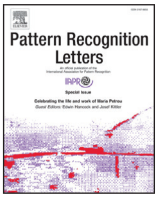
Pattern Recognition Letters 000 (2017) 1–9

Contents lists available at ScienceDirect

Pattern Recognition Letters

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


Pattern Recognition Letters

Active Incremental Recognition of Human Activities in a Streaming Context

Rocco De Rosa[a,*], Ilaria Gori[b], Fabio Cuzzolin[c], Nicolò Cesa-Bianchi[d]

[a] Rank Group, Data Science Lab, UK
[b] iCub Facility, Istituto Italiano di Tecnologia, Italy
[c] Department of Computing and Communication Technologies, Oxford Brookes University, UK
[d] Dipartimento di Informatica, Università degli Studi di Milano, Italy

ARTICLE INFO

Article history:
Available online xxx

Keywords:
Activity recognition
Streaming video
Active learning
Incremental learning
Constant Budget
Continuous action recognition
Non-parametric classification

ABSTRACT

Recognising human activities from streaming sources poses unique challenges to learning algorithms. Predictive models need to be scalable, incrementally trainable, and must remain bounded in size even when the data stream is arbitrarily long. In order to achieve high accuracy even in complex and dynamic environments methods should be also nonparametric, i.e., their structure should adapt in response to the incoming data. Furthermore, as tuning is problematic in a streaming setting, suitable approaches should be parameterless (as initially tuned parameter values may not prove optimal for future streams). Here, we present an approach to the recognition of human actions from streaming data which meets all these requirements by: (1) incrementally learning a model which adaptively covers the feature space with simple and local classifiers; (2) employing an active learning strategy to reduce annotation requests; (3) achieving good accuracy within a fixed model size. Although in this work we focus on human activity recognition, our approach is completely independent from the feature extraction and can deal with any supervised matrix (set of feature vectors). Hence, it can be adapted to a wide range of applications (e.g., speech recognition, image classification, object recognition, pose recognition, and image matching). Extensive experiments on standard benchmarks show that our approach is competitive with state-of-the-art non-incremental methods, while outperforming the existing active incremental baselines.

1. Introduction

The pervasive presence of cameras and mobile devices in our everyday lives has created a strong demand for automated methods able to analyse data streams in real time. This is especially challenging in the case of videos capturing human activities, as in TV footages and videos from surveillance cameras. Another natural application is human robot interaction, which requires the machine to learn and recognise human behavioural patterns in real time. Nevertheless, the mainstream approaches to action and activity recognition are typically based on an offline training phase (for a review of previous work in activity recognition we refer the reader to Sec. 2). Such a setting leads to several critical issues when dealing with streaming videos:

How to incrementally learn activities from the incoming data? The dynamic nature of the streaming video setting implies that, at each time instant, new data is made available to the system, which needs to incrementally learn from it. This implies both refining the current models of known human activities and adding on the fly new models of previously unseen activities.

How to minimise the required annotation effort? The issue of how many video fragments should be annotated is strictly related to the ability of learning new activities. While for newly observed activities one might assume that all video frames should be —at least initially— manually annotated, when analysing footage of known action classes only a fraction of the video input will likely bring in new information. In this context, the system should automatically select which video fragments are the most informative, and asks human annotators for help only in those cases.

How to optimise the algorithm heuristics dynamically? System components, such as the chosen feature representation and the learning algorithm parameters, have a crucial impact on the final performance of any framework. In a continuous learning setting, however, design choices and parameter tuning are —if possible at all— more difficult than in offline settings, as we just cannot anticipate what new activities the system will be asked to learn.

[*] Corresponding author:
E-mail addresses: rocco.derosa1982@gmail.com, Rocco.DeRosa@rankinteractive.com (R. De Rosa).

http://dx.doi.org/10.1016/j.patrec.2017.03.005
0167-8655/© 2017 Elsevier B.V. All rights reserved.

The main contribution of this paper is an approach for dealing with human activity recognition in a streaming context. To the best of our knowledge, our approach is the first one to address all the challenges listed above in a principled manner. Our starting point is a recently proposed local algorithm for classification of data streams [5] which is incrementally trainable and nonparametric (i.e., the model structure is not specified a priori, but determined by the data in such a way that the number of parameters is not fixed in advance), while exhibiting theoretical guarantees on its performance. Here we leverage on this result, and extend it to the active learning setting. This leads to a framework that meets all the above requirements: (1) it incrementally and efficiently learns the incoming data stream while being robust to the addition of new classes; (2) the active learning component evaluates the informative content of the incoming data items with respect to the current level of confidence, thus allowing to decide when the cost of manual annotation is worthwhile; (3) the nonparametric nature of the approach allows for fully data-driven learning.

Next, we illustrate the workflow of our approach in the specific case of activity recognition from streaming videos:

1. Each video is associated with a variable number of feature vectors in a given feature space.
2. The feature space gets sequentially covered with balls centered on samples selected from the stream.
3. Each ball is associated with an estimate of the conditional class probabilities obtained by collecting statistics around its centre; a new unlabeled sample is predicted using the estimate of the closest ball.
4. A sample falling outside its closest ball becomes the center of a new ball.
5. The radius of each ball is adjusted according to how well each ball predicts the class label of the new samples that fall close to it.
6. Ball centres are incrementally adjusted to fit the actual data distribution.
7. The set of balls is organized in a tree-like structure [17], so that the ball nearest to the current sample can be found in time logarithmic in the number of balls.

We call our algorithm Fast active Incremental Visual covERing (FIVER). Extensive experiments on several publicly available databases show that our approach outperforms all existing algorithms for activity recognition from streaming data. Furthermore, we show that by combining FIVER with the robust temporal segmentation algorithm presented in [7], we obtain a system able to deal, in a straightforward manner, with a realistic continuous active recognition scenario. A significant contribution of this work is the extension of [5] to an active learning setting. This is key to the practical application of incremental learning in streaming settings for at least two reasons. Firstly, active learning systems allow to substantially save on costly ground truth annotations. Secondly, the confidence score plays a crucial role in continuous activity recognition tasks in domains such as surveillance and human-robot interaction (see Sec. 4.5).

2. Related Work

Within the vast literature related to action recognition — see [25] and references therein— research focusing on the streaming setting has gained momentum only recently [8]. Desirable features in this context are: *(1) Incremental Updating:* typically, a large amount of data is sequentially presented in a stream, and so it is desirable for algorithms to incrementally update the model rather than re-training it from scratch. *(2) Incremental Learning of New Classes (activities):* algorithms should be able to accommodate on the fly any new class. *(3) Bounded Size Models:* as the data stream may be very large, models should keep a bounded memory footprint, allowing for real-time prediction while avoiding storage issues. This implies the ability of discarding useless or old data, and is critical to the tracking of *drifting concepts* (i.e., settings where the optimal decision surface changes over time, requiring repeated adjustments of the model [28]). *(4) Data-Driven behaviour:* because parameter tuning is problematic in streaming settings, systems with few or no parameters are preferable. *(5) Nonparametric behaviour:* since the true structure of the data is progressively revealed as more examples from the stream are observed, nonparametric algorithms [10], which are not committed to any specific family of decision surfaces, are preferable. *(6) Active Learning:* in a streaming setting, the system needs to learn from each incoming data point. However, training labels are provided by human annotators, who should be invoked only when the system has low confidence in its own prediction for the current label. *(7) Bounded Request rate:* since querying human annotators is expensive, any practical active learning system for streaming settings should impose a bound on the query rate.

Table 1 lists the previous efforts in human activity recognition involving incremental and/or active learning components which, due to their features, are the closest alternatives to our approach.

A feature tree-based incremental recognition approach was proposed in [26], where the tree is free to grow without bounds as more examples are fed to the learner. As this requires to store all the presented instances, the method is infeasible for continuous recognition from streaming videos, where the number of activities can get very large over time. A human tracking-based incremental activity learning framework was proposed in [23] which, however, requires annotation on the location of the human body in the initial frame, heavily restricting its applicability. For these reasons [26] and [23] are not listed in Table 1. Our work shares similarities with the incremental algorithm in [3] upon which, to some extent, we build our proposal. Both methods adopt a nonparametric, incremental ball covering of the feature space strategy. FIVER, however, brings to the table crucial new features that makes it uniquely suitable for dealing with streaming data. Firstly, it does not rely on any input parameters, which are inconvenient to tune in streaming settings. Secondly, it limits the model size, thus allowing the tracking of drifting concepts. More precisely, when the number of allocated balls exceeds a given budget, FIVER discards each ball with a probability proportional to its error rate. Thirdly, it dynamically adjusts the ball centres, thus yielding very compact models while improving performance. The resulting covering resembles a visual dictionary, learned incrementally and directly usable for predictions, where the balls play the role of visual codewords. Finally, the active learning module defines the interaction between the learning system and the labeler agent, limiting the number of annotations requested.

The use of incremental active learning for activity recognition tasks was recently investigated in [11,12], where an ensemble of linear SVM classifiers is incrementally created in a sequence of mini-batch learning phases. These methods, however, are not designed to operate on individual data elements in streams, as it is required in our setting. A confidence measure over the SVM outputs is defined, where each individual classifier output is weighted by the training error. Two user-defined thresholds control the query rate of labeled videos. Non-confident instances, which are close to a class boundary, are forwarded to the annotator, while the others are discarded. Note that the set of ensemble classifiers can become very large, as an arbitrary number of SVMs can be added in each batch phase. Furthermore, the method requires model initialisation, and several parameters need to be tuned at validation time, thus making the approach unsuitable to a truly streaming context. The method in [11] initially learns features in an unsupervised manner using a deep neural network. Then, a multinomial

Please cite this article as: R. De Rosa et al., Active Incremental Recognition of Human Activities in a Streaming Context, Pattern Recognition Letters (2017), http://dx.doi.org/10.1016/j.patrec.2017.03.005