Signal Processing 124 (2016) 220–232

Contents lists available at ScienceDirect

Signal Processing

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

Investigating the impact of frame rate towards robust human action recognition

Fredro Harjanto [a], Zhiyong Wang [a,*], Shiyang Lu [a], Ah Chung Tsoi [b], David Dagan Feng [a]

[a] School of Information Technologies, The University of Sydney, NSW 2006, Australia
[b] Faculty of Information Technology, Macau University of Science and Technology, Avenida Wai Long, Taipa, Macau SAR, China

ARTICLE INFO

Article history:
Received 1 May 2015
Received in revised form
3 August 2015
Accepted 4 August 2015
Available online 24 August 2015

Keywords:
Evaluation
Human action recognition
Frame rate
Key-frame selection

ABSTRACT

Human action recognition from videos is very important for visual analytics. Due to increasing abundance of diverse video content in the era of big data, research on human action recognition has recently shifted towards more challenging and realistic settings. Frame rate is one of key issues in diverse and realistic video settings. While there have been several evaluation studies investigating different aspects of action recognition such as different visual descriptors, the frame rate issue has been seldom addressed in the literature. Therefore, in this paper, we investigate the impact of frame rate on human action recognition with several state-of-the-art approaches and three benchmark datasets. Our experimental results indicate that those state-of-the-art approaches are not robust to the variations of frame rate. As a result, more robust visual features and advanced learning algorithms are required to further improve human action recognition performance towards its more practical deployments. In addition, we investigate key-frame selection techniques for choosing a set of suitable frames from an action sequence for action recognition. Promising results indicate that well designed key-frame selection methods can produce a set of representative frames and eventually reduce the impact of frame rate on the performance of human action recognition.

1. Introduction

Human action recognition from videos is very important for visual analytics. Vision-based human action recognition is also one of the foremost challenging tasks in the area of computer vision and pattern recognition, which comprises multiple research issues such as view dependence, ambiguity, illumination variations, occlusion, extendibility, computational complexity, and robustness. Although there have been many approaches devised to enable a machine/computer to understand human actions and to perform the task autonomously [1–4], it is in a particularly increasing demand in the era of big data when accessing a large amount of video content acquired with diverse settings (e.g., frame rate and illumination condition) has never been easier. As a result, research on human action recognition has recently shifted towards more challenging and realistic settings than well controlled laboratory settings.

Among all the methods, local spatio-temporal features have been successful in achieving remarkable performance for action recognition. Laptev and Lindeberg [5] first introduced the concept of space-time interest points by extending a 2D Harris-Laplace detector. Schuldt et al. [6] proposed to detect salient sparse spatio-temporal features with automatic scale selection. To produce denser space-

* Corresponding author.
E-mail address: zhiyong.wang@sydney.edu.au (Z. Wang).

http://dx.doi.org/10.1016/j.sigpro.2015.08.006
0165-1684/

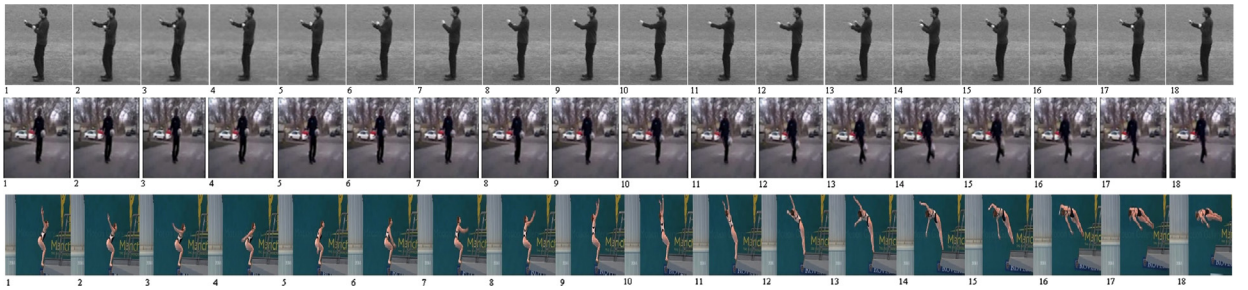

Fig. 1. Sample actions frames. The first row: the first 18 out of 75 frames from action class *Boxing* in the KTH dataset [6]. The second row: the first 18 out of 60 frames from action class *Soccer Juggling* in the UCF YouTube Action dataset [24]. The third row: the first 18 out of 60 frames from action class *Diving* in the UCF Sports Action dataset [22]. Motion redundancy exists in those sample actions frames, such as frames 2, 3 and 4 in the first row, frames 3, 4 and 5 in the second row, and frames 1 and 2 in the third row.

time feature points, Dollar et al. [7] used a pair of 1D Gabor-filters to convolve with a 2D spatial Gaussian function to select local maximal cuboids. Willems et al. [8] proposed a Hessian 3D detector and extended the SURF (Speeded Up Robust Features) descriptor to detect relatively denser and computationally efficient space-time points. A recent trend is the use of densely sampled feature points [9] and trajectories [10] for action recognition. Interest point based features [11] selected through unsupervised learning, History Trace Templates (HTTs) and History Triple Features (HTFs) [12] capturing the spatio-temporal information have also shown to be very useful for simple action datasets with single, staged human actions and uncorrelated backgrounds, such as the KTH dataset [6]. A hierarchical codebook model of local spatio-temporal video volumes [13] has also been used for action recognition.

While there have been impressive progresses, most of the aforementioned methods are still under intensive discussions concerning their practicability for the real-life challenging datasets [14–18], because most interest point detectors utilized are extended from 2D image space domain, which were originally designed for feature matching, not for selecting the most discriminate patches for the human action recognition task. While in realistic videos, some background features are highly correlated with the foreground actions (e.g. diving with water background and skiing with snow background), the proposed features do not provide sufficiently discriminative information for the foreground/background categories. In order to overcome these challenges raised from the realistic videos, several research studies have taken efforts towards providing a more discriminative method by identifying local motions of interest using group sparsity [19], including the global structure information and the ordering of local events [20–22]. A very recent work, using the so-called local part model, and random sampling of the dense trajectory approach, demonstrates state-of-the-art results on the different human action benchmark datasets, including realistic video datasets [23].

Despite numerous research efforts on human action recognition, the impact of video frame rate has rarely been investigated, no matter with laboratory datasets or realistic datasets. Most of the existing approaches perform human action recognition by assuming that different videos are under the same frame rate. However, videos are often acquired by different video capture devices with different frame rates. This problem gets more serious for visual analytics with large scale and heterogeneous action videos which are captured using various imaging devices with diverse settings. For example, if the human action models are learned through a dataset collected from laboratory settings (e.g., the videos at 50 frames per second (fps)), the performance of recognition approach will be expected at the best with the videos captured at the same frame rate, rather than the videos with other frame rates (e.g., surveillance videos with the frame rate at 15 fps), since local features, such as the descriptors of Space-Time Interest Points (STIPs) [5] derived from adjacent frames could be different at different frame rates. It would be impractical to deploy a human action recognition approach, without taking the variations of frame rate from different video capture devices into account. Therefore, it is essential to develop novel human action recognition algorithms in the ear of big data to take the frame rate issue into account.

It is true that a higher frame rate allows capturing more temporal information. However, it also introduces too much redundancy, which may suppress those discriminative information, as shown in Fig. 1. Meanwhile, several research studies demonstrate that human beings are even able to recognize an action when a video is captured at a very low frame rate, or from only a subset of video frames of an entire video sequence. Keval and Sasse [25] indicated that humans are able to complete an action detection task (e.g., identifying a stealing action) at 8 fps. Schindler and van Gool [26] proposed an algorithm to achieve very good recognition performance using only a small number of frames from action sequences of the KTH [6] and Weizmann [27] datasets. However, the impact of video frame rate was not investigated.

The above observations motivate us to investigate the impact of the underlying video frame rate on the performance of human action recognition. One straightforward question is: how well will the state-of-the-art human action recognition approaches perform on human action videos with different frame rates? In this paper we investigate the impact of frame rate on the performance of human action recognition by evaluating four state-of-the-art approaches under three benchmark action video datasets. The four state-of-the-art approaches are: (1) Bag of Visual Words (BoVW) model based method [28], (2) ActionSnippets based method