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/

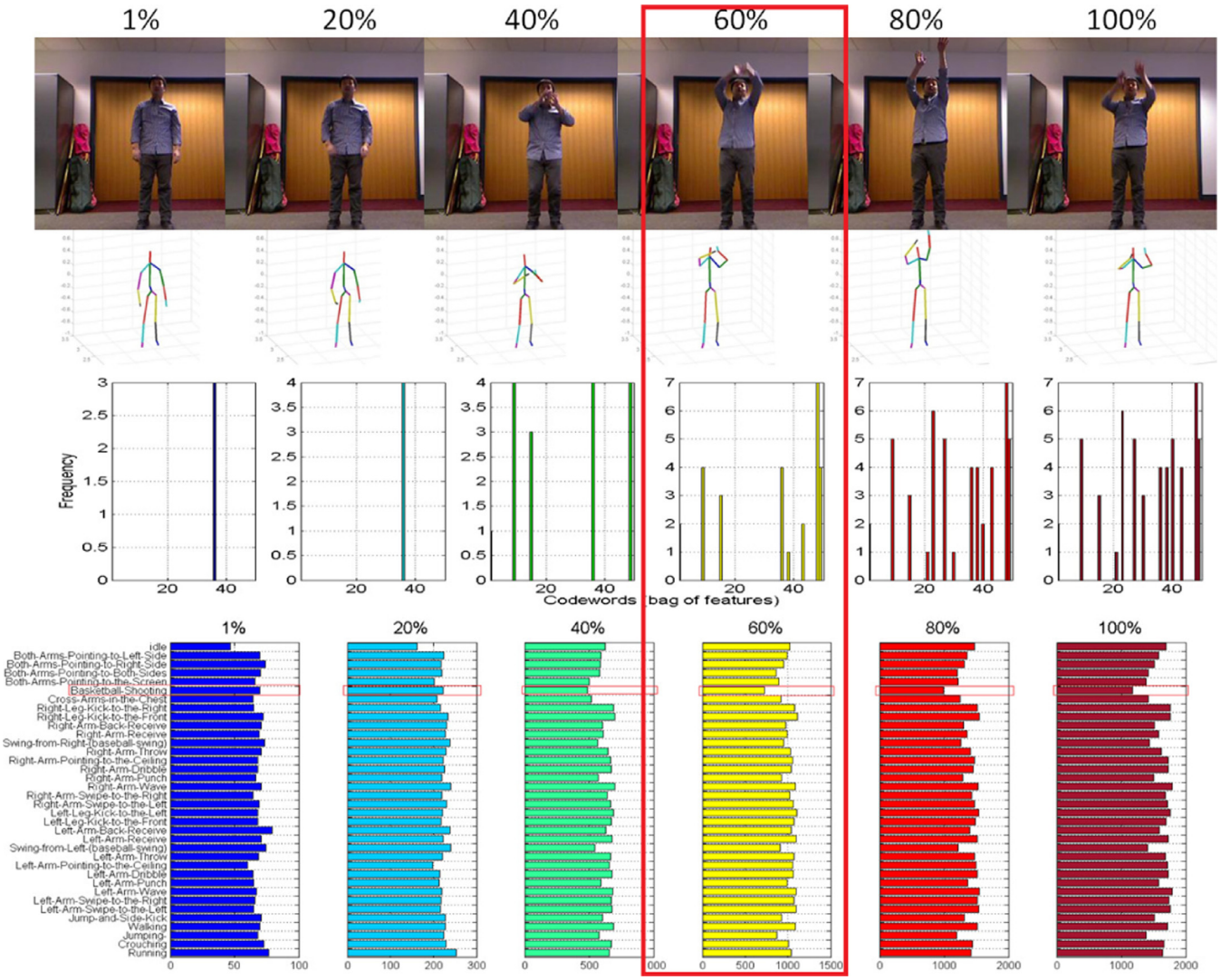

Fig. 1. Overview of the proposed approach. From top to bottom: a video is analyzed sequentially (RGB video and skeleton data are shown in rows 1 and 2), each time building the bag of features representation for the partial sequence (third row, for clarity we show the bag of features representation for 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

the aim of making predictions as early as possible (i.e., once all but one class has been discarded). The main advantage of these margin-maximization techniques is that they are very competitive (in terms of both recognition and earliness-of-recognition performances), besides they are based on principled learning procedures and they are trained discriminatively and specifically for the early-recognition task. A problem with these methods is the complexity of the learning problem and the computational cost when training the models. More importantly, these models assume the input sequences have a similar duration or can be aligned easily, see e.g. Fig. 2 in [18].

This paper introduces the use of naïve Bayes for early gesture recognition. We show that the standard naïve Bayes learning algorithm equipped with a partial-information inference-test procedure, is capable of obtaining early-recognition performance that compares favorably against recent state of the art methods. The naïve baseline is also based on principled learning procedures and is very efficient (inference complexity of $O(q)$, with q the number of attributes). Besides, it is very simple and can be extended in a number of ways.

To the best of our knowledge, naïve Bayes has not been used previously for early recognition. However, a wide variety of improvements and extensions of naïve Bayes related to our proposal have been proposed (in parentheses we highlight differences with early recognition): (i) those for alleviating independence assumption of Naïve Bayes: i.e., modeling feature-dependencies for a subset of features [14,41,46] (but partial information cannot be processed); (ii) anytime formulations, i.e., for making predictions under time/resource constraints [19,43] (but the complete instance/object to be predicted is known in advance); (iii) incremental learning algorithms [2,23], (a sequence/stream of *complete* objects is processed, i.e., do not deal with objects described with partial information); and (iv) techniques for dealing with incomplete information [15,34,45], e.g., smoothing techniques, (attributes are not known sequentially, and can deal with a few missing values). The previous extensions aim at making naïve Bayes more robust against certain limitations, however, to the best of our knowledge, it has not been used for early classification before. This is somewhat surprising given that, as shown in the next section, naïve Bayes classifiers can naturally deal with partial information.

3. Early gesture/action recognition

This section describes the way we use naïve Bayes classifier for early gesture recognition. First we outline the bag of features representation, next we present the standard naïve Bayes classifier, finally, the early recognition procedures are discussed.

3.1. Bag of features formulation

The bag of features representation is one of most effective nowadays for several computer vision tasks, including gesture and action recognition [1,11,18,26,31,42]. This representation is the analogous to the bag of words representation widely used in text mining [33]. In fact, this analogy has motivated pretty much research on the use of text mining methodologies within computer vision, for instance: visual ngrams [3,24], visual phrases [47], visual weighting-schemes [12], and visual topic modeling [8]. At the same time, the computer vision community has evolved in a number of ways the bag of features representation, clear examples are VLAD [20], temporal bag of features [38], strings on *aclets* [4], etc.

The intuitive idea in the bag of features is to generate a codebook of features (playing the role of the vocabulary in text analysis), and represent each video/image with an histogram that accounts for the frequency of occurrence of elements of the codebook in the video/image [39]. The visual codebook is generally built by clustering descriptors extracted from the objects of interest, the centers of the clusters are considered the visual words (codewords). In this work, we adopt the bag of features formulation for representing sequences of video (see Section 4 for a description of the considered descriptors).

3.2. Naïve Bayes classifier

Consider a data set: $\mathcal{D} = (\mathbf{x}_i, y_i)_{\{1,\ldots,N\}}$ with N pairs of attributes $(\mathbf{x}_i)$ and labels (y_i) associated to a supervised classification problem. Assuming that $\mathbf{x}_i \in \mathbb{R}^q$ and $y_i \in C = \{1, \ldots, K\}$ is a K-class classification problem with numeric attributes. Under the Bayesian classifier, the probability of each class C_i for an unseen instance $\mathbf{x}_T = \langle x_{T,1}, \ldots, x_{T,q} \rangle$ is given by:

$$P(C_i|\mathbf{x}_T) \approx P(C_i)P(\mathbf{x}_T|C_i) \tag{1}$$

The class of instance $\mathbf{x}_T$ is given by $y_T = \arg\max_i P(C_i|\mathbf{x}_T)$. The assumption of naïve Bayes is that the probability of occurrence of attributes is independent given the class, that is:

$$P(C_i|\mathbf{x}_T) \approx P(C_i) \prod_{j=1}^{q} P(x_{T,j}|C_i) \tag{2}$$

The maximum likelihood estimation for the prior of class C_i is given by:

$$\hat{P}(C_i) = \frac{|X_i|}{N} \tag{3}$$

where X_i is the set of all instances in $\mathcal{D}$ that are labeled with class C_i. Hence the key of the naïve Bayes classifier lies in the estimation of $P(\mathbf{x}_T|C_i)$, or more precisely of $P(x_{T,j}|C_i)$. Depending on the type of data (e.g., binary, discrete, or real) a different probability distribution may be assumed for computing $P(x_{T,j}|C_i)$ (e.g., Bernoulli, Multinomial, or Gaussian, respectively).

In text classification one of the most effective estimates is based on the multinomial distribution [25]: i.e., each visual word is seen as an independent multinomial trial with r possible outcomes. Because we are using a bag of features representation (i.e., samples are represented by histograms) it makes sense to use the multinomial naïve Bayes classifier: the frequency of occurrence of each visual word is the number of outcomes of the multinomial trial. Thus, assuming a multinomial distribution for the model we have that the maximum likelihood estimation for the term of interest is:

$$P(\mathbf{x}_T|C_i) \approx \prod_{j=1}^{q} \hat{P}(x_{T,j}|C_i)^{f_{j,T}} \tag{4}$$

where $f_{j,T}$ is the value of the jth attribute in instance $\mathbf{x}_T$ (under the bag of features, it can be interpreted as the frequency of occurrence of the jth codeword in video T), and

$$\hat{P}(x_{T,j}|C_i) = \frac{1 + F_{j,C_i}}{q + \sum_k^q F_{k,C_i}} \tag{5}$$

where F_{l,C_i} is the sum of values of the lth attribute in objects of class C_i. The derivation from Eqs. (4) and (5) is straightforward assuming by assuming a multinomial distribution for the bag of features of a video. Please note that factorial terms that do not affect the final decision are removed. For our implementations we take logarithms to prevent the underflow problem. For more details we refer the reader to [25].

3.3. Early naïve Bayes

In early gesture recognition we assume that during training we have full videos/sequences, therefore, the same training procedure as the standard naïve Bayes classifier is performed for estimating

the necessary probabilities. The difference comes at inference time: when classifying a new sequence we process it in sequential order starting from the beginning. W.l.o.g.[3], at time t we assume we have read the first t-codewords[4] in the video (i.e., one codeword is read at each time). Let v_T denote the video sequence we want to classify, where it contains M_{v_T} words, then, $v_T = w_1, w_2, \ldots, w_{M_{v_T}}$. Please note that we are implicitly assuming that we know the start of a gesture, although this is a somewhat strong assumption, there are methods that allow us to identify the starting point of gestures, see our experiments and results on gesture spotting in Section 4.

We notice from Eqs. (3)–(5) that in fact we can predict the class of sequence v_T regardless of the amount of information we have read from it: at time t we know that $v_T = w_1, \ldots, w_t$, therefore, we can generate a bag of features $\mathbf{x}_T$ representation for v_T as follows $\mathbf{x}_T = \langle \mathbf{x}_{T,1}, \ldots, \mathbf{x}_{T,q} \rangle$, where $\mathbf{x}_{T,j}$ indicates the frequency of occurrence of the jth codeword in video sequence v_T. One should note that regardless of the number of codewords seen, the representation for $\mathbf{x}_T$ has length q: it is just the bag of features obtained with the codewords seen so far. Thus, terms not occurring in v_T or not seen so far at time t are assigned values of $\mathbf{x}_{T,j} = 0$. With this representation we can use Eq. (1) directly to classify the sequence. Actually, we can attempt to classify sequence v_T without having read any information at all! (i.e., with $t = 0$). Of course, the a posteriori probability will be dominated by the priors, see Eq. (3). Simply as this, we can use naïve Bayes to perform early classification.

We now briefly analyze what are the main components that come into play when making early prediction. At time t one can rewrite Eq. (4) as:

$$P(C_i|\mathbf{x}_T) \approx P(C_i) \prod_{j:j\in v_T} P(x_{T,j}|C_i) \prod_{k:k\notin v_T} P(x_{T,k}|C_i) \quad (6)$$

the second product accounts for the codewords that we have seen so far from the video (probabilities are affected by the frequency of occurrence of such terms in v_T so far); the third product accounts for terms not seen so far, this term resolves to 1, because the use of the frequency of the codeword in the video as power in Eq. (4). Clearly, for small values of t, the priors will dominate the decision, as t increases the content of the document will dominate the whole product. Therefore, the way these three components are estimated can be crucial for improving the performance of naïve Bayes in early classification.

Despite the simplicity of this early gesture recognition approach, we will see in the next section that it compares favorably with more complex solutions from the state of the art. We show its validity in three benchmark data sets. The main goal of this paper is to show that naïve Bayes can be used for early gesture recognition and that its performance is competitive with the best existing solutions to this problem. We foresee our work will pave the way for development of a new type of models for this problem.

4. Experiments and results

This section evaluates the performance of the naïve Bayes model in data sets that have been previously used for gesture and action recognition. First we describe the considered data sets, next we report experiments on early gesture recognition and finally we present results on gesture spotting.

Table 1
Data sets considered for experimentation. The number of training/test gestures, categories (K) and features (Feats.) is given. Cont. column indicates whether the data set contains continuous video. The +1 means the no-gesture/idle category.

Data set	Train	Test	K	Feats.	Cont.
MSRDaily3D [40]	192	48	16	600	No
MAD [18]	2,265	578	35 + 1	300	Yes
Montalbano [10]	10,304	3,579	20 + 1	2,000	Yes

4.1. Data sets

Because our interest is in interactive systems, the considered data sets are associated to tasks in which the user is near the sensor. Data sets for both, gesture recognition and human action analysis, task were used. The main characteristics of considered data sets are shown in Table 1, below we provide a detailed description. The three data sets were captured with the Kinect [37]. Two data sets have been used for spotting and one for recognition. In each data set, training videos are represented with their bag of features (using the whole information in the video), with this representation the naïve Bayes classifier is trained. Likewise, test sequences are represented by its bag of features, however, this time we generate bag of features representations that capture partial information from the sequence: at each time t the representation includes information from the features observed up to time t, see Section 3. Then we make predictions with the trained classifier for different times t. As explained below, for each data set we use visual descriptors that were used in previous work in order to compare our results, in every data set, we replicate the corresponding evaluation protocol.

4.1.1. MSRDaily3D

The MSRDaily3D data set comprises 16 actions associated to daily activities, where there are objects in the background and most actions involve human-object interaction. For comparison with previous work on early gesture recognition we consider 12 out of the 16 actions, as in [16–18]. Video sequences were represented with a bag of depth cuboid similarity features (DCSF), the same parameters for the descriptor as in [16,18,42] were used. A vocabulary of $q = 600$ words was obtained by applying $k-means$ to descriptors extracted from training sequences (50 visual words for each category, a preliminary study on the sensitivity to the value of q show that, although it is suboptimal, it represents a good tradeoff between performance and high dimensionality). Five-fold cross validation (over subjects) was used for evaluation.

4.1.2. MAD

The Multimodal Action Data set (MAD) was introduced in [18], it contains 35 different actions (plus the idle/no-action category). The data set was proposed to evaluate gesture spotting, therefore it is composed of sequences of continuous video containing gestures and no-gestures (idle). A total of 40 sequences of continuous video were recorded by 20 subjects. Each frame was represented by vectors of features extracted from the skeleton, the descriptor includes: bone angles between joints, differences between joints (the current and previous one and the current and 10th previous frame). Frame descriptors were clustered with k-means ($q = 300$), each video is represented by its bag of features. The choice of descriptor and its parameters were taken from previous work, in such a way that we used exactly the same descriptors and evaluation protocol as in [18].

[3] One should note that we can take steps of any length, instead of processing codeword-by-codeword.

[4] Usually, a codeword is associated to every frame of the video sequence, thus the t-codeword is associated to the tth-frame. Nevertheless, there are cases in which this does not hold, e.g., when using STIP-like descriptors.

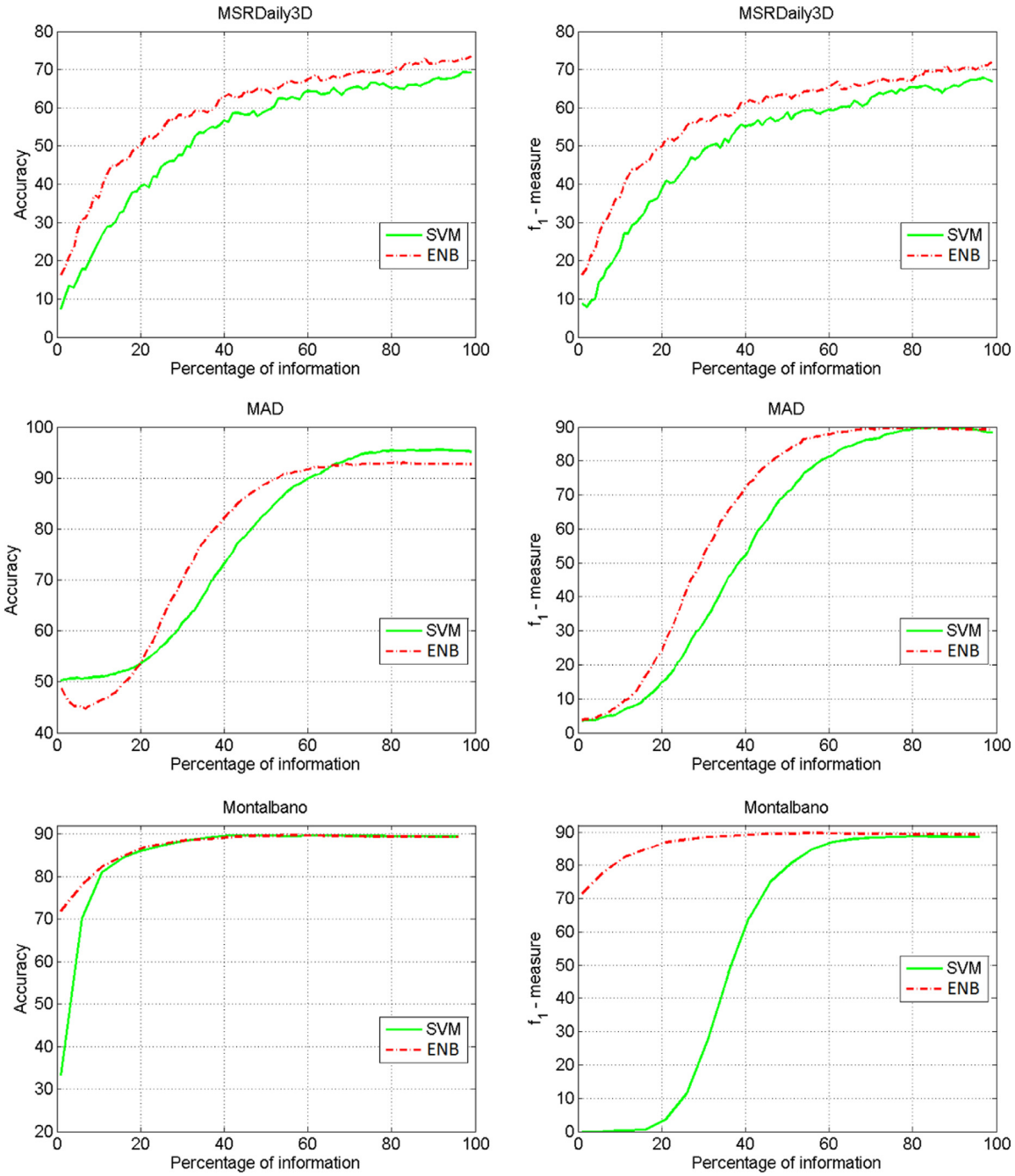

Fig. 2. Comparing the early recognition performance of our method (ENB) against support vector machines (SVM) for the three data sets. The left column reports accuracy and the right column f_1 measure.

4.1.3. Montalbano

The last data set we considered is Montalbano [10]. This data set has been used in two challenges on gesture recognition and spotting [10,11]. To the best of our knowledge, this is the largest data set for gesture recognition available nowadays. The task consist of recognizing gestures from 20 categories (Italian cultural gestures). The available data is depth and RGB video together with skeleton information. For our experiments we used the features proposed in [29], which combine depth, RGB video and skeleton information by means of convolutional nets and other deep learning mechanisms (For efficiency, we used a single channel out of the 4 that are used in [29,30]). Once each frame of each video is described with such features we learn a vocabulary of $q = 2000$ codewords. The choice for the value of q is justified by a previous study using this bag of features representation [12].

4.2. Recognition performance

First we evaluate the early recognition performance of early naïve Bayes (hereafter ENB) on segmented gestures for the three data sets. ENB was trained on the complete sequences and tested on sequences of increasing size: containing information from 1% to 100% of the sequences (step sizes of 1% were used). We report

Table 2
Early recognition results in the MSRDaily3D data set.

Segs./ Method (%)	[16] (%)	[18] (%)	[42] (%)	Ours (%)
[0–20]	50.4			**56.2**
[0–40]	63.8			**72.9**
[0–60]	65.8	73.2		**75.0**
[0–80]	68.8			**79.2**
[0–100]	68.3		83.6*	**79.2**

the recognition performance in terms of f_1 measure (leftmost column) and accuracy (rightmost column) for all of the data sets and different sizes in Fig. 2. For reference, we show the performance of a linear SVM trained and tested in the same conditions as ENB (below we provide a comparison with other state of the art techniques).

It can be seen from these plots that ENB outperforms SVM in most configurations. For MSRDaily3D (top plots), the improvement is consistent across different percentages of information and under both f_1 measure (left) and accuracy (right). For the MAD data set (mid plots), the improvement in f_1 measure of ENB over SVM is maintained up to near the end of gestures, whereas for accuracy, SVM outperforms ENB after 65% information has been read. This result, suggest SVM is making accurate predictions for majority-class gestures when enough information is given. However, this behavior is not necessarily desired when having a wide variety of actions (in this case there are 35 different actions). On the other hand, it is interesting that for the MAD data set ENB has an erratic behavior in terms of f_1 measure when less than 20% of information has been received (mid, leftmost plot). This behavior can be due to the fact that with very few information ENB relies to much on the priors (this explains the *normal* behavior in accuracy), which may affect its performance. Regarding Montalbano data set, ENB outperforms SVM in every setting, it is interesting to see that ENB has a high "jump-start": with 20% of the information we can classify about 85% of all of the gestures. This result can be due to the effectiveness of the features that were learned for this particular data set [29].

In general we can say that ENB outperforms SVM (most notably in terms of f_1). Also, it can be said that overall with about 60% of the information of gestures, we can make predictions with acceptable performance. Regarding efficiency, recognition for all of the considered data sets can be performed loosely in real time as the inference process of ENB is quite fast. In the following we report a per-data set analysis of results and comparisons with related work.

4.3. Results on MSRDaily3D

Early recognition results for the MSRDaily3D data set are shown in Table 2. For these results, we classified actions with the same interval predictions reported in previous work [16,18], see column 1.

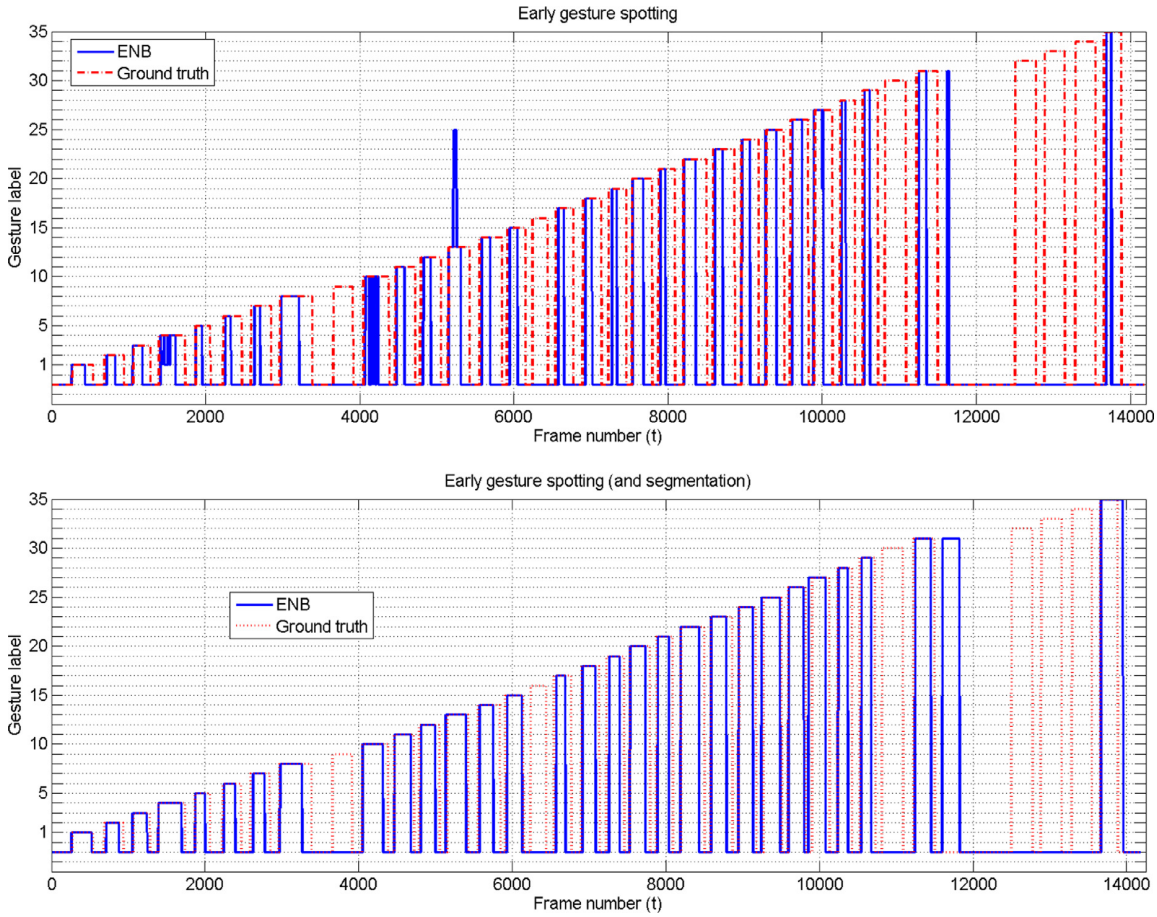

Fig. 3. ENB performance in a randomly selected test-sequence. *x*-axis denotes time (No. of processed frames), the *y*-axis denotes the label of the gesture (in MAD each sequence contains the 35 gestures performed in order); the idle/no-gesture category is indicated with $y = -1$. Ground-truth is shown with the red-dashed line, the predictions of ENB are shown in blue-solid line. In a perfect prediction, the solid line should cover the dashed one. Right: early gesture spotting (no attempt is made to detect the end of a gesture). Left: early gesture spotting and segmentation (the end of the gesture is detected). (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

One should note that for this experiment, the methods in [16,18] require that gestures are segmented in intervals, whereas ours does not: we have made predictions at such intervals for comparison, but ENB can make predictions at any time. We also include the result from [42], however, please note that such results were obtained using the 100% of information (i.e., no early prediction).

It can be seen from Table 2 that the best results were obtained with the proposed naïve baseline. ENB outperforms [16] with every amount of considered information. Besides, it obtained comparable performance to the SMMED method in [18] when reading only about 40% of the sequences (after reading 60% our method consistently outperforms SMMED). One should note that ENB classifies every gesture using the percentage of information indicated in column 1 of Table 2, whereas SMMED only was capable of early-classifying around 39% of the gestures/samples (for the rest the whole sequence of video was used). It is important to emphasize that we have used exactly the same data partitions, descriptor and settings as in previous work, thus our results are directly comparable to [16,18].

Finally, it is worth emphasizing that our results are very close to the best recognition (i.e., no early-classification) result reported for this data set when using the same descriptor (column 4) [42]. Therefore, naïve Bayes not only is appropriate for early recognition, but also for the standard gesture recognition task.

4.4. Results on MAD

In Fig. 2 we already reported recognition results on isolated sequences, in this section we evaluate the performance of ENB in continuous video. For gesture spotting we extended the ENB method as follows. First, we train a two-class naïve Bayes model for distinguishing gestures from no-gestures, where the complete sequences were used for training this classifier. At testing time, we process the sequence of continuous video with a sliding window of size ρ, we used ρ equal to half the average length of training gestures. Intuitively, we wanted to recognize a gesture by using 50% of information. Every window is classified with both the binary and multiclass ENB classifiers. Let p_g and p_n denote the probabilities of a window containing or not a gesture, respectively. Whenever $(p_g - p_n)$ exceeds a threshold we record the prediction of the multiclass classifier for that window. Whenever $(p_n - p_g)$ exceeds another threshold, we combine the predictions of all of the windows up to the previous one (i.e., we detected the end of a gesture). Both thresholds are learned from training videos. In this way we are able to perform early recognition and spotting.

Table 3
Early recognition spotting results in the MAD data set.

Segs./ Method	[16] (%)	[18] (%)	Ours (%)
Precision	28.7	59.2	**76.1**
Recall	51.4	57.4	**73.6**
f_1 measure	36.8	58.2	**74.8**

To better appreciate the early recognition performance in spotting, it is worth evaluating the performance of the early recognition mechanism alone (i.e., without attempting to detect the end of gestures). Fig. 3 shows a comparison of the recognition mechanisms with and without trying to detect the end of the gesture. It can be seen from the top plot, that ENB correctly recognizes most of the gestures in the sequence, by using around 50% of the information. Although part of the sequence cannot be correctly detected, it mostly corresponds to the second part of gestures, which is expected because we do not attempt to segment gestures.

On the other hand, if we attempt to detect the end of gestures (see bottom plot in Fig. 3), it is clear that the performance of ENB is very close to the ground truth. This figures show evidence of the usefulness of ENB for early spotting, in the following we compare the performance of ENB with reference methods in the task of gesture spotting.

Table 3 shows a quantitative evaluation and comparison of ENB. For this evaluation we report the average precision and recall, as reported in [18]: a gesture is considered correctly detected if it overlaps with at least 50% of the ground truth gesture. A false positive is counted whenever the number of incorrect predicted frames overpass 50% of the average length of training gestures. The reference methods, MMED and SMMED were extended to deal with continuous video and perform spotting as described in [16,18]. It is clear from Table 3 that our method outperforms MMED [16] and SMMED [18] approaches. The difference in performance is considerable. The absolute improvement in terms of f_1 measure is of more than 15%. This is a very interesting result: the naïve baseline, outperforms max-margin based models.

A final qualitative comparison is shown in Fig. 4. We reproduced the plots from Figure 6 in [18], and generated similar plots from predictions obtained with our method. The color codes the actual label of each gesture, GT depicts the ground truth. It can be seen from these plots that our method indeed performs

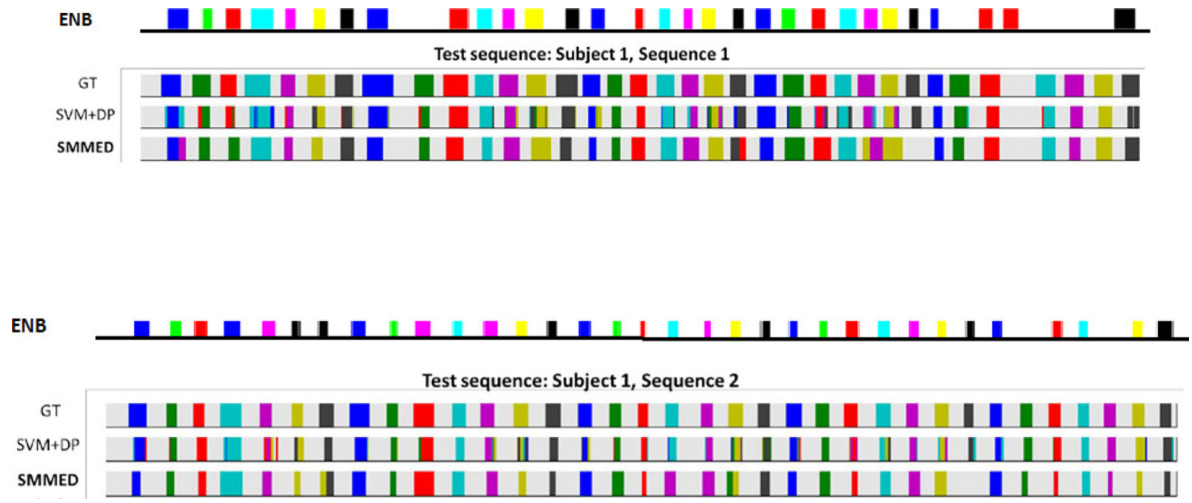

Fig. 4. Qualitative comparison of ENB with the methods reported in [18]. We reproduce here the figures from [18].

Table 4
Recognition-spotting results in the Montalbano data set. Undisclosed: no paper published about this method, see [10].

Method	Overlap (%)
[30]	**85.0**
[27]	83.4
[7]	82.7
[13]	74.5
[6]	74.7
Undisclosed	68.9
Ours	77.0

very close to the ground truth. Although it misses a few gestures, in most cases the correct gesture is detected and there are not incorrectly classified frames within a detected gesture. In our opinion, ENB provides better predictions than the reference techniques.

4.5. Montalbano data set

Recognition results on isolated gestures for this data set were reported previously, in this section we analyze spotting performance. A similar spotting mechanism as that used for the MAD data set has been adopted (i.e., gesture vs. no-gesture classifier, combined with a multiclass ENB). Results of this experiment are shown in Table 4. For this data set the (frame-level) overlap measure has been mostly used (see, e.g., [10]), accordingly we evaluate our method with such measure.

From Table 4 it can be seen that the ENB approach is competitive with state of the art methods in the Montalbano data set. An overlap of $\approx$ 77% is acceptable for a number of applications, the difference between the top method and ours is of around 8%, but ENB can provide anticipated predictions. It is worth mentioning that because Montalbano data set was released in the context of two academic challenges, many methods have been using it, we report in this paper only the best performing methods. See [10,11,30] for further details. Also, it is important to emphasize that our method performs online gesture recognition, that is, it could be used directly in a real gesture recognition application (e.g., human robot interaction). However, most of the reference methods perform spotting in an offline setting: they analyze the whole sequence of video to segment the video [10].

5. Conclusions

We introduced a new baseline for early gesture recognition. The proposed method takes advantage of naïve Bayes cumulative-evidence property and adapts it to gesture recognition and spotting. A comparison with state of the art methods in standard data sets shows that *Early naïve Bayes* compares favorably with a number of more complex methodologies that have been specifically designed for early gesture recognition. It is important to emphasize that we have obtained these outstanding results with a kind-of straight implementation of naïve Bayes. Therefore, even better results are expected when incorporating improvements/extensions/adaptations of this simple classifier. Future work includes: improving the early-predictive capabilities of naïve Bayes in several ways. Specifically: one can define adaptive priors that change as the value of t increases; we can implement the same idea with methods that take into account term-dependencies (see e.g., [41,46]) in order to increase the predictive power of the classifier; also one can adopt advanced/alternative smoothing techniques to account for partial and missing information properly [5,34].

Acknowledgments

CONACyT – project grant no. CB-2014-241306 (Clasificación y Recuperación de Imágenes Mediante Técnicas de Minería de Textos) and CONACyT - project grant no. PN-215546 *(ViVA: Video vigilancia automática: hacia un sistema genérico de análisis inteligente de videos)*.

References

[1] J.K. Aggarwal, M.S. Ryoo, Human activity analysis: A review, ACM Comput. Surv. 43 (2011).
[2] J.R. Alcobé, Incremental learning of tree augmented naive Bayes classifiers, in: Proceedings of the IBERAMIA'02, Springer, 2002, pp. 32–41.
[3] V. Bettadapura, G. Schindler, T. Plotz, I. Ess, Augmenting bag-of-words: data-driven discovery of temporal and structural information for activity recognition, in: Proceedings of the CVPR'13, 2013, pp. 1–8.
[4] L. Brun, G. Percanella, A. Saggese, M. Vento, Action recognition by using kernels on aclets sequences, Comput. Vis. Image Underst. (2015), doi:10.1016/j.cviu.2015.09.003.
[5] J.M. Cabrera, H.J. Escalante, M.M. y Gómez, Distributional term representations for short-text categorization, in: Proceedings of the CICLING, 2013.
[6] N.C. Camgoz, A.A. Kindiroglu, L. Akarun, Gesture recognition using template based random forest classifiers, in: Proceedings of the ECCV- Workshops, 2014.
[7] J.Y. Chang, Nonparametric gesture labeling from multi-modal data, in: Proceedings of the ECCV- Chalearn Workshop, 2014.
[8] A. Cruz, E. Díaz, F. González, Automatic annotation of histopathological images using latent topic model based on nonnegative matrix factorization, Int. J. Pathol. Inf. 2(4) (2011).
[9] C. Ellis, S.Z. Masood, M.F. Tappen, J.J. Laviola Jr., R. Sukthankar, Exploring the trade-off between accuracy and observational latency in action recognition, Int. J. Comput. Vision 101 (2013) 420–436.
[10] S. Escalera, X. Baro, J. Gonzalez, M.A. Bautista, M. Madadi, M. Reyes, V. Ponce, H.J. Escalante, J. Shotton, I. Guyon, ChaLearn looking at people challenge 2014: Dataset and results, in: Proceedings of the ECCV Workshops, 2014.
[11] S. Escalera, J. Gonzalez, X. Bar, M. Reyes, O. Lopes, I. Guyon, V. Athistos, H.J. Escalante, Multi-modal gesture recognition challenge 2013: dataset and results, in: Proceedings of the ICMI '13, 2013, pp. 445–452.
[12] H.J. Escalante, J. Marínez-Carranza, S. Escalera, V. Ponce-López, X. Baró, Improving bag of visual words representations with genetic programming, in: Proceedings of the IJCNN 2015, 2015, pp. 3674–3681.
[13] G.D. Evangelidis, G. Singh, R. Horaud, Continuous gesture recognition from articulated poses, in: Proceedings of the ECCV- Chalearn Workshop, 2014.
[14] N. Friedman, D. Geiger, M. Goldszmidt, Bayesian network classifiers, Mach. Learn. 29 (1997) 131–163.
[15] F. He, X. Ding, Improving naive Bayes text classifier using smoothing methods, in: Proceedings of the of European conference on IR research, 2007, pp. 703–707.
[16] M. Hoai, Z. Lan, F.D. la Torre, Joint segmentation and classification of human actions in video, in: Proceedings of the CVPR, 2011.
[17] M. Hoai, F.D. la Torre, Max-margin early event detectors, in: Proceedings of the CVPR, 2012.
[18] D. Huang, S. Yao, Y. Wang, F.D.L. Torre, Sequential max-margin event detectors, in: Proceedings of the ECCV, 2014.
[19] B. Hui, Y. Yang, G.I. Webb, Anytime classification for a pool of instances, Mach. Learn. 77 (2009) 61–102.
[20] H. Jegou, M. Douze, C. Schmid, P. Perez, Aggregating local descriptors into a compact image representation, in: Proceedings of the CVPR2010, 2010, pp. 3304–3311.
[21] M. Kawashima, A. Shimada, H. Nagahara, R.I. Taniguchi, Adaptive template method for early recognition of gestures, in: Proceedings of the Joint Workshop on Frontiers of Computer Vision (FCV), 2011 17th Korea-Japan, 2011, pp. 1–6.
[22] D. Kim, J. Song, D. Kim, Simultaneous gesture segmentation and recognition based on forward spotting accumulative HMMs, Pattern Recognit. 40 (2007) 3012–3026.
[23] F. Klawonn, P. Angelov, Evolving extended naïve Bayes classifiers, in: Proceedings of the of 6th Intl. Conference on Data Mining Workshops, 2006, pp. 643–647.
[24] P. López, M. Montes, H.J. Escalante, A. Cruz, F. González, Improving the BoVW with discriminative n-grams and MKL, Neurocomputing (2015).
[25] A. McCallum, K. Nigam, A comparison of event models for naive bayes text classification, in: Proceedings of the AAAI/ICML-98 Workshop on Learning for Text Categorization, AAAI, 1998, pp. 41–48.
[26] S. Mitra, T. Acharya, Gesture recognition: a survey, Trans. Syst. Man Cybern. C 37 (2007) 311–324.
[27] C. Monnier, S. German, A. Ost, A multi-scale boosted detector for efficient and robust gesture recognition, in: Proceedings of the ECCV workshops, 2014.
[28] A. Mori, S. Uchida, R. Kurazume, R.I. Taniguchi, Early recognition and prediction of gestures, in: Proceedings of the ICPR, 2006, pp. 560–563.
[29] N. Neverova, C. Wolf, G.W. Taylor, F. Nebout, Multi-scale deep learning for gesture detection and localization, in: Proceedings of the ECCV 2014 Workshops, 2014.

[30] N. Neverova, C. Wolf, G.W. Taylor, F. Nebout, ModDrop: Adaptive Multi-Modal Gesture Recognition, Technical Report, 2015 arXiv:1501.00102.
[31] J. Niebles, H. Wang, L. Fei-Fei, Unsupervised learning of human action categories using spatial-temporal words, Int. J. Comput. Vis. 79 (2008) 299–318.
[32] Y. Sabinas, E. Morales, H.J. Escalante, A one-shot DTW-based method for early gesture recognition, in: Proceedings of the CIARP, 2013, pp. 439–446.
[33] F. Sebastiani, Machine learning in automated text categorization, ACM Comput. Surv. 34 (2008) 1–47.
[34] D. Shen, J. Wu, B. Cao, J. Sun, Q. Yang, Z. Chen, Y. Li, Exploiting term relationship to boost text classification, in: Proceedings of the CIKM, 2009.
[35] A. Shimada, M. Kawashima, R. Taniguchi, Improvement of early recognition of gesture patterns based on a self-organizing map, Artif. Life Robot. 16 (2011) 198–201.
[36] A. Shimada, M. Kawashima, R.R. Taniguchi, Early recognition based on cooccurence of gesture patterns, in: Proceedings of the ICONIP, 2010, pp. 431–438.
[37] J. Shotton, A. Fitzgibbon, M. Cook, T. Sharp, M. Finocchio, R. Moore, A. Kipman, A. Blake, Real-time human pose recognition in parts from a single depth image, in: Proceedings of the CVPR, 2011.
[38] P. Shukla, K. Biswas, P. Kalra, Action recognition using temporal bag-of-words from depth maps, in: Proceedings of the MVA, 2013, pp. 41–44.
[39] J. Sivic, A. Zisserman, Video Google: a text retrieval approach to object matching in videos, in: Proceedings of the ICCV, 2003, pp. 1470–1477.
[40] J. Wang, Z. Liu, Y. Wu, J. Yuan, Mining action let ensemble for action recognition with depth cameras, in: Proceedings of the CVPR, 2012, pp. 1290–1297.
[41] G. Webb, J.R. Boughton, Z. Wang, Not so naive Bayes: aggregating one-dependence estimators, Mach. Learn. 58 (2005) 5–24.
[42] L. Xia, J.K. Aggarwal, Spatio-temporal depth cuboid similarity feature for activity recognition using depth camera, in: Proceedings of the CVPR, 2013, pp. 2834–2841.
[43] Y. Yang, G.I. Webb, K. Korb, K. Ting, Classifying under computational resource constraints: anytime classification using probabilistic estimators, Mach. Learn. 69 (2007) 35–53.
[44] G. Yu, J. Yuan, Z. Liu, Predicting human activities using spatio temporal structure of interest points, in: Proceedings of the ACM Multimedia, 2012.
[45] Q. Yuan, G. Cong, N.M. Thalmann, Enhancing naive Bayes with various smoothing methods for short text classification, in: Proceedings of the WWW, 2012.
[46] N. Zaidi, J. Cerquides, M. Carman, G. Webb, Alleviating nb attribute independence assumption by attribute weighting, JMLR 14 (2013) 1947–1988.
[47] Q. Zheng, W. Wang, W. Gao, N.M. Thalmann, Effective and efficient object-based image retrieval using visual phrases, in: Proceedings of the ACM Multimedia, 2006, pp. 77–80.