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

Eric Lopez-Lopez, Xose M. Pardo, Carlos V. Regueiro

PII: S0031-3203(22)00366-1
DOI: https://doi.org/10.1016/j.patcog.2022.108885
Reference: PR 108885

To appear in: *Pattern Recognition*

Received date: 9 December 2020
Revised date: 20 May 2022
Accepted date: 2 July 2022

Please cite this article as: Eric Lopez-Lopez, Xose M. Pardo, Carlos V. Regueiro, Incremental Learning from Low-labelled Stream Data in Open-Set Video Face Recognition, *Pattern Recognition* (2022), doi: https://doi.org/10.1016/j.patcog.2022.108885

Highlights

- A online approach to unsupervised instance-incremental learning with stream data.
- Adaptation from pseudo-labels, which are the own predictions of the system.
- A strategy to deal with catastrophic forgetting and the effect of wrong pseudo-labels.
- Designed to operate in the open-set, extendable to the class-incremental problem.
- Method for person re-identification based on face without a reservoir of face images.

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

Eric Lopez-Lopez[a,*], Xose M. Pardo[b], Carlos V. Regueiro[a]

[a] *Universidade da Corua, CITIC, Computer Architecture Group*
[b] *CiTIUS, Universidade de Santiago de Compostela*

Abstract

Deep Learning approaches have brought solutions, with impressive performance, to general classification problems where wealthy of annotated data are 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.

Keywords: Open-set face recognition, Incremental Learning, Self-updating, Adaptive biometrics, Video-surveillance

*Corresponding author
Email address: eric.lopez@udc.es (Eric Lopez-Lopez)

1. Introduction

Deep Learning approaches have brought solutions, with impressive performance, to general classification problems where a wealthy set 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, 5, 6, 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

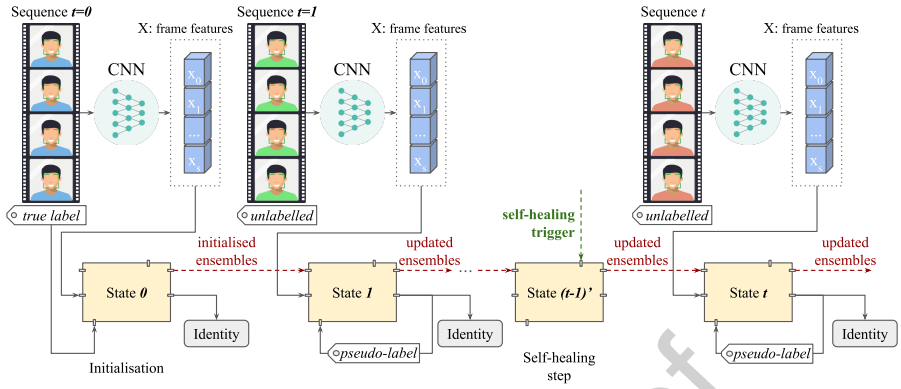


Figure 1: Open-Set Dynamic Ensembles of SVM (OSDe-SVM) is able to incorporate new knowledge and correcting wrong updates by adding and removing classifiers in an unsupervised way. The system is designed to work under open-set recognition conditions.

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 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 [15, 6, 7]. 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 (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 Sec. 2 we perform an extensive study of the existent literature related to the problem. After that, we move to present the proposed approach in Sec. 3 and a set of experiments to study its behaviour, in Sec. 4 and 5. Finally, in Sec. 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 with 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 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 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 [27, 1].

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, 34, 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* [38, 1], when dealing with data streams [39, 33]. 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

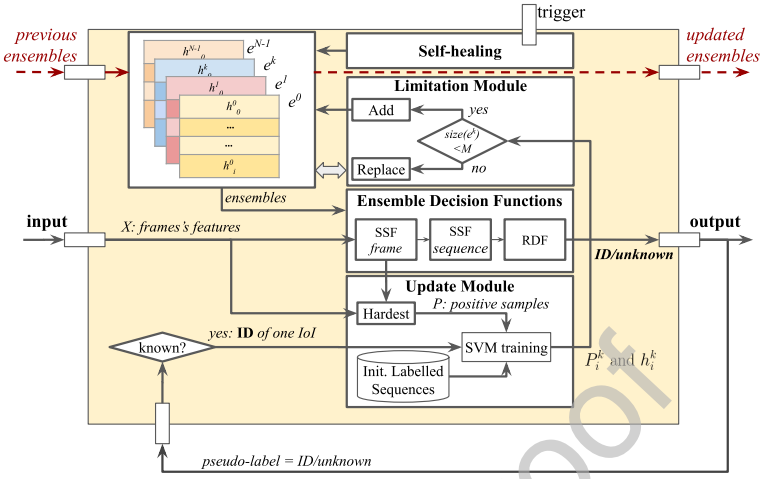


Figure 2: The pipeline of OSDe-SVM. After being processed by a deep feature extractor, the sequence's frames pass through the Ensemble Decision Function (EDF). This function assigns an identity label (either one of the IoI or *unknown*) based on the scores given by the ensembles at the moment. If recognised as one of the IoI, OSDe-SVM will add an additional classifier to the associated ensemble. Additionally, the limitation module controls ensembles to not exceed maximum size and the self-healing module helps to correct possible wrong updates *a posteriori*.

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.

3. Proposed Method: Deep Embeddings + Open-Set Dynamic Ensembles of SVM (OSDe-SVM)

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

[1]Implementation can be found on: `https://gitlab.citius.usc.es/eric.lopez/osde-svm`

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^k. This ensemble is updated whenever the system is queried. The update mechanism consists of adding classifiers based on the Ensembles Decision Functions, Sec. 3.1, following the self-training paradigm (Sec. 3.2). Besides, OSDe-SVM can remove classifiers when the maximum number of classifiers is reached (Limitation Module, Sec. 3.3) or when a possible mistake is detected (Self-healing, Sec. 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 (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 IoI. 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).

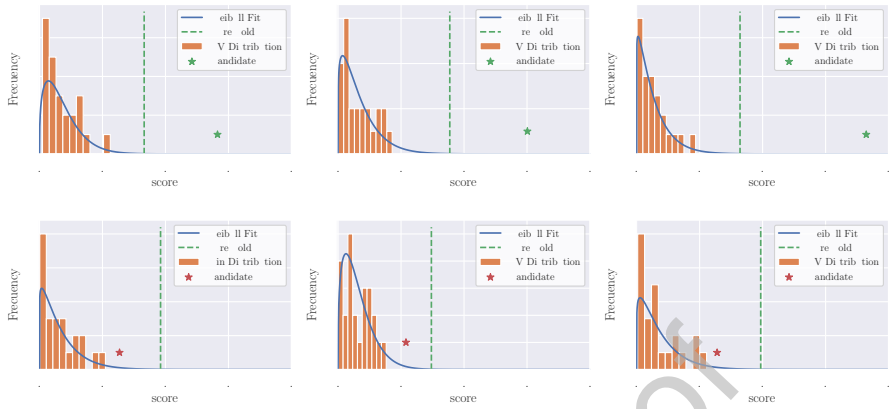


Figure 3: Examples of Extreme Values (EV) distributions and application of RDF. A Weibull function is fitted to the EV distribution to see whether the candidate belongs or not to this distribution. First row illustrates examples of *known* identities and second row does the same with the *unknown* ones.

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 Sec. 2).

Here, we follow an approach similar to [48]. As any input sequence belongs to a unique identity, the ensembles 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 (x) score (i.e. $x < 0$ for similar input sequences). So, we need to perform a variable change ($\hat{x} = m - x$, 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 discrimi-

Algorithm 1 Recognition Decision Function (RDF) based on EVT.

S is the input sequence, T_W is the threshold in the Weibull function
$E = \{e^0, e^1, \ldots, e^{N-1}\}$ set of ensembles associated to *known* identities
$R = \varnothing$ set of scores given by each ensemble to a candidate
for e^i in E **do**
    $R \leftarrow SSF(e^i, S)$
end for
$c = \min(R)$; $m = median(R \setminus \{c\})$
$V = \{(m - x) \mid x \in (R \setminus \{c\}) \wedge (x < m)\}$
Fit V to a Weibull function, W
if then$W(m - c) < T_W$
    ID $= arg(c)$
else
    ID $=$*unknown*
end if

nate 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 Alg. 1. More importantly, there we also show a way of distinguishing the known from the unknown by thresholding the Weibull distribution (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 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_j^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* $D(\cdot)$. Given an ensemble, e^k, composed by M_k SVM classifiers, $\left\{h_0^k, h_1^k, ..., h_{M_k-1}^k\right\}$, $D(h_i^k)$, is computed from the binary response of each of the classifiers of the ensemble over a certain set of video frame features $\{x_0, x_1, ..., x_{Q-1}\}$:

$$D(h_i^k) = \sum_{j=0; j\neq i}^{M_k-1} d\left(h_i^k, h_j^k\right) \tag{1}$$

$$d\left(h_i^k, h_j^k\right) = -\frac{1}{Q}\sum_{q=0}^{Q-1} sgn\left(h_i^k(x_q)\right) \cdot sgn(h_j^k(x_q)), \tag{2}$$

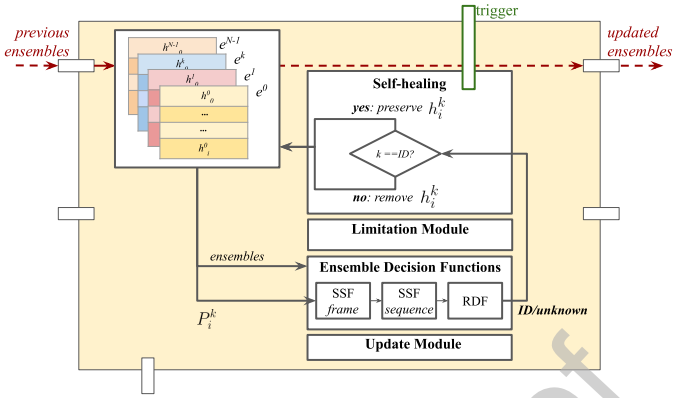


Figure 4: The pipeline of OSDe-SVM when self-healing is performed. The frames used to create each of the ensemble's classifiers are passed again through the Ensemble Decision Function to check if they are still recognised with the same identity. If not, this classifier is removed.

where $h_i^k(x_q)$ is the response of the SVM classifier h_i^k to the frame feature x_q, and $sgn(\cdot)$ is the sign function.

Whenever an ensemble e^k reaches the maximum size, the classifier h_*^k 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_i^k, we store the positive samples used to create it, P_i^k, which, in practice, can be considered a sequence. Therefore, we can pass every set (for all k and i) again through the *EDF* for a re-evaluation. If the system assigns the same identity as before, the classifier is maintained. Otherwise, the classifier is removed (Figs. 4). The self-healing module triggers after a certain period which is adjustable (see Fig. 1).

4. Experimental Preliminars

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

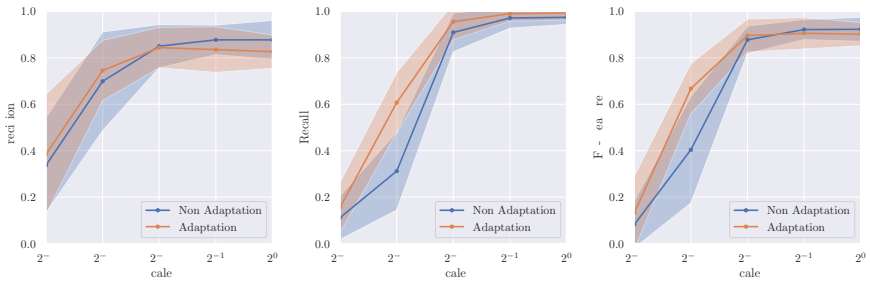
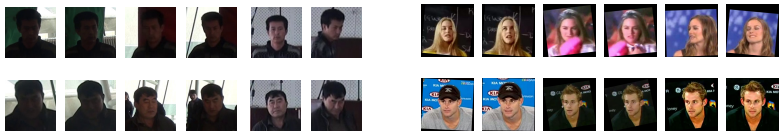


Figure 5: Performance versus image resolution, scales 1, 1/2, 1/4, 1/8 and 1/16 (that is, 112x112, 56x56, 28x28, 14x14, and 7x7 pixel image sizes) for CMU FiA database.

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 IoI 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 IoI 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 downscale



(a) COX Face Database. (b) YouTube Face Database.

Figure 6: Some samples of the main datasets used during the experiments.

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 are 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_i^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*).

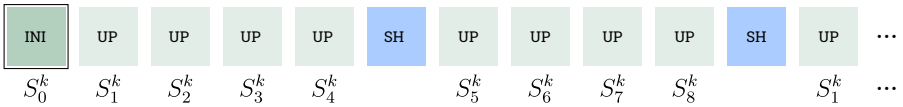


Figure 7: Adaptation steps performed during the experiments. INI stands for initialisation, UP for update and SH for self-healing. The last step corresponds to the beginning of the second iteration. Each sub-sequence contains about 20 to 60 frames.

4.2. Experimental Setup

We designed the experimental setup to simulate the stream data scenario of V2V-FR. First, initial models of the IoIs are created, which consist of one-classifier ensembles. This classifier is created using samples from the *initially labelled training sequences* (S_0^k): 5 frames of the actual identity as a positive set and a 100 frames from other IoI 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 IoI 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 $t = \{1, 2, ..., 8\}$ 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 IoI 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. Alg. 2 outlines the whole procedure.

Both the size of the identity universe and the number of IoIs vary with the experiment. For universe sizes smaller than 1000, the experiment is repeated

Algorithm 2 Experimental procedure and testing protocol.

S_t^k is the sequence t of the identity k, L is the number of iterations
f number of different sub-sequences per identity
N = number of IoI, N_U = number identities in the universe
for each *split* **do**
    for $k = 0$ **to** $N - 1$ **do**
        Initialise ensemble k using S_0^k
    end for
    Perform testing using the set of $S_f^{k=\{0,1,\ldots,N-1\}}$
    for $lap = 0$ **to** $L - 1$ **do**
        for $t = 1$ **to** $f - 1$ **do**
            for $k = 0$ **to** $N_U - 1$ **do**
                Perform adaptation using S_t^k.
            end for
            Perform testing using the set of $S_f^{k=\{0,1,\ldots,N-1\}}$
        end for
    end for
end for

for different splits of identities (following Alg. 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 IoI constant (Sec. 5.1). After that, we perform a comprehensive analysis of the temporal evolution of one of the

Algorithm 3 Algorithm to create the splits

1: N_U is the number of identities in each experiment universe
2: N_D is the number of identities in the dataset
3: i = 0
4: **while** $i + N_U < N_D$ **do**
5: *splits* $\leftarrow$ Samples with ID $\in \left[\frac{i}{2}, \frac{i}{2} + N_U\right]$
6: $i+ = N_U$
7: **end while**

Table 1: Performance over different universe sizes (N_U), while preserving the ratio with the number (N) of IoI. Values are expressed as $\mu(\sigma)$, where μ stands for mean and σ for standard deviation. Test performed on COX Face Database.

		Precision		Recall		F1	
N	N_U	Initial	Final	Initial	Final	Initial	Final
10	20	75 (12)	71 (12)	79 (17)	88 (11)	76 (14)	78.5 (9.7)
20	40	86.9 (8.2)	85.1 (6.3)	74 (15)	91.1 (6.7)	79 (11)	87.8 (5.4)
30	60	88.5 (6.4)	89.7 (5.5)	72 (10)	94.3 (3.7)	78.8 (7.8)	91.8 (3.7)
50	100	91.2 (4.7)	91.9 (3.8)	70 (13)	94.2 (4.1)	78.8 (9.0)	93.0 (3.2)
100	200	92.6 (3.0)	92.6 (2.1)	68 (10)	95.1 (1.9)	77.8 (7.8)	93.79 (0.97)
200	400	92.0 (1.8)	93.5 (1.6)	66.5 (8.5)	95.7 (1.0)	76.8 (5.6)	94.6 (1.1)
300	600	90.3 (1.3)	91.9 (1.3)	63.8 (4.8)	95.6 (1.0)	74.6 (2.9)	93.8 (1.1)
500	1000	84.6 (1.4)	89.33 (0.76)	63.3 (1.7)	95.33 (0.59)	72.4 (1.1)	92.23 (0.64)

previous configurations (Sec. 5.2.1). Then, we compare the performance of our approach against state-of-the-art face recognition methods (Sec. 5.3). Finally, the effect of openness is assessed (Sec. 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 IoI to N_U 1:2. Results are shown in Tabs. 1 and 2. We measure initial (non-adaptation) and final (after adaptation) performance of OSDe-SVM, using the previously described experimental set-up (Sec. 4.2). It is important to remark that non-adaptation means that ensembles do not incorporate new SVMs apart from the initial one. Thus,

Table 2: Performance over different universe sizes (N_U), while preserving the ratio with the number (N) of IoI. Values are expressed as $\mu(\sigma)$, where μ stands for mean and σ for standard deviation. Test performed on YouTube Faces dataset.

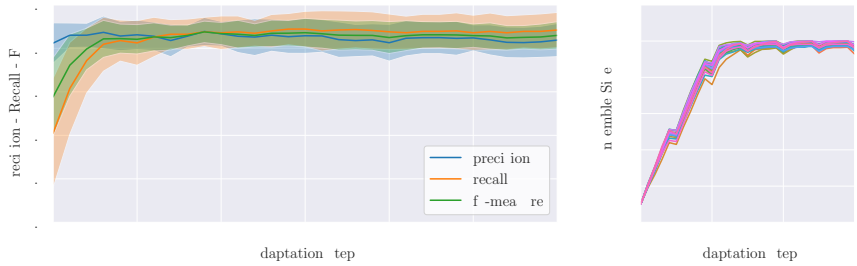
		Precision		Recall		F1	
N	N_U	Initial	Final	Initial	Final	Initial	Final
10	20	79 (10)	73.3 (9.5)	86 (11)	89.4 (9.4)	81.9 (8.2)	80.2 (8.1)
20	40	89.4 (5.8)	87.0 (6.2)	86.0 (8.2)	91.2 (6.4)	87.3 (5.0)	88.8 (4.9)
30	60	91.8 (4.4)	90.5 (5.1)	84.6 (7.5)	91.8 (5.7)	87.8 (4.5)	91.0 (4.4)
50	100	93.6 (2.9)	91.7 (3.7)	84.1 (5.4)	90.3 (4.0)	88.5 (3.2)	90.9 (3.2)
100	200	93.7 (1.5)	91.9 (1.7)	84.3 (6.1)	90.8 (2.8)	88.6 (3.4)	91.3 (1.8)
200	400	92.3 (1.6)	93.81 (0.48)	84.1 (2.7)	91.5 (1.2)	87.94 (0.71)	92.62 (0.88)

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 IoI 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, confusions 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, L8 in Alg. 1), when the IoI is 10 the Weibull fit is done with only 5 points.

Finally, while OSDe-SVM presents benefits in both datasets, COX is the



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

Figure 8: Evolution of OSDe-SVM for the case of 50 IoI over an universe of 100 identities. Test performed on COX Face Database.

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 IoI in a universe of 100

This experiment was aimed at performing a detailed study of one of the previous cases (50 IoI in a universe of 100) to fully understand OSDe-SVM behaviour. First, we will study its detailed temporal evolution in Sec. 5.2.1. After that, we complement this study by exploring the behaviour of two fundamental parts of OSDe-SVM: self-healing (Sec. 3.4) and the decision threshold T_W (Sec. 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. 8a) 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

Table 3: Comparison of OSDe-SVM with and without the self-healing module. Test performed on COX Face Database. Performance is expressed as $\mu(\sigma)$, where μ stands for mean and σ for standard deviation. The number of tests used to compute these values are the ones explained in Section 4.2.

	Precision		Recall		F1	
Conf.	Initial	Final	Initial	Final	Initial	Final
With SH	91.2 (4.7)	91.9 (3.8)	70 (13)	94.2 (4.1)	78.8 (9.0)	93.0 (3.2)
Without SH	91.2 (4.0)	90.8 (3.7)	70 (11)	94.5 (3.5)	79.0 (7.8)	92.6 (2.9)

lesser extent. Besides, Fig. 8b shows the evolution of the average ensemble size for each of the splits (Alg. 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 Tab. 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 Sec. 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*.

Table 4: Initial and final performances for different values of the decision threshold (T_W) over the Weibull distribution. Test performed on COX Face Database.

	Precision		Recall		F1	
T_W	Initial	Final	Initial	Final	Initial	Final
0.100	69.5 (5.2)	70.9 (4.7)	91.4 (5.8)	97.3 (2.3)	78.8 (4.8)	81.9 (3.4)
0.010	91.2 (4.7)	91.9 (3.8)	70 (13)	94.2 (4.1)	78.8 (9.0)	93.0 (3.2)
0.001	97.0 (3.6)	95.7 (3.2)	48 (12)	83.2 (5.7)	63 (11)	88.9 (4.1)

Table 5: Comparison against state-of-the-art face recognition models: FaceNet [25] and RN100-AF (ArcFace) [46], (for the case of 50 IoI in an universe of 100) using their proposed metrics for classification and a simple threshold (TH) to determine the unknown. We represent standard deviation within brackets. Test performed on COX Face Database.

Method	Precision	Recall	F1-measure
FaceNet +Euclidean+TH	38.7 (7.4)	71.3 (9.4)	49.0 (8.7)
RN100-AF +Cosine+TH	77.3 (9.9)	86 (11)	80.7 (6.5)
RN100-AF +OSDe-SVM, Initial	91.2 (4.7)	70 (13)	77.8 (7.8)
RN100-AF +OSDe-SVM, After Adapt.	91.9 (3.8)	94.2 (4.1)	93.0 (3.2)

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 performance for the case of 50 IoIs in a universe of 100, for 3 different values of T_W.

Results in Table 4 show that overall the properties of OSDe-SVM are maintained over each value, keeping its ability to improve performance without supervision. Therefore, varying T_W only move the point of the precision-recall curve in which OSDe-SVM operates.

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 (Tab. 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 Tab. 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 OSDe-SVM with RN100-AF embeddings preserves most of the discrimination power of the original decision function (cosine similarity). Finally, the enhanced performance provided by OSDe-SVM 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%:

$$O = 1 - \sqrt{2 \cdot \frac{N_{\text{training}}}{N_{\text{target}} + N_{\text{testing}}}}, \tag{3}$$

where N_{training} is the number of identities used on training (in our case, N), N_{target} is the number of identities to recognise (in our case, N as well) and

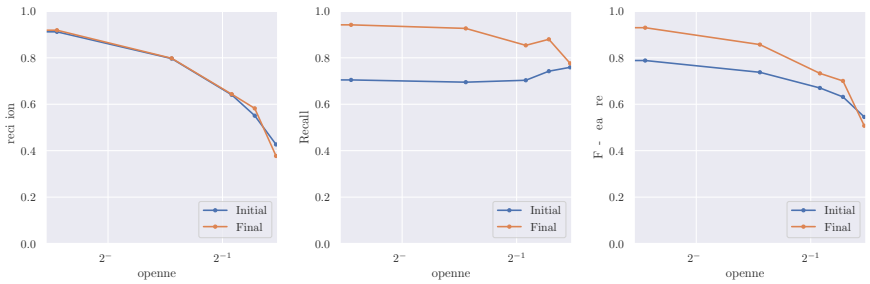


Figure 9: Performance openness dependence with fixed IoI (50), and universe ∈ {50, 100, 200, 400, 600, 1000}.

N_{testing} are the number of identities used on testing (in our case, N_U). Thus, Eq. 3 simplifies to:

$$O = 1 - \sqrt{2 \cdot \frac{N}{N + N_U}}. \tag{4}$$

To have a wide range of openness values, we selected a relatively low number of IoI (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 IoI), 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 $\approx 94\%$ F1-measure, depending on the amount of IoI 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 IoI in a universe of 1000 identities), where *unknown* identities appear many more times than IoI. 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 IoI into the recognition system. And, since classifiers are independently created, these additions would not have any adverse effect on previous knowledge.

Acknowledgements

This work has received financial support from the Spanish government (project PID2020-119367RB-I00); from the Xunta de Galicia, Consellara de Cultura, Educacin e Ordenacin 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 Lpez-Lpez has received financial support from the Xunta de Galicia and the European Union (European Social Fund - ESF).

References

[1] R. Kemker, M. McClure, A. Abitino, T. L. Hayes, C. Kanan, Measuring catastrophic forgetting in neural networks, AAAI'18/IAAI'18/EAAI'18, AAAI Press, 2018. doi:10.5555/3504035.3504450.

[2] M. Wang, W. Deng, Deep visual domain adaptation: A survey, Neurocomputing 312 (2018) 135 – 153. doi:10.1016/j.neucom.2018.05.083.

[3] L. Ren, X. Yuan, J. Lu, M. Yang, J. Zhou, Deep reinforcement learning with iterative shift for visual tracking, in: V. Ferrari, M. Hebert, C. Sminchisescu, Y. Weiss (Eds.), Computer Vision – ECCV 2018, Springer International Publishing, Cham, 2018, pp. 697–713. doi:10.1007/978-3-030-01240-3_42.

[4] J. He, R. Mao, Z. Shao, F. Zhu, Incremental learning in online scenario, in: IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2020, pp. 13923–13932. doi:arXiv:2003.13191.

[5] D. Sahoo, Q. Pham, J. Lu, S. C. H. Hoi, Online deep learning: Learning deep neural networks on the fly, in: International Joint Conference on Artificial Intelligence (IJCAI), 2018, pp. 2660–2666. doi:10.24963/ijcai.2018/369.

[6] X. Tao, X. Hong, X. Chang, S. Dong, X. Wei, Y. Gong, Few-shot class-incremental learning, in: IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2020, pp. 12180–12189. doi:10.1109/CVPR42600.2020.01220.

[7] J. M. Prez-Ra, X. Zhu, T. M. Hospedales, T. Xiang, Incremental few-shot object detection, in: IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2020, pp. 13843–13852. doi:10.1109/CVPR42600.2020.01386.

[8] M. McCloskey, N. J. Cohen, Catastrophic interference in connectionist networks: The sequential learning problem, Vol. 24 of Psychology of Learning and Motivation, Academic Press, 1989, pp. 109–165. doi:10.1016/S0079-7421(08)60536-8.

[9] Z. Huang, S. Shan, R. Wang, H. Zhang, S. Lao, A. Kuerban, X. Chen, A benchmark and comparative study of video-based face recognition on cox face database, IEEE Transactions on Image Processing 24 (12) (2015) 5967–5981. doi:10.1109/TIP.2015.2493448.

[10] Y. Guo, L. Zhang, Y. Hu, X. He, J. Gao, MS-Celeb-1M: A Dataset and Benchmark for Large-Scale Face Recognition, in: European Conference on Computer Vision (ECCV), 2016, pp. 87–102. doi:arXiv:1607.08221.

[11] M. Günther, L. E. Shafey, S. Marcel, Face Recognition in Challenging Environments: An Experimental and Reproducible Research Survey, Springer International Publishing, 2016, pp. 247–280. doi:10.1007/978-3-319-28501-6_11.

[12] E. Lpez-Lpez, X. M. Pardo, C. V. Regueiro, R. Iglesias, F. E. Casado, Dataset bias exposed in face verification, IET Biometrics 8 (4) (2019) 249–258. doi:10.1049/iet-bmt.2018.5224.

[13] G. Guo, N. Zhang, A survey on deep learning based face recognition,

Computer Vision and Image Understanding 189 (2019) 102805. doi:10.1016/j.cviu.2019.102805.

[14] S. Disabato, M. Roveri, Learning convolutional neural networks in presence of concept drift, in: International Joint Conference on Neural Networks (IJCNN), 2019, pp. 1–8. doi:10.1109/IJCNN.2019.8851731.

[15] D. Maltoni, V. Lomonaco, Continuous learning in single-incremental-task scenarios, Neural Networks 116 (2019) 56 – 73. doi:10.1016/j.neunet.2019.03.010.

[16] W. J. Scheirer, A. de Rezende Rocha, A. Sapkota, T. E. Boult, Toward open set recognition, IEEE Transactions on Pattern Analysis and Machine Intelligence 35 (7) (2013) 1757–1772. doi:10.1109/TPAMI.2012.256.

[17] M. Gnther, S. Cruz, E. M. Rudd, T. E. Boult, Toward open-set face recognition, in: IEEE Conference on Computer Vision and Pattern Recognition Workshops (CVPRW), 2017, pp. 573–582. doi:10.1109/CVPRW.2017.85.

[18] D. Yarowsky, Unsupervised word sense disambiguation rivaling supervised methods, in: Annual Meeting on Association for Computational Linguistics (ACL), 1995, pp. 189–196. doi:10.3115/981658.981684.

[19] C. Geng, S. Huang, S. Chen, Recent advances in open set recognition: A survey, IEEE Transactions on Pattern Analysis and Machine Intelligence (2020) 1–1doi:10.1109/TPAMI.2020.2981604.

[20] Z. Ge, S. Demyanov, Z. Chen, R. Garnavi, Generative openmax for multi-class open set classification, in: British Machine Vision Conference Proceedings (BMVC), 2017. doi:10.5244/C.31.42.

[21] P. Perera, V. I. Morariu, R. Jain, V. Manjunatha, C. Wigington, V. Ordonez, V. M. Patel, Generative-discriminative feature representations for open-set recognition, in: Conference on Computer Vision and Pattern Recognition (CVPR), 2020, pp. 11811–11820.

[22] S. Coles, Classical Extreme Value Theory and Models, Springer London, London, 2001, pp. 45–73. doi:10.1007/978-1-4471-3675-0_3.

[23] E. M. Rudd, L. P. Jain, W. J. Scheirer, T. E. Boult, The extreme value machine, IEEE Transactions on Pattern Analysis and Machine Intelligence 40 (3) (2018) 762–768. doi:10.1109/TPAMI.2017.2707495.

[24] G. Salomon, A. Britto, R. H. Vareto, W. R. Schwartz, D. Menotti, Open-set face recognition for small galleries using siamese networks, in: International Conference on Systems, Signals and Image Processing (IWSSIP), 2020, pp. 161–166. doi:10.1109/IWSSIP48289.2020.9145245.

[25] F. Schroff, D. Kalenichenko, J. Philbin, Facenet: A unified embedding for face recognition and clustering, in: IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2015, pp. 815–823. doi:10.1109/CVPR.2015.7298682.

[26] G. I. Parisi, R. Kemker, J. L. Part, C. Kanan, S. Wermter, Continual lifelong learning with neural networks: A review, Neural Networks 113 (2019) 54 – 71. doi:10.1016/j.neunet.2019.01.012.

[27] T. L. Hayes, C. Kanan, Lifelong machine learning with deep streaming linear discriminant analysis, in: IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW), 2020, pp. 887–896. doi:10.1109/CVPRW50498.2020.00118.

[28] J. Kirkpatrick, R. Pascanu, N. Rabinowitz, J. Veness, G. Desjardins, A. A. Rusu, K. Milan, J. Quan, T. Ramalho, A. Grabska-Barwinska, D. Hassabis, C. Clopath, D. Kumaran, R. Hadsell, Overcoming catastrophic forgetting in neural networks, Proceedings of the National Academy of Sciences 114 (13) (2017) 3521–3526. doi:10.1073/pnas.1611835114.

[29] J. Zhang, J. Zhang, S. Ghosh, D. Li, S. Tasci, L. Heck, H. Zhang, C. . Jay Kuo, Class-incremental learning via deep model consolidation, in: IEEE

Winter Conference on Applications of Computer Vision (WACV), 2020, pp. 1120–1129. doi:10.1109/WACV45572.2020.9093365.

[30] G. M. van de Ven, H. T. Siegelmann, A. S. Tolias, Brain-inspired replay for continual learning with artificial neural networks, Nature Communications 11 (1) (2020) 4069. doi:10.1038/s41467-020-17866-2.

[31] S. Ud Din, J. Shao, J. Kumar, W. Ali, J. Liu, Y. Ye, Online reliable semi-supervised learning on evolving data streams, Information Sciences 525 (2020) 153 – 171. doi:10.1016/j.ins.2020.03.052.

[32] Y. Li, Y. Wang, Q. Liu, C. Bi, X. Jiang, S. Sun, Incremental semi-supervised learning on streaming data, Pattern Recognition 88 (2019) 383 – 396. doi:10.1016/j.patcog.2018.11.006.

[33] M. D. la Torre, E. Granger, P. V. Radtke, R. Sabourin, D. O. Gorodnichy, Partially-supervised learning from facial trajectories for face recognition in video surveillance, Information Fusion 24 (2015) 31 – 53. doi:10.1016/j.inffus.2014.05.006.

[34] P. H. Pisani, A. Mhenni, R. Giot, E. Cherrier, N. Poh, A. C. P. d. L. Ferreira de Carvalho, C. Rosenberger, N. E. B. Amara, Adaptive biometric systems: Review and perspectives, ACM Comput. Surv. 52 (5) (2019) 102:1–102:38. doi:10.1145/3344255.

[35] G. Orr, G. L. Marcialis, F. Roli, A novel classification-selection approach for the self updating of template-based face recognition systems, Pattern Recognition 100 (2020) 107–121. doi:10.1016/j.patcog.2019.107121.

[36] A. Franco, D. Maio, D. Maltoni, Incremental template updating for face recognition in home environments, Pattern Recognition 43 (8) (2010) 2891 – 2903. doi:10.1016/j.patcog.2010.02.017.

[37] F. Pernici, A. D. Bimbo, Unsupervised incremental learning of deep descriptors from video streams, in: IEEE International Conference on Mul-

timedia Expo Workshops (ICMEW), 2017, pp. 477–482. doi:10.1109/ICMEW.2017.8026276.

[38] R. Coop, A. Mishtal, I. Arel, Ensemble learning in fixed expansion layer networks for mitigating catastrophic forgetting, IEEE Transactions on Neural Networks and Learning Systems 24 (10) (2013) 1623–1634. doi:10.1109/TNNLS.2013.2264952.

[39] B. Krawczyk, L. L. Minku, J. Gama, J. Stefanowski, M. Woniak, Ensemble learning for data stream analysis: A survey, Information Fusion 37 (2017) 132 – 156. doi:10.1016/j.inffus.2017.02.004.

[40] H. M. Gomes, J. P. Barddal, F. Enembreck, A. Bifet, A survey on ensemble learning for data stream classification, ACM Comput. Surv. 50. doi:10.1145/3054925.

[41] N. Liang, G. Huang, P. Saratchandran, N. Sundararajan, A fast and accurate online sequential learning algorithm for feedforward networks, IEEE Transactions on Neural Networks 17 (6) (2006) 1411–1423. doi:10.1109/TNN.2006.880583.

[42] N. Dvornik, J. Mairal, C. Schmid, Diversity with cooperation: Ensemble methods for few-shot classification, in: 2019 IEEE/CVF International Conference on Computer Vision (ICCV), 2019, pp. 3722–3730. doi:10.1109/ICCV.2019.00382.

[43] X. Zhang, F. Yan, Y. Zhuang, H. Hu, C. Bu, Using an ensemble of incrementally fine-tuned cnns for cross-domain object category recognition, IEEE Access 7 (2019) 33822–33833. doi:10.1109/ACCESS.2019.2903550.

[44] Y. Guo, X. Wang, P. Xiao, X. Xu, An ensemble learning framework for convolutional neural network based on multiple classifiers, Soft Computing 24. doi:10.1007/s00500-019-04141-w.

[45] J. Wang, Z. Mo, H. Zhang, Q. Miao, Ensemble diagnosis method based on transfer learning and incremental learning towards mechanical big data,

Measurement 155 (2020) 107517. doi:10.1016/j.measurement.2020.107517.

[46] J. Deng, J. Guo, N. Xue, S. Zafeiriou, Arcface: Additive angular margin loss for deep face recognition, in: IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2019, pp. 4685–4694. doi:h10.1109/CVPR.2019.00482.

[47] H. Liu, X. Zhu, Z. Lei, S. Z. Li, Adaptiveface: Adaptive margin and sampling for face recognition, in: IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2019, pp. 11939–11948. doi:10.1109/CVPR.2019.01222.

[48] W. Scheirer, A. Rocha, R. Micheals, T. Boult, Robust fusion: Extreme value theory for recognition score normalization, in: European Conference on Computer Vision (ECCV), 2010, pp. 481–495. doi:10.1007/978-3-642-15558-1_35.

[49] N. Li, Y. Yu, Z.-H. Zhou, Diversity regularized ensemble pruning, in: Machine Learning and Knowledge Discovery in Databases, 2012, pp. 330–345. doi:10.1007/978-3-642-33460-3_27.

[50] Z. Cheng, X. Zhu, S. Gong, Surveillance face recognition challenge, arXiv preprint arXiv:1804.09691.

[51] R. Goh, L. Liu, X. Liu, T. Chen, The CMU Face In Action (FIA) Database, in: W. Zhao, S. Gong, X. Tang (Eds.), Analysis and Modelling of Faces and Gestures, Springer, 2005, pp. 255–263. doi:10.1007/11564386_20.

[52] L. Wolf, T. Hassner, I. Maoz, Face recognition in unconstrained videos with matched background similarity, in: CVPR 2011, 2011, pp. 529–534. doi:10.1109/CVPR.2011.5995566.

[53] K. Zhang, Z. Zhang, Z. Li, Y. Qiao, Joint face detection and alignment using multitask cascaded convolutional networks, IEEE Signal Processing Letters 23 (10) (2016) 1499–1503. doi:10.1109/LSP.2016.2603342.

Biographies

Eric Lopez-Lopez received his Master Degree in Computer Vision from Autonumous University of Barcelona. He is now a PhD canditate at University of A Coruña, working around Adaptive Learning topics. Apart from that, his research interests include Face Biometrics, Online Learning and Self-supervised learning techniques.

Carlos V. Regueiro received the B.S. and Ph.D. degrees in Physics from the University of Santiago de Compostela, Spain, in 1992 and 2002, respectively. Since December 1993 he has been an Associated Professor in the Faculty of Computer Science at the University of A Coruña, Spain, where he teaches undergraduates and graduates courses on computer architecture. His research interests focus on control architectures, perception, control, localization, navigation and machine learning in mobile robotics.

Xose M. Pardo is an associate professor of Software and Computer Systems at the University of Santiago de Compostela (Spain). He received a Ph.D. in Physics from this university in 1998, for his research on 3D medical image analysis. He has been a postdoctoral research fellow at the Computer Vision Center of Barcelona (Spain) and the INRIA Sophia Antipolis (France), between 1998 and 2000. In recent years, his interest has shifted to biologically inspired computer vision, and includes visual saliency, object and scene recognition, human activity recognition and machine learning. At the moment, they are mostly working on projects related to robot vision, object and scene recognition, photogrammetry and visual inspection.

Declaration of interests

☒ 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.

☐The authors declare the following financial interests/personal relationships which may be considered as potential competing interests: