Incremental Learning from Low-labelled Stream Data in Open-Set Video Face Recognition

Eric Lopez-Lopez\textsuperscript{a,}\textsuperscript{*}, Xose M. Pardo\textsuperscript{b}, Carlos V. Regueiro\textsuperscript{a}

\textsuperscript{a} Universidade da Coruña, CITIC, Computer Architecture Group, Spain
\textsuperscript{b} CITIUS, Universidade de Santiago de Compostela, Spain

\begin{abstract}
Deep Learning approaches have brought solutions, with impressive performance, to general classification problems where a wealth of annotated data is provided for training. In contrast, less progress has been made in continual learning of a set of non-stationary classes, mainly when applied to unsupervised problems with streaming data.

Here, we propose a novel incremental learning approach which combines a deep features encoder with an Open-Set Dynamic Ensembles of SVM, to tackle the problem of identifying individuals of interest (IoI) from streaming face data. From a simple weak classifier trained on a few video-frames, our method can use unsupervised operational data to enhance recognition. Our approach adapts to new patterns avoiding catastrophic forgetting and partially heals itself from miss-adaptation. Besides, to better comply with real world conditions, the system was designed to operate in an open-set setting. Results show a benefit of up to 15\% F1-score increase respect to non-adaptive state-of-the-art methods.

© 2022 The Author(s). Published by Elsevier Ltd. This is an open access article under the CC BY-NC-ND license (http://creativecommons.org/licenses/by-nc-nd/4.0/)
\end{abstract}

1. Introduction

Deep Learning approaches have brought solutions, with impressive performance, to general classification problems where a wealth of annotated data is provided for training. Given the fact that in real-world applications specific data are many times scarce, very costly to label, non-stationary (i.e. data distributions changing over time), or streaming, new and classical learning strategies have been incorporated to the realm of Deep Learning, to deal with these challenges [1]. Thus, topics as transfer learning [2], reinforcement learning [3], or incremental learning [4–7], both supervised and unsupervised, have gained new momentum.

Incremental learning is the ability of a classifier to evolve by continuously integrating information from new instances or new classes, and without resorting to full retraining [1]. Currently, incremental and online machine learning are receiving more and more attention, especially in the context of learning from real-time data streams [4,5]. In particular, rehearsal-free incremental learning techniques have also demonstrated their abilities to extend the class-set of a classifier considering only labels from the new classes, while avoiding the problem of catastrophic forgetting [6,7]. Catastrophic forgetting is the tendency of an artificial neural network to completely and abruptly forget previously learned information upon learning new information [8]. Overcoming this issue is of special interest when computational capacities do not allow full retraining, or confidentiality issues impede new access to old samples during the process of extending the class set. In contrast, less progress has been made in incremental learning of a set of non-stationary classes, mainly when applied to tasks involving unsupervised streaming data.

A paradigmatic example of the application of incremental learning, dealing with unsupervised, non-stationary and streaming data is the case of video-to-video face recognition (V2V-FR) in video surveillance [9]. Usually, video-frame are captured with a broad range of individual pose, camera position, resolution, and illumination, which often exceeds the diversity available in datasets used to train deep networks (generally focused on web extracted images) [10]. Transfer learning to specific task domains in V2V-FR has proven to be challenging even for Deep Learning encoders [11,12] since image quality factors are still decisive for performance [13]. Besides, since the data are received continuously in a stream fashion, individual appearance could change when switching be-

\begin{doi}
https://doi.org/10.1016/j.patcog.2022.108885
\end{doi}

\begin{license}
0031-3203/© 2022 The Author(s). Published by Elsevier Ltd. This is an open access article under the CC BY-NC-ND license (http://creativecommons.org/licenses/by-nc-nd/4.0/)
\end{license}
between different cameras, which could also operate in changing conditions over time [14]. While, in theory, all of these issues can be solved with further labelling, the task of having addressed every possible variation in a training dataset is, in practical terms, infeasible [4]. Then, a more efficient and scalable approach is needed [6,7,15]. In this regard, what truly represents the application context and the changes that appear over time is the actual data incrementally extracted during the operation of the system, and so without labels.

Another characteristic of real-world applications of V2V-FR relates to their intrinsic open-set nature [16,17]. Open-set recognition refers to the classification problem aimed at identifying a specific set of known classes over an undetermined number of unknown ones [16]. This type of recognition exactly corresponds to some of the most common scenarios in which FR is demanded. Take for example the case of an airport video-surveillance aimed to track some individuals of interest (IoI) who have not been collaboratively enrolled in the system, e.g. those exhibiting suspicious behaviours, among a larger number of unknown non-target identities, that should be identified as unknown non-target identities.

In this paper, we propose a novel incremental learning system, the Open-Set Dynamic Ensembles of SVM (OSDe-SVM) (see Fig. 1). Using a deep feature encoder as a basis, the system is capable of using operational data to enhance and improve the recognition of target identities in an unsupervised way. Additionally, guided by real-world necessities, the system is designed to operate in a completely open set setting. Rooted on the power of a Deep features encoder, trained for the general face recognition problem, an incremental learning module, fed with stream data, simultaneously predict and update classifiers, while dealing with catastrophic forgetting issues. The incremental module is based on dynamic ensembles of SVM classifiers, which from a single SVM built from a few labelled video-frames directly taken from the footage, can acquire and adapt to additional information by adding/removing classifiers to/from ensembles. It follows the self-training strategy [18] in which predictions also play the role of pseudo-labels, which are used to update and improve the classifiers. Based on the modular nature of ensembles, adaptations consists of either adding or removing classifiers. Contributions of this paper can be summarised as:

- An approach to unsupervised incremental learning designed to operate online with stream data. During its operation, predictions also play the role of pseudo-labels.
- A strategy to deal with both catastrophic forgetting issues and the effect of mistaken pseudo-labels.
- An approach to instance-incremental learning in the open-set, which could be extended to cope with the class-incremental problem.
- A method for person re-identification based on face, which is not directly based on a reservoir of face images.

The rest of the paper is organised as follows. First, in Section 2 we perform an extensive study of the existent literature related to the problem. After that, we move to present the proposed approach in Section 3 and a set of experiments to study its behaviour, in Section 4 and 5. Finally, in Section 6, we reflect on the conclusions we can extract from the work.

2. Related Work

Open-set Recognition. In open set recognition, training is performed on a dataset with samples of some known classes, while samples of both known and unknown classes are presented for testing. Therefore, classifiers should appropriately deal with all of them. Within this approach, closer to real-world applications, decision boundaries not only separate instances of different known classes, but they separate the known from the unknown as well [16]. A recent survey [19] distinguishes between discriminative and generative approaches to open set recognition. Discriminative classifiers are trained to discriminate between the known classes, and then, given the most likely class label, to decide whether a test sample was in fact drawn from the distribution of known class samples or not [17]. Meanwhile, generative methods try to provide explicit probability estimation over unknown categories, most of them based on deep networks [20,21]. Plenty of methods in both sets of approaches, leverage Extreme Value Theory (EVT) to tackle the unknown [22]. EVT is a branch of statistics aimed to assess the probability of observing an event more extreme than any previously observed. It has been widely used for outlier detection in open-set recognition [23].

In face recognition, the most realistic scenario corresponds to an open-set setting (e.g. criminal watch-lists, restricted areas access control, smart-homes, etc.) [17]. In this domain, apart from EVT based methods, solutions based on siamese networks have been proposed to address the open-set as they are metric learning methods, and their similarity scores can be thresholded to perform
recognition [24]. Although they do not fit the data stream context, they could be used as a baseline for comparison purposes [25].

**Incremental Learning.** The main goal of incremental (a.k.a. lifelong, continuous or continual) learning is to learn from data as they are provided by real-world dynamic sources, usually at a low pace, including noisy samples and, in general, exhibiting non-stationarity. As data distributions change over time, computational systems have to deal with the stability-plasticity dilemma. This dilemma consists of finding a balance between models’ plasticity allowing them to adapt to changes (i.e. concept drift) [5] and stability to avoid that new knowledge erases old one.

In the context of deep approaches, incremental learning has been focused on learning new tasks/classes, more than on enhancing the performance of classifiers (fixed number of classes) as new instances arrive [15,26]. Among common strategies are the exploitation of, at least, partial or full rehearsal (looping over old data) [1,27], dynamic changes in architectures (retraining after pruning/increasing the number of neurons, filters or layers), and regularisation (updating weights in order not to forget previous knowledge) [28]. Among the last are usually also included a wide range of knowledge distillation methods, in which a teacher network transfers knowledge to a student network [29]. However, the drawback of distillation is that it generally needs to retain big past memories [6]. Notwithstanding the progress made in supervised incremental learning in recent years, there is still a substantial gap between the performance of batch offline learners on stationary data and the performance of the incremental learners that deal with non-stationary data [1,27].

Most of the work carried out to date regarding incremental learning is focused on batches. So, they need to wait for a batch of data to accumulate before a new adaptation can take place. Only a bunch of approaches were really designed to tackle the problem of incremental learning from streaming data, which is considered a more challenging task [30]. One of its critical difficulties is the infeasibility of complete manual labelling of streaming data in real-world applications. A more realistic approach should only assume that a few instances in data streams are labelled [31].

Most of the semi-supervised methods leverage unlabelled examples by making some assumptions, using label propagation or generating pseudo-labels during the learning process [32]. Some approaches are based on keeping a set of dynamic clusters to summarise class distributions and model their evolution over time [31]. Others use a few labelled data to initialise a set of models, which are afterwards sequentially updated based on pseudo labelled data [33–35]. In the specific case of video recognition, weak labels can be provided by the temporal tracking [36,37], but also co-training or predictions of the own classifiers can provide pseudo-labels.

Ensemble methods have been acknowledged as powerful tools to overcome catastrophic forgetting [1,38], when dealing with data streams [33,39]. Moreover, ensemble algorithms can be integrated with drift detection algorithms and incorporate dynamic updates, such as selective removal or addition of classifiers [40]. In the semi-supervised scenario, it must be taken into account that any kind of weak labelling or pseudo labelling is prone to error. So, dynamic updates can be also useful for healing from the effect of mislabelling. Unlike other incremental learning approaches (either classic [41] or DL-based [4]), ensembles provide a simple way to isolate updates and, consequently, make changes reversible. And not only that, since decisions are based on majorities, ensembles are robust to outliers.

In [42] ensembles of deep networks have been proposed to encourage networks to cooperate and take advantage of their prediction diversity, in the context of few-shot classification. Besides, to deal with tasks where training data are inadequate, the training of a collection of incrementally fine-tuned CNN models and their combination using an ensemble, was presented [43]. In [44], the authors propose an ensemble learning framework based on multiple CNN classifiers. The CNN acts as a feature extractor for the posterior use of different ensemble frameworks to classify its content. Recently, already in the context of incremental learning, an approach based on ensembles, which is close to ours, was proposed for tackling the problem of mechanical fault diagnosis [45].

Although there are propositions of end-to-end deep learning approaches for incremental semi-supervised learning [32,35], their inherent characteristics make them yet unsuitable to operate online with streaming data. Therefore, for this specific context, we propose to combine the good characteristics of a deep feature encoder, which transfers knowledge from the source domain, with an ensemble method able to provide adaptation to the target domain.


In this work we present the Open-Set Dynamic Ensemble of SVM (OSDe-SVM)\(^1\) for the problem of V2V-FR in the open-set context (Fig. 1). This method takes advantage of transfer learning from large labelled datasets, to get discriminant feature embeddings that feed an instance-incremental learning module. The complete method is able to achieve online adaptation to the task domain from unlabelled streaming data. To do so, OSDe-SVM relies on the self-training strategy and the modular nature of ensembles to add and remove classifiers in a totally autonomous way.

OSDe-SVM uses features taken after the last pair of convolutional and batch normalisation layers of the ResNet100-ArcFace (RN100-AF) network trained on MS1MV2 dataset [46]. This is one of the top-performing CNN in the face recognition state-of-the-art. ArcFace is a loss-function specifically designed to enhance the discriminative power of face recognition models. In this sense, the deepest networks, like ResNet-100, are the ones that take the most advantage of it [46]. The encoding transforms a 112x112 face crops into a 512-D feature embedding.

The general structure of OSDe-SVM is depicted in Fig. 2. Each individual of interest (IoI), \(k\), has an associated ensemble, \(E_k\), composed of a set of SVM classifiers, \(h_i\). This ensemble is updated whenever the system is queried. The update mechanism consists of adding classifiers based on the Ensembles Decision Functions, Section 3.1, following the self-training paradigm (Section 3.2). Besides, OSDe-SVM can remove classifiers when the maximum number of classifiers is reached (Limitation Module, Section 3.3) or when a possible mistake is detected (Self-healing, Section 3.4). Each SVM classifier is trained with a small number of positive samples (face crops of the first 5 frames containing each IoI) against a pool of initially labelled training samples (specifically a total of 100 frames) randomly drawn from other IoIs.

3.1. Ensemble Decision Functions

OSDe-SVM builds, and keeps updated, ensembles aimed at the re-identification of each IoI within the area of a camera network (Fig. 2). Ensemble’s decisions are made in a two-step process. Firstly, the **Sequence Scoring Function** assigns a certain score to the query sequence. Secondly, the **Recognition Decision Function** uses these scores to assign an identity label (either as one of the IoI or as an unknown).

3.1.1. Sequence Scoring Function

When making decisions, it is convenient that each ensemble gives a unique score to each incoming sequence. Nevertheless, both

\(^1\) Implementation can be found on: https://gitlab.citius.usc.es/eric.lopez/osde-svm.
(sequences and ensembles) are composed elements. Being \(n_f\) the number of sequence's frames and \(M^k\) the number classifiers of ensemble \(k\), we would have a total of \(n_f \times M^k\) different responses. We call the Sequence Scoring Function (SSF) to the process of combining all of these different responses into a unique score. This process consists of two levels:

- **At frame level**, we combine the responses of the ensemble's classifiers to give a unique score to each frame. The function used here is the median of the individual ensemble's SVM scores. If we then make decisions by establishing a threshold on the median score, in practice, we will be performing a majority voting based on the binary responses (using this same threshold) of the ensemble's classifiers.
- **At sequence level**, we take advantage of the temporal coherence assumption to assign a unique identity to the whole input sequence. This assumption allows us to combine all the frame's scores into a unique one. The function used here is the median.

3.1.2. Recognition Decision Function based on Extreme Value Theory

Once every ensemble delivers its prediction score about an input query, the next step is to combine all the predictions to decide the underlying identity. The identity assignment based on the best score is the usual procedure in a closed-set scenario \([46,47]\). That is because input sequences always belong to a known IoL. In an open-set scenario, assigning identities becomes trickier because non-match responses, corresponding to unknown identities, are also expected \([17]\). To tackle these scenarios, OSDe-SVM was endowed with a Recognition Decision Function (RDF) based on Extreme Value Theory (EVT).

EVT is a statistical theory aimed at estimating the probability of observing events more extreme than any previously observed. In practice, it has been widely used for reliability applications, as well as outlier detection \([48]\). In the frame of open-set recognition, EVT has been successfully applied on numerous occasions (see Section 2).

Here, we follow an approach similar to \([48]\). As any input sequence belongs to a unique identity, the sequences associated with other identities should deliver non-match outputs. According to the Fisher-Tippet-Gnedenko Theorem of EVT \([22]\), the distribution of these non-match scores is modelled by some particular functions.

In this case, for left bounded positive samples, the distribution of the extreme values \(G(z)\), is given by the Weibull distribution. For OSDe-SVM the greater similarity, the smaller the SSF \(\alpha\) score (i.e. \(x < 0\) for similar input sequences). So, we need to perform a variable change \((\bar{x} = m - x\) being \(m\) the median of the non-match scores) to satisfy the previous conditions and be able to fit the \(n\)-top scores to a Weibull distribution as described in \([48]\). Then, to discriminate between unknown and known identities, the best ensemble response (the best score) can be checked whether it comes from the Weibull Extreme Value distribution (Fig. 3) or not.

The complete decision process is depicted in Algorithm 1. More importantly, there we also show a way of distinguishing the known from the unknown by thresholding the Weibull distribution \((\bar{T}_W)\), instead of the actual scores which can be uncalibrated. Since the fitted function is different depending on the input sequence, we are implicitly personalising the threshold to each input sequence, as is depicted in Fig. 3.

3.2. Update Module: Incremental learning based on Self-Updating

OSDe-SVM was conceived to operate in the context of a shortage of labelled data. Only the first classifier of each ensemble is trained with a very short labelled sequence extracted from the input. The first five frames have proven to be the bare minimum for
Algorithm 1 Recognition Decision Function (RDF) based on EVT.

1: $S$ is the input sequence, $T_W$ is the threshold in the Weibull function
2: $E = \{e^1, e^2, \ldots, e^{N-1}\}$ set of ensembles associated to known identities
3: $R = \emptyset$ set of scores given by each ensemble to a candidate
4: for $e^i$ in $E$ do
5:     $R \leftarrow \text{SSF}(e^i, S)$
6: end for
7: $c = \min(R); m = \text{median}(R \setminus \{c\})$
8: $V = \{(m - x) \mid x \in (R \setminus \{c\}) \wedge (x < m)\}$
9: Fit $V$ to a Weibull function, $W$
10: if $\text{then}(m - c) < T_W$ then
11:     $\text{ID} = \text{arg}(c)$
12: else
13:     $\text{ID} = \text{unknown}$
14: end if

our method. From that point on, incremental learning is exclusively based on pseudo-labels (Fig. 1).

After an ensemble is initialised, our method decides whether a new classifier must be added to enhance future performance, each time a sample of the same identity is identified. OSDe-SVM follows a self-updating strategy based on pseudo-labels provided by EDF to input sequences. Whenever an identity, $k$, is identified in an input sequence, a new SVM is created using as (pseudo-labelled) positive samples, $P_k$, the 5 hardest frames of the sequence, namely those which got the lowest scores returned by the SSF (see Fig. 2). This way, diversity within each ensemble is encouraged.

3.3. Limitation Module

In a self-updating context where each ensemble is initialised with only one classifier trained with a few labelled frames, further updates can only occur when close samples of the same identity query the system. If they are almost identical, there is nothing to be learnt. However, if they are very different, there is a danger of not being identified. So, the model can only learn from samples on the borderline, i.e. samples that can still be recognised by the ensemble of the corresponding identity, but which also include some level of novelty in their features. However, ensembles’ size should not grow indefinitely whenever the EDF recognised their target identities in input sequences. As ensembles’ performance relies on diversity, we have chosen a solution inspired in [49], to decide which classifiers are to be removed once the maximum size is reached.

Classifies are compared against each other to obtain a measurement of their relative relevance, the diversity score $D(-)$. Given an ensemble, $e^k$, composed by $M_k$ SVM classifiers, $\{h^k_0, h^k_1, \ldots, h^k_{M_k-1}\}$, $D(h^k)$, is computed from the binary response of each of the classifiers of the ensemble over a certain set of video frame features $\{x_0, x_1, \ldots, x_{Q-1}\}$:

$$D(h^k) = \sum_{j=0, j \neq i}^{M_k-1} d(h^k_i, h^k_j)$$

$$d(h^k_i, h^k_j) = \frac{1}{Q} \sum_{q=0}^{Q-1} \text{sgn}(h^k_i(x_q)) \cdot \text{sgn}(h^k_j(x_q)),$$

where $h^k_i(x_q)$ is the response of the SVM classifier $h^k_i$ to the frame feature $x_q$, and $\text{sgn}(-)$ is the sign function.

Whenever an ensemble $e^k$ reaches the maximum size, the classifier $h^k_i$ with the lowest diversity will be removed.

3.4. Self-healing: Correcting Wrong Updates

Since the whole adaptation process performs without supervision, wrong updates, provoked by errors in pseudo-labelling, should be expected. This behaviour may affect re-identification performance, mainly in the long term. The self-healing procedure is designed to mitigate this problem.

Self-healing relies on the fact that the ensembles build their decisions based on majorities. Therefore, if an ensemble reaches a relatively high accuracy in the first classifications, it should be difficult for wrong classifiers to take over very soon. This fact opens the possibility of detecting wrong updates before it becomes irreversible. We expect that, with a limited amount of wrong updates, ensembles are still able to recognise their target identity. Consequently, the future detection of the target identity can build a stronger majority capable of detecting the previous wrong update.

To implement these ideas, along with each SVM classifier, $h^k_i$, we store the positive samples used to create it, $P^k_i$, which, in prac-
4. Experimental Preliminaries

4.1. Database Selection

To test any V2V-FR system, we must rely on video datasets to perform our experiments, and they are not too abundant [50]. In this sense, frames’ quality (especially in terms of resolution), which can vary substantially depending on the context, can have a substantial impact on the performance of recognition methods [13]. This fact is even more important when we aim to work in video-surveillance scenarios.

We have included experiments in three different video datasets. On the one hand, CMU FIA [51] and COX Face Database [9], are datasets specifically designed for video-surveillance scenarios. However, their frame quality is radically different, as is the demand for adaptation.

On the other hand, we have also performed experiments on YouTube Faces Database [52] to test how the proposed method performs in other FR video contexts.

4.1.1. CMU Face in Action (FIA) database

The CMU FIA database contains 20-second videos of more than 200 different individuals simulating a passport checking scenario in both indoor and outdoor environments [51]. Data was acquired by six synchronised cameras from 3 different angles, 2 focal lengths per angle, in 3 different sessions (3-months span between each pair of sessions). FIA video-frames present a considerable high quality, specifically in terms of resolution, since they were captured in a relatively controlled scenario. This dataset has been used to assess other adaptive methods, like the one in [33].

In our experiments, we have used the videos provided by the smaller focal length of the frontal camera, both indoor and outdoor, and only considered the 70 identities present in all sessions. Given the high quality of frames, the initial performance of our method reached values of +92% in F1-score, which widely surpasses the ones observed in [33]. In this sense, with these performance rates, it is difficult for any unsupervised adaptation method to increase them. Even more, if we take into account that OSDe-SVM mainly provides improvements in recall (as we will see in the following sections), which here reach almost perfect values from the beginning.

To challenge our method by emulating more realistic conditions, we decided to down-sample the video-frames before entering the feature encoder. In Fig. 5 performance results for 5 different downscaling ratios are shown for the case of 35 lol in a universe of 70 identities. Without having the possibility of averaging performance under different universes, we decided to randomly draw 20 different sets of 35 lol for average and deviation computations. We measure OSDe-SVM performance before and after adaptation. Results in Fig. 5 show the performance degradation as the resolution decrease, which OSDe-SVM alleviates with its unsupervised adaptation. It must be taken into account that a 1/16 down-scale gives face crops of size 7x7 (for an original size of 112x112); such low resolutions make identification almost unfeasible.

4.1.2. COX Face Database

COX Face database [9] was specifically designed for the context of video-surveillance. There is a total of 1000 different identities in the dataset. The creators of the database asked to each of individual to follow an S-path while they capture video from 3 different viewpoints (cam1, cam2 and cam3). Samples can be seen in Fig. 6. Despite being taken in an interior setting, the resulting video frames present important variations in terms of both illumination and pose and especially low resolution. Samples provided
by the database are the output of a commercial face tracker with a partially removed background. Nevertheless, to fine-tune this background removal and for alignment purposes, faces are passed through a face detection module for the proper performance of the feature encoder module [53]. This database will be the one in which the core parts of OSDe-SVM are tested.

4.1.3. YouTube Faces Dataset

YouTube Faces database [52] is a widely used database for the context of video face recognition. While not being designed for the case for the specific case of video-surveillance, it will provide insights into how well OSDe-SVM generalise to other video contexts. This database contains 3.425 videos of 1,595 different people. Each identity contains from 1 to 6 different sequences. Since we wanted to have room to perform adaptation during operation, we will only keep the identities containing ≥3 videos. This gives us a total of 533 different identities. Samples can be seen in Fig. 6.

4.1.4. Dataset adjustments for the experiments

Given the specific context of our application, it was necessary to perform some adjustments to adapt the data provided by databases to how we operate. First, to increase the number of sequences per identity (and so the possibilities to update), we split each of the available videos to have an initial mini-sequence of 5 frames and 9 additional sub-sequences. All this process keeps intact the temporal order. Second, we organised the data into different sets depending on their role in the experiments. These different sets are:

- The **initially labelled training sequences** are labelled video-frames of target identities used to create the first classifier of each ensemble (the sets positive samples, \( P_k \)). They consist of the first 5 frames from the first available database video.
- The **operational sequences** simulate input sequences which would be received in the operational phase. They consist of the first eight sub-sequences of the available sequences.
- The **testing sequences** are used to assess performance. They correspond to the last sub-sequence of 9 sub-sequences of each of the available individuals.

Hereafter, each sequence will be noted by \( S_t^k \), where \( t \) refers to temporal order and \( k \) refers to the identity. Following this notation, \( t = 0 \) corresponds to the 5-frame sequences of the **initially labelled training sequences** used in the initialisation, \( t = 1, 2, \ldots, 8 \) correspond to the streaming of sequences (operational sequences), and \( t = 9 \) corresponds to a sequence for performance assessment (testing sequences).

4.2. Experimental Setup

We designed the experimental setup to simulate the stream data scenario of V2V-FR. First, initial models of the lols are created, which consist of one-classifier ensembles. This classifier is created using samples from the **initially labelled training sequences** (\( S_{t=0}^k \)): 5 frames of the actual identity as a positive set and a 100 frames from other lol as a negative set (randomly drawn without restrictions for each classifier from other subject’s samples pool). The size of the negative set is maintained for future classifier additions to have the same balance in each of the ensemble’s classifiers. After this initialisation process, the system is repeatedly queried with unlabelled sequences. These sequences have a variable number of frames (from 20 to 60). Since we are working in an open-set scenario, these input sequences can belong to one of the lol or not.

Experiments are organised in **adaptation steps**, after which performance is measured. An **adaptation step** corresponds to either the initialisation, a complete iteration over the \( k \) available identities with the same \( t \), or a process of self-healing (See Fig 7). Additionally, we fully iterate over \( t = 1, 2, \ldots, 8 \) a total of 3 times (iterations), always preserving the temporal order. This way, we can increase the number of possible updates and study the system’s behaviour with redundant data of both lol and unknowns. Self-healing was performed at adaptation steps multiples of \( 5 \), and
the maximum number of classifiers per ensemble, \( M \), was fixed to 10. This gives us a total of 31 adaptation steps per experiment. Algorithm 2 outlines the whole procedure.

**Algorithm 2** Experimental procedure and testing protocol.

1. \( S_0^k \) is the sequence \( t \) of the identity \( k \), \( L \) is the number of iterations
2. \( f \) number of different sub-sequences per identity
3. \( N = \) number of lol, \( N_u = \) number identities in the universe
4. for each split do
5. \( \text{for } k = 0 \text{ to } N - 1 \text{ do} \)
6. Initialise ensemble \( k \) using \( S_0^k \)
7. end for
8. Perform testing using the set of \( S_f^{k=\{0,1,...,N-1\}} \)
9. \( \text{for } \) lap \( = 0 \text{ to } L - 1 \text{ do} \)
10. \( \text{for } t = 1 \text{ to } f - 1 \text{ do} \)
11. \( \text{for } k = 0 \text{ to } N_u - 1 \text{ do} \)
12. Perform adaptation using \( S_f^k \)
13. end for
14. Perform testing using the set of \( S_f^{k=\{0,1,...,N-1\}} \)
15. end for
16. end for
17. end while

Both the size of the identity universe and the number of lols vary with the experiment. For universe sizes smaller than 1000, the experiment is repeated for different splits of identities (following Algorithm 3) to compute an average performance. A partial overlapping between splits was considered to get a more comprehensive sampling. For the case of 1000 identities, we repeat the experiment 5 times to address the variations provoked by the random set of negatives. For example, in the case of a universe with 100 identities, we would have a total of 19 different splits. As for metrics, we measure precision, recall and F1-measure, using a \( T_w \) fixed to 0.01.

**5. Experiments and Results**

The experimental part of the paper is organised as follows. First, we study the dependence of performance against the size of the universe, while maintaining the ratio with respect to the number of lol constant (Section 5.1). After that, we perform a comprehensive analysis of the temporal evolution of one of the previous configurations (Section 5.2.1). Then, we compare the performance of our approach against state-of-the-art face recognition methods (Section 5.3). Finally, the effect of openness is assessed (Section 5.4).

### 5.1. Performance vs. Universe Size

This experiment shows the performance behaviour of OSDe-SVM under different universe sizes \( (N_u) \) while keeping the proportion of lol to \( N_U \) 1:2. Results are shown in Tables 1 and 2. We measure initial (non-adaptation) and final (after adaptation) performance of OSDe-SVM, using the previously described experimental set-up (Section 4.2). It is important to remark that non-adaptation means that ensembles do not incorporate new SVMs apart from the initial one. Thus, performance is quite similar to the one provided by the original network [46].

From the experimental results on both datasets, the benefits provided by the adaptive nature of the OSDe-SVM are patent. F1-scores increase in all but one case (the case of having 10 lol in a Universe of 20 for YTF), mainly due to the impact on recall (9-30% improvement). OSDe-SVM helps to enhance and enrich the existent face models, being able to recognise what previously were unrecognisable. This improvement is even more remarkable accounting for the challenging experimental conditions. First, only 5 low-quality frames are provided with true labels to create the initial models. After that, no additional labelling is provided. Second, we use the same identities (both known and unknown) to perform the queries in each adaptation step. Therefore, confusion between identities could reinforce the impostor and eventually provoke a complete identity theft.

Although overall the behaviour observed is stable, the highest improvement in performance corresponds to larger universes. This behaviour can be explained by how we use the EVT. The quality of the Weibull fit in RDF (Section 3.1.2) increases as the number of samples to fit do so. For instance, since just half of the data is used in this process (those greater than the median, \( L_8 \) in Algorithm 1), when the lol is 10 the Weibull fit is done with only 5 points.

Finally, while OSDe-SVM presents benefits in both datasets, COX is the one which presents the best result. This can be attributed to fact that their sequences are more contiguous.

### 5.2. Comprehensive study of the 50 lol in a universe of 100

This experiment was aimed at performing a detailed study of one of the previous cases (50 lol in a universe of 100) to fully understand OSDe-SVM behaviour. First, we will study its detailed temporal evolution in Section 5.2.1. After that, we complement this study by exploring the behaviour of two fundamental parts of OSDe-SVM: self-healing (Section 3.4) and the decision threshold \( T_w \) (Section 5.2.3).

#### 5.2.1. Temporal evolution

The results of the temporal evolution are shown in Fig. 8. The first thing we can extract from the experiments (Fig. 8 a) is that the performance improvement is higher in the first steps. This is something which could be expected as adding individual classifiers has a higher impact when the size of the ensemble is lower.
Besides, this behaviour shows the system's robustness against repeated unknown queries.

These figures also allow observing in a more detailed manner the remarkable recall improvement provided by OSDe-SVM. Precision is also improved but to a lesser extent. Besides, Fig. 8 b shows the evolution of the average ensemble size for each of the splits (Algorithm 3). We can see the effect of self-healing (every 5 steps) and the limitation module. First, drops in size correspond to the triggering of the self-healing process. Second, the size of each ensemble, $M^k$, is effectively restricted by the limitation module to 10 SVM classifiers.

5.2.2. The effects of self-healing

Here we wanted to see the effects on the performance of using or not the self-healing module. Therefore, we have repeated the previous experiment for 50 IoI in a universe of 100 with the self-healing module disabled. The results are in Table 3.

Self-healing provides limited performance increases. Consequently, the results do not achieve enough significance. Precision is not degraded as much with this module activated (+1% difference). All of this suggests that the module may be helping to partially eliminate wrong classifiers from the ensemble. This improvement, while being small, opens an interesting research line to further optimise ensemble errors self-correction in the future.

5.2.3. Dependence on $T_W$

OSDe-SVM makes its decisions by establishing a threshold ($T_W$) on the Weibull distribution fitted to the ensemble's scores of an incoming sequence (see Section 3.1.2). This approach allows the threshold to better generalise to each specific context. Even more, when data scarcity forces us to fix its value a priori.

In this section, though, we wanted to explore the influence of $T_W$ in OSDe-SVM performance, to better understand its behaviour. Note that such an exploration could not be performed in a real-world application. We are going to measure initial and final perfor-

| Table 1 |

| Table 2 |

<table>
<thead>
<tr>
<th>$N$</th>
<th>$N_0$</th>
<th>Precision</th>
<th>Recall</th>
<th>F1</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td>Initial</td>
<td>Final</td>
<td></td>
</tr>
<tr>
<td>10</td>
<td>20</td>
<td>75 (12)</td>
<td>71 (12)</td>
<td></td>
</tr>
<tr>
<td>20</td>
<td>40</td>
<td>86.9 (8.2)</td>
<td>85.1 (6.3)</td>
<td></td>
</tr>
<tr>
<td>30</td>
<td>60</td>
<td>88.5 (6.4)</td>
<td>89.7 (5.5)</td>
<td></td>
</tr>
<tr>
<td>50</td>
<td>100</td>
<td>91.2 (4.7)</td>
<td>91.9 (3.8)</td>
<td></td>
</tr>
<tr>
<td>100</td>
<td>200</td>
<td>92.6 (3.0)</td>
<td>92.6 (2.1)</td>
<td></td>
</tr>
<tr>
<td>200</td>
<td>400</td>
<td>92.0 (1.8)</td>
<td>93.5 (1.6)</td>
<td></td>
</tr>
<tr>
<td>300</td>
<td>600</td>
<td>90.3 (1.3)</td>
<td>91.9 (1.3)</td>
<td></td>
</tr>
<tr>
<td>500</td>
<td>1000</td>
<td>84.6 (1.4)</td>
<td>89.33 (0.76)</td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>$N$</th>
<th>$N_0$</th>
<th>Precision</th>
<th>Recall</th>
<th>F1</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td>Initial</td>
<td>Final</td>
<td></td>
</tr>
<tr>
<td>10</td>
<td>20</td>
<td>79 (10)</td>
<td>73.3 (9.5)</td>
<td></td>
</tr>
<tr>
<td>20</td>
<td>40</td>
<td>89.4 (5.8)</td>
<td>87.0 (6.2)</td>
<td></td>
</tr>
<tr>
<td>30</td>
<td>60</td>
<td>91.8 (4.4)</td>
<td>90.5 (5.1)</td>
<td></td>
</tr>
<tr>
<td>50</td>
<td>100</td>
<td>93.6 (2.9)</td>
<td>91.7 (3.7)</td>
<td></td>
</tr>
<tr>
<td>100</td>
<td>200</td>
<td>93.7 (1.5)</td>
<td>91.9 (1.7)</td>
<td></td>
</tr>
<tr>
<td>200</td>
<td>400</td>
<td>92.3 (1.6)</td>
<td>93.81 (0.48)</td>
<td></td>
</tr>
</tbody>
</table>

| (a) Performance evolution. |
| (b) Ensemble size evolution. |

Fig. 8. Evolution of OSDe-SVM for the case of 50 IoI over an universe of 100 identities. Test performed on COX Face Database.
5.3. Comparison against state-of-the-art face recognition models.

Here, we compare the performance of OSDeSVM against two other well-known methods for face recognition (Table 5). In these two methods, the focus was on obtaining the most widely separated classes in feature space, to make the classification as easy as possible. On the one hand, FaceNet [25] feature embedding is designed to distinguish faces by computing the euclidean distance between two features (99.6% accuracy on LFW). On the other hand, ArcFace [46] embeddings are designed to distinguish features by using cosine similarity. All of this makes them suitable for application in any face related task (either verification, identification or general recognition) or, as in our case, to use as a basis for the development of an adaptive method.

Both euclidean distance and cosine similarity are used to compare two single features. Since here we work with the features of all the frames in each query sequence, the centre of this cluster of features is computed as proposed in the original paper [46], to obtain a unique feature per sequence. Besides, the thresholds were tuned offline to get the best F1-scores, which are used as baselines. This would be impossible to do in stream learning conditions.

Results on Table 5 allow us to gain insights into the issues addressed in this paper. First, the performance of FaceNet shows the difficult endeavour of transitioning to real-world problems (low-quality, open-set considerations, etc.). Second, our initialisation OSDeSVM with RN100-AF embeddings preserves most of the discrimination power of the original decision function (cosine similarity). Finally, the enhanced performance provided by OSDeSVM is put into perspective against other state-of-art static face recognition models. This improvement translates into a 15% higher F1-score.
5.4. Performance vs. Openness

The goal of this experiment is to study how the behaviour of OSDe-SVM changes with the openness ratio, \( O \), that is the ratio of known to unknown identities [16]. This measure goes from 0% openness (closed-set recognition) to, theoretically, 100%:

\[
O = 1 - \sqrt{\frac{N_{\text{training}}}{N_{\text{target}} + N_{\text{testing}}}},
\]

where \( N_{\text{training}} \) is the number of identities used on training (in our case, \( N_{\text{training}} = 50 \)) and \( N_{\text{target}} \) is the number of identities to recognise (in our case, \( N = 1000 \) identities). Thus, Eq. 3 simplifies to:

\[
O = 1 - \sqrt{\frac{N}{N + N_0}},
\]

To have a wide range of openness values, we selected a relatively low number of lol (50) and then vary the size of the universe from 50 identities (0% openness, i.e. closed-set) to 1000 identities (≈70%). Experimental results are shown in Fig. 9, where performance is represented in terms of precision, recall and F1-scores.

The performance graphs show a clear decay of F1 performance as openness increases, because of the loss of precision. It must be noted that openness affects both the unsupervised adaptation and the testing process. An increase in openness provokes a decay in precision, which also entails making more mistakes during the self-adaptation. Accordingly, the drop in precision leads to a decay in recall after the adaptation process. Against all odds, the system proves its robustness until almost 60% of openness.

6. Conclusions

In this work, we propose a novel system, the OSDe-SVM as an instance-incremental learning approach to the problem of open-set face recognition in video surveillance. This system aims to operate in real-world non-stationary environments where the availability of labelled data is quite limited.

OSDe-SVM design uses the power of deep face representations as a basis. Once initialised (using 5 labelled frames per lol), the proposed method creates and updates an ensemble of SVM classifiers using samples directly taken from the input sequence which effectively deals with catastrophic forgetting. These updates are performed following the self-training paradigm in which OSDe-SVM predictions are used as pseudo-labels to incorporate new knowledge without additional supervision. In this regard, we achieve update reversibility by encapsulating each update into an individual SVM classifier. By the use of EVT, OSDe-SVM can make decisions in open-set conditions.

Experiments were mainly performed on COX Face Database, to our knowledge the most challenging video-surveillance database available. Guided by real-world necessities, the set-up simulates open set recognition conditions. Results show up to a 15% F1-measure (achieving up to a ≈ 94% F1-measure, depending on the amount of lol to recognise) increase respect to the closest static state-of-the-art (ResNet100+AF) face recognition model. Furthermore, the proposed system’s performance is tested under different degrees of openness, proving to be reliable up to +60% openness (50 lol in a universe of 1000 identities), where unknown identities appear many more times than lol. Additionally, CMU FiA and YouTube Face databases are also successfully used to test the generalisation capabilities of the proposed OSDe-SVM.

In future work, apart from translating OSDe-SVM to other related machine learning applications, an interesting line of research would be to extend the proposed system to the unsupervised class-incremental problem. Following the same self-training paradigm, unknown responses could be used to incorporate additional lol into the recognition system. And, since classifiers are independently created, these additions would not have any adverse effect on previous knowledge.

Declaration of Competing Interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgements

This work has received financial support from the Spanish government (project PID2020-119367RB-I00); from the Xunta de Galicia, Consellería de Cultura, Educación e Ordenación Universitaria (accreditations 2019-2022 ED431G-2019/04 and ED431G-2019/01, and reference competitive groups 2021-2024 ED431C 2021/48 and ED431C 2021/30), and from the European Regional Development Fund (ERDF). Eric López-López has received financial support from the Xunta de Galicia and the European Union (European Social Fund - ESF).

References
