Learning from uneven video streams in a multi-camera scenario

1 Samaneh Khoshrou
sikhoshrou@inescporto.pt
2 Jaime S. Cardoso
jaime.cardoso@inescporto.pt
3 Luís F. Teixeira
luistf@fe.up.pt
4 INESC TEC
Porto, Portugal
5 Faculdade de Engenharia
Universidade do Porto

Abstract

We present a semi-supervised incremental learning algorithm for evolving visual data in order to develop a robust and flexible track classification system in a multi-camera surveillance scenario. Most existing methods, which are variations on static learning schemes, cannot cope with many real-life challenges. The scarcity of labelled data in real applications ends up generating poor classifiers. Furthermore, labelling the whole data (possibly massive in such applications) imposes a high cost to the system, rendering the technology impractical. So, there is an increasing interest on exploiting unlabelled data.

Our proposed method learns from consecutive batches by updating an ensemble in each time. It tries to achieve a balance between performance of the system and amount of data which needs to be labelled. As no restriction is considered, the system can address many practical problems in an evolving multi-camera scenario, such as concept drift, class evolution and various length of video streams which have not been addressed before.

1 Introduction

Over the last decades, video surveillance began to spread rapidly, specifically targeted at public areas. Recording for hours, days, and possibly years provides a massive amount of information coming from an evolving environment in which traditional learning methods fail to reflect the evolution taking place. In such environments, the underlying distribution of data changes over time - often referred to as concept drift - either due to intrinsic changes (pose change, movement, etc.), or extrinsic changes (lighting condition, dynamic background, complex object background, changes in camera angle, etc.). Thus, models need to be continually updated to represent the latest concepts. The problem is further aggravated when new objects enter the scene - referred to as class evolution in machine learning literature - as new models need to be trained for the novel classes.

Figure 1 demonstrates a typical surveillance scenario. Depending on the angle of the view and quality of the camera, every surveillance camera covers an area called Field of View (FoV). Often the fields of view are disjoint due to budget constraints, whereas they overlap in some scenarios. When entering the scene, the object will enter the coverage area of at least one of the cameras. The surveillance system will have to track that object from the first moment it was captured by a camera and across all cameras whose field of view overlaps the object’s path. In such environments where objects move around and cross the FOV of multiple cameras, it is more likely to have multiple streams, potentially overlapping in time, recorded at different starting points with various lengths, for the same individual object (Figure 1).

Yet, much of the learning literature is concerned with a stationary environment, where fixed and known number of categories to be recognized [6] and enough resources (labelled data, memory and computational power) are available [4]. Learning from time-changing data has mostly appeared in data mining context and various approaches have been proposed. Ensemble-based approaches constitute a widely popular group of these algorithms to handle concept drift [1] and in some recent works class evolution [3], as well. Learn++.NSE [3] is one of the latest ensemble-based classification methods in literature, that generates a classifier using each batch of training data and applies a dynamic weighting strategy to define the share of each ensemble in the overall decision. As success is heavily dependent on labelled data, this method would not be applicable in wild scenarios. Although a considerable body of research has emerged from stream mining, learning from multiple streams in wild environments, that views whole or segments of a stream as a unique element to cluster (or classify), is a less explored area. The methods that have been proposed [2], require equal length streams coming from a fixed number of sources. Thus, they would fail to leverage information from time-varying video tracks. An effective and appropriate algorithm to fit in our scenario is required to: a) learn from multiple streams; b) mine streams with various length and starting points (uneven streams); c) handle the concept drift; d) accommodate new classes; e) deal with partially labelled or unlabelled data; f) be of limited complexity; g) handle multi-dimensional data.

To the extent of our knowledge, no such method has been introduced in the literature.

2 Learning algorithm

In this section we present our Never Ending Visual Information Learning (NEVIL) framework. NEVIL is designed for non-stationary data environments in which no labelled data is available but the learning algorithm is able to interactively query the user to obtain the desired outputs at carefully chosen data points. The NEVIL algorithm is an ensemble of classifiers that are incrementally trained (with no access to previous data) on incoming batches of data, and combined with some form of weighted majority voting. A sketch of the proposed method is shown in Figure 2. A typical tracking algorithm analyses sequential video frames and outputs the movement of targets between the frames, generating multiple streams of visual data. Environmental challenges such as varying illumi-
nation, lack of contrast, bad positioning of acquisition devices, blurring caused by motion as well as occlusion make data often noisy and/or partially missing. We address these challenges by a batch divisive strategy, as learning from a data batch may reduce the noise and fill the gap caused partially missing. Note that a stream corresponds to a track generated by the tracking system and a single camera can yield multiple streams. A single batch aggregates \( B \) frames. The starting time of each stream is potentially different from stream to stream but batches are aligned between streams. Inside each frame the data corresponds to some pre-selected object representation (e.g. bag of words, histogram) and is out of the scope of this paper.

The ensemble obtained by all models generated up to the current time slot \( T_S \) is named the composite hypothesis \( H_{t-1} \). With the arrival of the current data batches \( D^n_t \), \( i = 1 \cdots M \), NEVIL tries to predict the class label for each of the batches in current \( T_S \) based on the probability estimate \( p(C_i|D^n_t, H_{t-1}) \), where \( C_i \) runs over all the class labels observed so far.

This kind of online-learning approach addressed in this work can suffer if labelling errors accumulate, which is inevitable. Unrelated objects will sooner or later be assigned the same label or different labels will be assigned to different views of the same object. To help mitigate this issue, we allow the system to interact with a human, to help it stay on track. Initially, the composite model is initialized to yield the same probability to every possible class (uniform prior). When the batches \( D^n_t \) in time slot \( t \) become available, NEVIL starts with computing the probabilities \( p(C_i|D^n_t, H_{t-1}) \) for each batch \( D^n_t \) in the time slot. Once \( p(C_i|D^n_t, H_{t-1}) \) are obtained, a batch confidence label (BCL) is estimated; if BCL is high (above a presupervised threshold), the predicted label

\[
\arg \max_{C_i} p(C_i|D^n_t, H_{t-1})
\]

is accepted as correct, otherwise the user is requested to label the data batch. The labelled batches (either automatically or manually) are used to generate a new multiclass classifier that is integrated in the composite model, yielding \( H_{t+1} \).

3 Experimental Methodology

3.1 Datasets

A series of experiments were conducted to explore the capabilities of the proposed framework. We tested the NEVIL framework with real video data from CAVIAR dataset [5] including: OneLeaveShopReenter1, EnterExitCrossingPaths1, OneShopOneWait1, OneStopEnter2 and WalkByShop1/front. Due to the presence of different perspectives of the same person, streams are drifting in time (see Fig. 3). These sequences present challenging situations with cluttered scenes, high rates of occlusion, different illumination conditions as well as different scales of the person being captured. We employ an automatic tracking approach to track objects in the scene and generate streams of bounding boxes, which define the tracked objects’ positions. As the method may fail to perfectly track the targets, a stream often includes frames of distinct objects [7]. An hierarchical bag-of-visters method is applied to represent the tracked objects, resulting in a descriptor vector of size 11110 for each frame.

Figure 3: An example of diversity in appearance

3.2 Evaluation Criteria

Active learning aims to achieve high accuracy using as little annotation effort as possible. Thus, a trade-off between accuracy and proportion of labelled data can be considered as one of the most informative measures. Let \( N \) denotes the total number of batches, \( MC \) refers to misclassified

\[\text{Due to space limitations, we have chosen not to include more details in the paper.}\]

Figure 4: Performance of NEVIL on multiple CAVIAR sequences. The results were obtained by applying SVM as the classifier.

3.3 Results

In our framework, Support Vector Machine (SVM) is applied as base learner. The size of the batches is optimized for each dataset. Figure 4 presents the results obtained from various CAVIAR sequences. In the OneStopMoveEnter1, the most complex scenario with 42 streams from 14 classes, NEVIL achieves 80% accuracy with 30% annotation effort.

4 Conclusion and Future work

In this paper, we proposed a semi-supervised incremental learning framework (NEVIL) to mine visual data coming from non-stationary environments. Inspired from active learning strategies, an oracle provides labelled batches. We have empirically shown that NEVIL can achieve high accuracy with fairly little human effort.

Although, framework fulfills most of the characteristics listed in Sec. 1, as the size of ensemble is growing in time, complexity is still the main concern and main direction for future research.

Acknowledgements

This work is financed by the ERDF -European Regional Development Fund through the COMPETE Programme (operational programme for competitiveness) and by National Funds through the FCT -Fundação para a Ciência e a Tecnologia within project FCOMP-01-0124-FEDER-037281.

References


