (TSS) is one of the attractive cues for vision based action recognition by virtue of its following properties:

- TSS has a better capacity of absorbing intra-class variations, e.g., the different visual appearances between individuals, performing styles of actions [31,32,41,42].
- By the use of TSS features, less training samples are required for action modeling, e.g., in [19], just only one video sequence is used for modeling one action.

Typical TSS seeks the action patterns with original frame description [10,41], human silhouettes [10], histogram of gradient [31,32], optical flow [31,32] or body trajectory [31,32], with a requirement of subject detection in each frame. A new TSS, called GTSS, has been proposed in our previous work [25], which seeks motion patterns between all pairs of difference images in a given video sequence without requiring finding bounding boxes and has been demonstrated advantages to some typical TSSs. It, however, still needs some improvements for practical usage, considering the two following aspects: (1) the GTSS has succeeded in bypassing time-consuming subject detection for extracting action patterns, but it still requires a relatively large computation cost, that is $O(N^2)$ for a video sequence with N frames. This is not suitable for real-time applications especially for a long-term video. (2) In [25], we have justified the use of GTSS in *dynamic sequence matching* based action recognition. The experimental results showed its priority, in recognition rate, to the conventional TSSs. The recognition rate is, however, very low, that is only 77.8% in Weizmann dataset, which is far away from real-world applications.

1.2. Contributions of this paper

To cope with these two aspects, in this paper, by extending the previous work, we propose a new pattern extraction method using local temporal self-similarities (LTSSs) derived from difference images, and introduce the *bag-of-word* framework to assemble these obtained LTSSs for action recognition. Our main contributions are described in two fold:

- The newly proposed LTSS, as well as GTSS, does not require a time-consuming subject detection, and meanwhile inherits the general properties of conventional TSSs, e.g., the robustness against intra-class variations. On the other hand, it needs less computation cost than the GTSS (see Section 3.4 for details), which satisfies the requirement of practical usage.
- We employ the *bag-of-words* framework [15,24] for action classification by representing the action as a collection of numerous extracted LTSSs, i.e., codebook, trained from all training video sequences. It is experimentally shown that the proposed LTSS and *bag-of-words* expression is comparable, in recognition rate, with *state-of-the-art* work, but has promisingly higher efficiency in execution (see Section 5.3 for comparison).

The rest of the paper is organized as follows. Overview of the proposed framework is given in Section 2. Section 3 describes the proposed LTSS. In Section 4, the *bag-of-words* framework for action recognition is introduced, followed by experimental results and analysis in Section 5. In Section 6, we discuss our proposed approach. Finally, we conclude the paper and present one possible improvement direction in Section 7.

2. Overview of the proposed scheme

First, we show the overview of the proposed framework in Fig. 1. In the training phase, the LTSSs are computed from every training video sequence and described by block-based descriptors, and then, the *bag-of-words* framework is employed to model actions. In the testing phase, LTSSs of one testing sequence are also computed in the same procedure as in the training, and then, action is classified with the trained human action models in the employed *bag-of-words* framework.

The detailed procedures are then described in the following.

3. Proposed local self-similarities in difference images

For computing LTSSs, we put the following assumptions: (1) the sampling rate (25 fps in experiments) is sufficiently high for

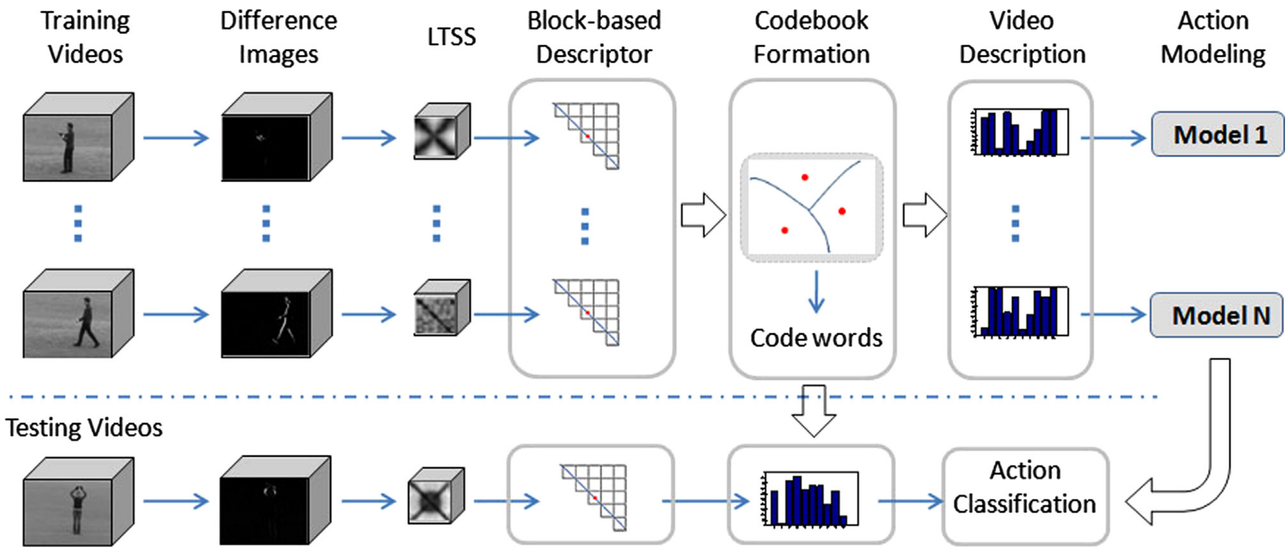

Fig. 1. Overview of the proposed framework.