Pattern Recognition Letters 73 (2016) 91–99

Contents lists available at ScienceDirect

Pattern Recognition Letters

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

A *naïve* Bayes baseline for early gesture recognition☆

Hugo Jair Escalante*, Eduardo F. Morales, L. Enrique Sucar

Instituto Nacional de Astrofísica, Óptica y Electrónica, Luis Enrique Erro No. 1, Tonantzintla, Puebla 72840, Mexico

ARTICLE INFO

Article history:
Received 18 September 2015
Available online 3 February 2016

Keywords:
Early recognition
Gesture recognition
Naïve Bayes

ABSTRACT

Early gesture/action recognition is the task of determining the identity of a gesture/action with as few information as possible. Although the topic is relatively new, there are some methods that address this problem. However, existing methods rely on complex modeling procedures, that do not necessarily paid off the computational effort. Thus, simple yet effective and efficient techniques are required for this task. This paper describes a new methodology for early gesture recognition based on the well known naïve Bayes classifier. The method is extremely simple and very fast, yet it compares favorably with more elaborated state of the art methodologies. The naïve baseline is based on three main observations: (1) the effectiveness of the naïve Bayes classifier in text mining problems; (2) the link between natural language processing and computer vision via the bag-of-words representation; and (3) the cumulative-evidence nature of the inference process of naïve Bayes. We evaluated the proposed method in several collections that included segmented and continuous video. Experimental results show that the proposed methodology compares favorably with state of the art methodologies that are more elaborated or were specifically designed for this purpose.

1. Introduction

Gesture and human action[1] recognition are two widely studied topics in computer vision that can have a huge impact in the field of human-computer interaction. Significant progress has been reported in the last few years [1], in large part because of the release of the Kinect [37]. Most methods tackle the problem in an offline setting, meaning that gestures most be segmented prior to its classification [11,26,40,42]. On the other hand, gesture spotting techniques aim at recognizing gestures online [10,22,29]. In both recognition and spotting, current methods usually segment and recognize a gesture once it has been finished; that is, the whole segment of video has to be seen before a prediction can be made. Hence, traditional methods are not suitable for systems requiring a real interactive experience.

Early gesture-recognition methods aim at identifying the category of a gesture before it has been finished [9,21,28,44]. This type of solutions can improve the interaction experience for users, because intelligent/anticipated decisions can be made (e.g., in response to the gesture that is about to finish). Besides, in certain scenarios these techniques could be used for prevention or alert emission, which could result in fast response against undesired behavior. Despite its potential impact, early gesture/action recognition is a research topic that is in its infancy. A few methods have been proposed, however most of them are based on strong assumptions (e.g., gestures can be clearly distinguished at their beginning, or one can know the duration of the gesture) and complex (yet very effective) modeling procedures (e.g., structured-output prediction models).

This paper describes a simple approach for early gesture recognition based on a well known classifier: naïve Bayes. In a nutshell, we apply the multinomial naïve Bayes model [25] to partial video sequences, where the video is represented under the bag of features representation. This proposal is grounded in the success that multinomial naïve Bayes has had in text mining [25,33], and in the analogy of the bag of words – bag of features representation [39]. Because of the nature of the inference process of naïve Bayes, we can make predictions after seeing any amount of information (even zero[2]!). This is illustrated in Fig. 1. We show that this *naïve baseline* can obtain superior performance to state of the art techniques and at the same time is more efficient. Because we are using a basic version of this classifier, our work can be extended in many ways. Hopefully, our research will pave the way

☆ This paper has been recommended for acceptance by Maria De Marsico.

* Corresponding author. Tel.: +52 222 2663100x8319; fax: +52 222 2663152.
E-mail address: hugojair@inaoep.mx (H.J. Escalante).

[1] We describe our methods in terms of gesture recognition, although most descriptions apply to action recognition as well. We report results in both tasks.

[2] Please note that it may not make sense to make predictions without seeing any evidence, nevertheless, we wanted to point out that with naïve Bayes it is possible to do this: we can use the prior probabilities (see Eq. (1)) for making predictions under total uncertainty.

http://dx.doi.org/10.1016/j.patrec.2016.01.013
0167-8655/

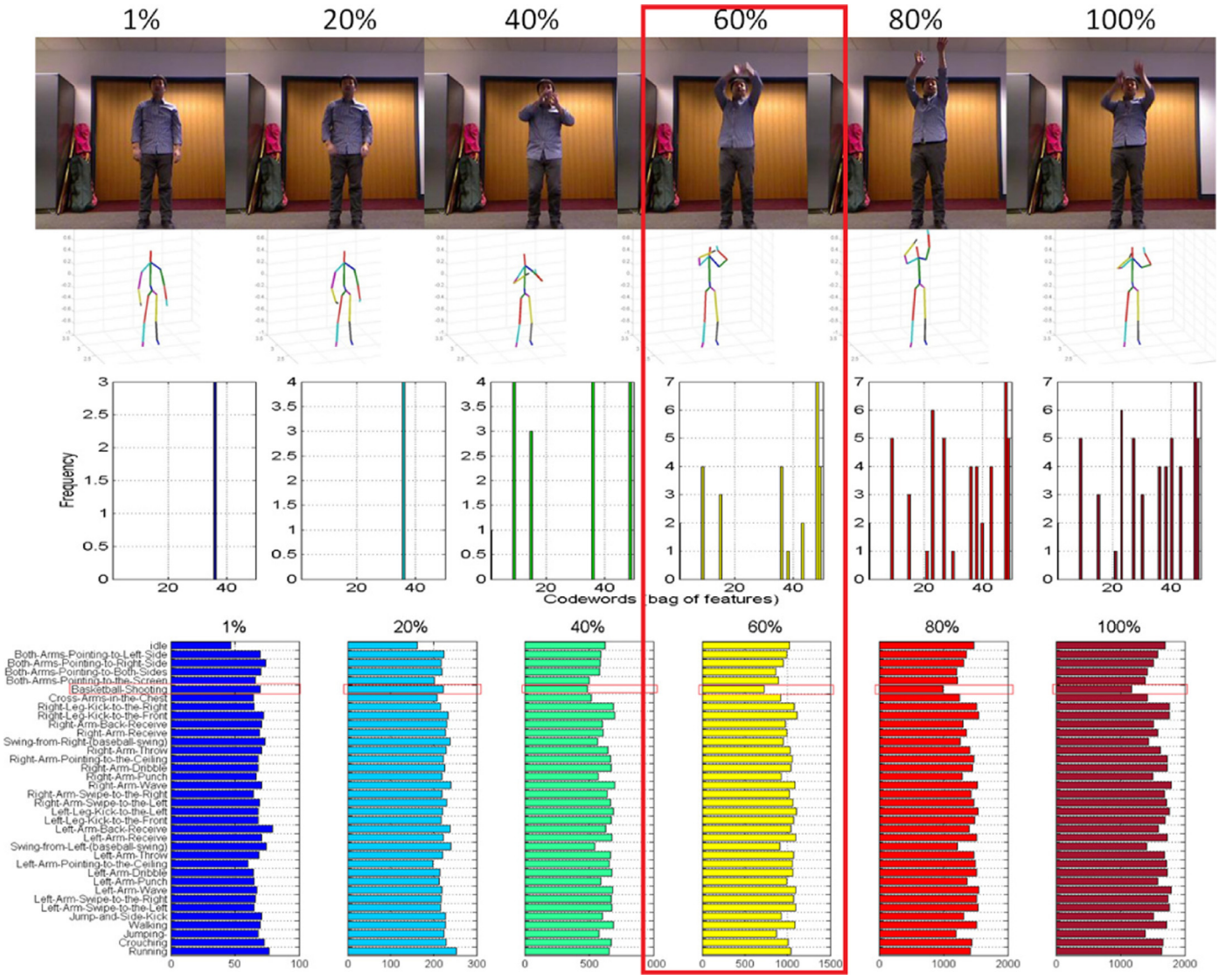

Fig. 1. Overview of the proposed approach. From top to bottom: a video is analyzed sequentially (RGB video and skeleton data are shown in rows 1 and 2), each time building the bag of features representation for the partial sequence (third row, for clarity we show the bag of features representation for 50 randomly selected bins/features). At each time t, the naïve Bayes classifier makes predictions (fourth row). We show the negative-log probabilities for the different gestures (the correct class is marked with a red rectangle). Under this scheme we can make predictions at any time t, (in this example, we recognize the gesture after processing 60% of the video). (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

for the development early-recognition methods based on more sophisticated generative models.

The remainder of this paper is organized as follows. Next section reviews related work on gesture recognition with emphasis on early-recognition methods. Section 3 describes the proposed approach to gesture recognition. Section 4 reports experiments and results on benchmark data sets and a comparison with state of the art techniques. Finally, Section 5 summarizes our findings and outlines future work.

2. Related work

Early gesture recognition is a relatively new research topic. The first attempts were published less than a decade ago [28], and it remains a somewhat unexplored topic. The first works on early recognition attempted to extend and apply standard methods (e.g., DTW) [36]. However, they were not specifically designed for early recognition, but just were evaluated for this setting. Truly extended methods have been proposed, see e.g., [32], although their performance is limited (e.g., around 50% of accuracy in the same data set we consider, MSRDaily3D). More recently, methods that identify templates on the execution of gestures at their beginnings were proposed (i.e., learn to model the initial parts of gestures) [21,28,35]. Template-based methods have the disadvantage that they do not perform well if there are gestures with similar *beginnings* (e.g., *come* and *go-away* hand gestures).

With the release of the Kinect [37], effective methods that take advantage of body-joints information and depth video have been proposed. For instance, in [17] the max-margin early event detector (MMED) method was proposed. MMED is based on structured-output SVMs with a training mechanism that shows sequences frame by frame to the model. MMED is able to make predictions for partial events, and it was proposed for early analysis of facial expressions. This method was later extended to actions in [16]. More recently, in [18], an improved formulation called SMMED for early event detection was proposed. Differently from the methods in [16,17], the margin-maximization formulation from [18] discards gesture-classes as the sequence of video is being processed, with