Expert Systems With Applications 174 (2021) 114734

Contents lists available at ScienceDirect

Expert Systems With Applications

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

Towards a self-sufficient face verification system

Eric Lopez-Lopez [a,*], Carlos V. Regueiro [a], Xosé M. Pardo [c], Annalisa Franco [b], Alessandra Lumini [b]

[a] Universidade da Coruña, CITIC, Computer Architecture Group, Spain
[b] DISI – Dept. of Computer Science and Engineering, Universitá di Bologna, Italy
[c] CiTIUS, Universidade de Santiago de Compostela, Spain

ARTICLE INFO

Keywords:
Adaptive biometrics
Video surveillance
Video-to-video face verification
Unsupervised learning
Incremental learning

ABSTRACT

The absence of a previous collaborative manual enrolment represents a significant handicap towards designing a face verification system for face re-identification purposes. In this scenario, the system must learn the target identity incrementally, using data from the video stream during the operational authentication phase. So, manual labelling cannot be assumed apart from the first few frames. On the other hand, even the most advanced methods trained on large-scale and unconstrained datasets suffer performance degradation when no adaptation to specific contexts is performed. This work proposes an adaptive face verification system, for the continuous re-identification of target identity, within the framework of incremental unsupervised learning. Our Dynamic Ensemble of SVM is capable of incorporating non-labelled information to improve the performance of any model, even when its initial performance is modest. The proposal uses the self-training approach and is compared against other classification techniques within this same approach. Results show promising behaviour in terms of both knowledge acquisition and impostor robustness.

1. Introduction

1.1. Motivation

Video-surveillance is one of the most demanding contexts of operation for face verification systems. As with any other biometric system, the standard approach is divided into two separated phases: *enrolment* and *test/verification* (Pisani et al., 2019). During the first phase, samples of the target identity (*genuine*) are registered into the system to create a model. During the second one, the system receives identity queries with the task of checking if identity is *genuine* or not (*impostor*).

The conditions in which each of these phases are conducted can drastically change the challenge we are facing. For instance, in terms of *enrolment* stage, some cases allow executing this phase in a separated way (user collaboration is often required) where high-quality photographs or videos are acquired (Huang et al., 2015; Bashbaghi, Granger, Sabourin, & Bilodeau, 2017; Wang, Shan, Chen, & Gao, 2008; Dewan, Granger, Marcialis, Sabourin, & Roli, 2016; Chen, Wang, Xiao, & Cai, 2014). In these conditions, state-of-the-art systems seem to perform astonishingly well (≈ 90–99% Rank-1 Identification Rates (Zhang et al., 2019; Bashbaghi, Granger, Sabourin, & Parchami, 2019)), something that makes this specific problem almost solved. Other cases do not allow this kind of *enrolment* (e.g. criminal watch-list, lost children, disoriented older people, etc.). Thus, the only option is to use data from the video stream as enrolment source (Huang et al., 2015; la Torre, Granger, Radtke, Sabourin, & Gorodnichy, 2015; Franco, Maio, & Maltoni, 2010) which substantially lowers the system performance.

1.2. The problem

Working with only video-frame data presents two major challenges. First, this type of data presents a wide range of context-dependent and time-dependent variations (e.g. camera parameters, poses, illumination, target distance, etc.). And second, the gathering of extensive amounts of labelled context-specific data is not very realistic.

Deep learning approaches have brought outstanding performance to general recognition applications, where plenty of annotated data are available. In contexts where target domain labelled data are scarce, knowledge transfer from source to target domains can alleviate the problem (Sohn et al., 2017; Crosswhite et al., 2018; Pernici, Bartoli,

* Corresponding author.
E-mail addresses: eric.lopez@udc.es (E. Lopez-Lopez), carlos.vazquez.regueiro@udc.es (C.V. Regueiro), xose.pardo@usc.es (X.M. Pardo), annalisa.franco@unibo.it (A. Franco), alessandra.lumini@unibo.it (A. Lumini).

https://doi.org/10.1016/j.eswa.2021.114734
Received 21 January 2020; Received in revised form 12 January 2021; Accepted 12 February 2021
Available online 23 February 2021

Bruni, & Del Bimbo, 2018; Wang & Deng, 2018). In these contexts, even deep learning methods suffer from performance degradation when the differences between both domains are highly marked (Pisani et al., 2019; López-López, Pardo, Regueiro, Iglesias, & Casado, 2019; Tommasi, Patricia, Caputo, & Tuytelaars, 2017; Li et al., 2018; Bianco, 2017). In any case, the need of at least a partial labelling process persists, and a general solution to transfer learning still remains open (Masi, Wu, Hassner, Natarajan, & del Rey, 2018).

One way of tackling both of the previous challenges at the same time is through adaptation (Pisani et al., 2019). Since target domain data becomes gradually abundant during the *test/verification* stage, it seems natural to conceive learning as an incremental process (Franco et al., 2010; Ditzler, Roveri, Alippi, & Polikar, 2015; Lomonaco & Maltoni, 2016) rather than other learning modes (Mian, 2011) (e.g. batch learning). Despite being possible to perform under different supervision levels, here, the real value of adaptation arises when it is considered as an unsupervised process (Pisani et al., 2019; Krawczyk, Minku, Gama, Stefanowski, & Woźniak, 2017). In this direction, the literature proposes semi-supervised incremental learning approaches (la Torre et al., 2015; Franco et al., 2010; Villamizar, Sanfeliu, & Moreno-Noguer, 2019). However, despite their reduced label requirements, they often require a human operator in the loop to assist with the most challenging samples (by given additional labels).

In the context of adaptive biometrics (Pisani et al., 2019), the self-training approach (Yarowsky, 1995) is an interesting strategy to reduce labels requirements drastically. Using an incremental learning point of view, it relies on pseudo-labels given by the classifier at the moment to decide whether to update the template or not. In the specific context of video-based face verification, this strategy, in combination with two established incremental learning techniques, has proven its viability in Lopez-Lopez, Franco, and Lumini (2019) work. Besides, self-training has also been recently used for domain adaptation purposes (Zou, Yu, Liu, Kumar, & Wang, 2019; Kim, Choi, Kim, & Kim, 2019), person Re-ID (Zhang, Cao, Shen, & You, 2019) or object detection (Roychowdhury et al., 2019).

1.3. The proposed approach and main contributions

This work aims to tackle the problem of progressively computing an efficient classifier for a video-to-video face verification (V2V-FV) setting where labels' availability is not enough to generate a robust model. For this purpose, we propose the Dynamic Ensemble of SVM, a method which creates and automatically improves/updates an ensemble of very-specific SVM classifiers. This kind of ensembles has proven to achieve remarkable results (Malisiewicz, Gupta, & Efros, 2011) on static supervised conditions. Here, we provide a novel decision mechanism aimed to incrementally generate the ensemble in a semi-supervised way (i.e. starting from a few labelled data, and then autonomously updating) and using only online target domain data. In this regard, the main contributions of the work are:

- The use of the self-updating approach in combination with the current most powerful feature representations as face re-identification system in the context of V2V-FV.
- The proposition of an ensemble-based adaptive biometric system called Dynamic Ensemble of SVM (De-SVM).

The rest of the paper is organised as follows. First, in Section 2, we cite the main approaches that have been of inspiration for our work. In Section 3, the proposed adaptive system is explained. After some technical details in Section 4, we show the experiments performed in Section 5. Finally, we expose the conclusions in Section 6.

2. Related work

Face verification in video-surveillance can be tackled from numerous points of view. Based on Huang et al. (2015), three main scenarios are distinguished. First, in the Still-to-Video face verification (S2V-FV) a system is queried using a still face image in order to find a video sequence where the same identity appears (Bashbaghi et al., 2017; Dewan et al., 2016; Chen et al., 2014; Bashbaghi et al., 2019). Second, in a Video-to-Still (V2S-FV) task, a system is queried using a video sequence to retrieve the same identity from a pool of still face images (Zhang et al., 2019; Bashbaghi et al., 2019; Wang et al., 2008). And finally, in Video-to-Video (V2V-FV) task, the task is to compare two video sequences to check if they contain the same identity or not (la Torre et al., 2015; Franco et al., 2010). As aforementioned, results achieved on recent databases like COX (Huang et al., 2015) when high-quality stills are used (either V2S or S2V conditions) convert the problem into an almost solved one (Bashbaghi et al., 2019; Zhang et al., 2019). However, V2V in low-labelled conditions adds an extra handicap not solved yet.

Template updating (a.k.a. incremental learning) is considered by this work as a strategy to take advantage of data acquired during the *test/verification* phase. Historically, this family of approaches has been focused on two slightly different tasks with different challenges (Gepperth & Hammer, 2016; Chefrour, 2019). On the one hand, modifying or adapting a complete model to deal with dynamic environments that can impair performance (Ditzler et al., 2015; la Torre et al., 2015); and, on the other hand, gradually improving the quality of a template created with a tiny amount of labelled data (Yarowsky, 1995). In any case, the fact of having changing models encloses new important challenges. The literature often refers to the *stability-plasticity* dilemma or the *exploitation-exploration* dilemma (Hoens, Polikar, & Chawla, 2012; Grossberg, 1988) as paradigmatic examples of these new challenges. Incremental learning systems should have plasticity for the integration of new relevant knowledge, but also stability to prevent the *catastrophic* forgetting of previous, but relevant, knowledge because of the addition of the new inputs. The proper balance between these two ideas is a highly challenging problem which, despite recent advances (Kirkpatrick et al., 2017), remains open (Kemker, McClure, Abitino, Hayes, & Kanan, 2018; Chefrour, 2019). In the unsupervised case, the template updating process (either adaptation or improvement) must be done without labels. Thus, the correct definition of a label inference procedure (to substitute human annotation) represents an additional handicap to the existing incremental learning challenges (Chefrour, 2019).

Self-training or self-updating is the approach used to update identity models in an unsupervised way. Firstly proposed in the scope of natural language processing (Yarowsky, 1995), these methods are rooted on the supposition that the classifier itself can do the genuine/impostor labelling, avoiding any supervision (Franco et al., 2010; Didaci, Marcialis, & Roli, 2014; Orrú, Marcialis, & Roli, 2020). This way, learning is also considered as an unsupervised incremental process where the actual model at the moment is the one that decides whether to update or not. Outside the biometric scope, self-training ideas help to minimise human annotation effort in network traffic classification (Fahad et al., 2019) or to incorporate unlabelled data from auxiliary information sources, like the internet, to improve object detectors (Radosavovic, Dollár, Girshick, Gkioxari, & He, 2018). The central dilemma with this family of approaches is about selecting the confidence level to update the template. A high confidence threshold could avoid accepting impostor identities but at the expense of only accepting too redundant information that does not contribute to improve performance. A low confidence threshold, on the contrary, would accept higher variance in samples, but at the cost of letting in more impostors in

the model.

Temporal Coherence in videos leads to the assumption that consecutive frames will almost always contain very similar information (Becker, 1999). The exploitation of temporal coherence is often highlighted as one of the keys for unsupervised learning (Franco et al., 2010; Pernici & Bimbo, 2017; Wang & Gupta, 2015). Based on the assumption that two patches connected by a track should have a similar visual representation in deep feature space since they probably belong to the same object, Siamese network architectures have been proposed to train DNNs in an unsupervised way (Wang & Gupta, 2015; Misra, Zitnick, & Hebert, 2016; Redondo-Cabrera & Lopez-Sastre, 2019). As temporal coherence assumption is valid in the video-surveillance scenario, once an identity is verified in a video frame, its visual tracking is enough to keep the target identified. Thus, visual tracking provides the supervision for labelling the identity in every frame of the sequence, even if pose and illumination change or partial occlusions happen. However, this does not apply to the transition between different sequences, i.e. when a cut takes place.

3. An adaptive biometric system for face verification in video-surveillance

In this paper, a novel adaptive biometric system is proposed. Using a general (hand-crafted or learned) feature extractor, the system self-updates classifiers using new target-identity samples without forgetting previously acquired knowledge. Since ensemble methods are well-acknowledged as approaches that fulfil these characteristics, we propose (and compare against other alternatives) a method based on the ensemble of SVM classifiers. Most of the studies in supervised contexts enhance SVM versions compared to the decision trees (Chefrour, 2019).

3.1. Self-updating general pipeline

From a general perspective, the self-updating approach is based on the following hypothesis: the use of pseudo-labels given by the model (M_{t-1}) at time t to drive the decision to update helps to improve performance. The considered scenario assumes that initially ($t = 0$) a few video frames of the *genuine* identity (short sequence given by a visual tracker) are selected to create the *template*. The quality of this template can be an important constraint to the performance of the system, as Lopez-Lopez et al. (2019) and Section 5.2.2 show. The availability of a group of negative samples to build the initial model (M_0) is also assumed.

As it is outlined in Algorithm 1 and the following Fig. 1, over time ($t = 1, 2, \ldots, T$), the system is queried with new video sequences (S_t) to verify the identity of the individuals (both *genuine* and *impostor*) appearing in them (Cohort Model, CM (la Torre et al., 2015)). Following an *query acceptance* adaptation criterion (Pisani et al., 2019), if M_{t-1} accepts the query sequence, following the previous hypothesis the sequence is used to create the model M_t. In the opposite case, M_t remains the same.

3.2. Decision rules

Given the previous pipeline and joined to the fact of working with videos as data source, several decision rules need to be defined. These rules control when an identity is verified or not, and if so how updates are performed.

- *Frame Decision Rule (FDR)*. This rule assigns a score to each frame of the query video sequence. It can be the outcome provided by a single classifier (e.g. SVM score, distance in a nearest-neighbour algorithm, or a softmax in a DNN), or the fused output in the case of an ensemble of classifiers.
- *Sequence Decision Rule (SDR)*. This rule assigns a unique score to the query video sequence based on the FDR individual scores in each frame. Identities will be verified by fixing a certain confidence level to this score. The greediness or cautiousness in this fixation impacts the *stability-plasticity* dilemma.
- *Update Rule (UR)*. This rule defines how new information is used to enhance the current model incrementally.

Algorithm 1 and Fig. 1 illustrates the role of each decision rule in the self-update pipeline. The actual implementation choices for each of the explored models (in the following Section 3.4) are shown in Table 1. Note that median is used as a central tendency measure for its analogy with a majority vote. However, mean was also tested with almost identical results (more detailed information on Appendix C).

3.3. The proposed dynamic ensemble of SVM (De-SVM)

The classification power of ensembles of very-specific classifiers was first proven by Malisiewicz et al. (2011). There, an ensemble of exemplar Support Vector Machines (SVM) classifiers, each of them (exemplar-SVM) trained with just one positive sample and a great number of negative samples, is used in the frame of object detection. The idea

Algorithm 1: The implementation of the self-updating strategy.

input : Query Sequences = $\{S_0, S_1, \ldots, S_T\}$, negativeSet, type_of_model, TH (operational threshold).
output: Self-updated model, M_T

M_0 = createModel (S_0, negativeSet, type_of_model)
for t = *1,2,...,T* **do**
  $score$ = evaluateSequence (S_t, M_{t-1}, SDR, FDR)
  if $score > TH$ **then**
    S_t *assumed to be a genuine sequence*
    M_t = updateModel (S_t, M_{t-1}, UR)
  else
    S_t *assumed to be an impostor sequence*
    $M_t = M_{t-1}$
  end
end

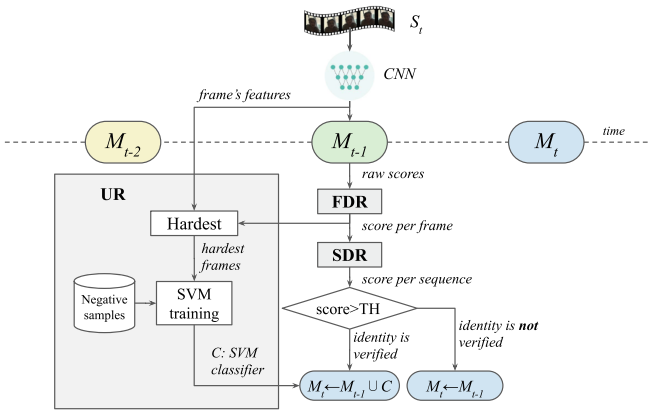


Fig. 1. The pipeline of De-SVM within the self-updating general strategy. The model (M_{t-1}) is updated based on its decision over the query sequences (S_t) to generate a new model (M_t). Three different rules are involved in the decision of updating (FDR and SDR) and the way of updating (UR).

behind this strategy is to have an ensemble of very-specific classifiers whose combined decision will be able to overcome the over-fitting. In the specific case of face verification in video-surveillance, a similar approach is used in Bashbaghi et al. (2017) where an identity-specific ensemble of exemplar SVMs has been proposed to recognise a target identity among distractors in the case of S2V face recognition. Each exemplar is built during enrolment from a single target sample and multiple distractors' samples, to represent the diversity of the same identity appearance due to various perturbation factors. Ensembles of (exemplar) SVMs leverage the intuition according to which a pool of simple classifiers, one for each training sample, can outperform a single and complex one (Becattini, Seidenari, & Del Bimbo, 2017). Besides, another advantage of ensemble-based methods is the addition of an extra-point of flexibility in the algorithms. One could potentially control how each member of the ensemble performs, allowing classifiers substitutions or removals whenever needed to keep the ensemble size bounded.

Our contribution uses these previous ideas as a basis. From them, the aim is to create the identity-specific ensemble incrementally using the self-updating strategy. Instead of using exemplar-SVM, each ensemble member will be trained using n positive samples instead of just one. In any case, the number of positive samples will remain low ($n = 5$ in our experiments) following the Exemplar-SVM philosophy. The need to use more positive samples than just one is shown in a previous work (Lopez-Lopez et al., 2019) where the quality of the initial template is crucial when designing a method based on the self-updating strategy. The transformation to an incremental classifier consists of adding new classifiers to the ensemble whenever the ensemble needs to be updated.

Table 1
Summary of the decision rules with each method.

Method	Decision rules	Method used
De-SVM	*Frame Decision Rule*	Ensemble fusion rule (Median)
	Sequence Decision Rule	Median
	Update Rule	Add a new classifier with the hardest samples
SVM	*Frame Decision Rule*	Raw score
	Sequence Decision Rule	Median
	Update Rule	Retrain the classifier with D^{t+1}
I-SVM	*Frame Decision Rule*	Raw score
	Sequence Decision Rule	Median
	Update Rule	Partial fit using the query sequence
OS-ELM	*Frame Decision Rule*	SoftMax
	Sequence Decision Rule	Median
	Update Rule	Partial fit using the query sequence

In our case, the number of possible updates keeps relatively bounded. This keeps out of the scope a procedure to limit the number of ensemble classifiers. Nevertheless, for a real application, one should limit it following either substitution or removing strategies (Krawczyk et al., 2017).

Fig. 1 depicts the pipeline of De-SVM. Following the self-updating paradigm, the update decision is made by the ensemble at the moment. The median is used as *FDR* to give a score to each frame (which in practice corresponds to a majority voting). Afterwards, the median of the sequence's frames FDR scores is computed again as SDR. If the identity is verified (based on the operational threshold), the ensemble adds a new classifier following the UR.

The $n = 5$ samples used to create the new member of the ensemble will be pooled from the query sequence's frames. In order to enhance diversity within the ensemble, the hardest frames (the ones with the *worst* score obtained using the current model, using the *FDR*) are selected as positive samples to train (against a large number of negative samples) the next classifier of the ensemble.

3.4. Other explored classification methods

The self-update approach has a 'wrapper algorithm' (Zhu, 2005) nature that in practice converts a supervised classification method in an unsupervised one (by using the previous decision rules). Therefore, one could use different classification methods within this same approach. In this work, we explore different supervised classification methods (both incremental and batch-based):

3.4.1. Linear soft-margin support vector machine (SVM)

The Linear Soft-Margin Support Vector Machine (SVM) is a batch-based binary classification method (Cortes & Vapnik, 1995) widely used in many applications. Given a set of N labelled training feature vectors $\mathbf{x}_i$, the classifier finds the optimal hyperplane which separates both classes of the binary problem. This classification technique has lost a bit of its prominence in favour of CNN's. Nevertheless, it continues to be used on top of CNN-based features. In the specific context of face recognition, there are numerous examples of SVM-based methods in the recent literature (Crosswhite et al., 2018; Wu, Zuo, Lin, Jia, & Zhang, 2018; Dhamecha, Noore, Singh, & Vatsa, 2019).

3.4.2. Incremental SVM (I-SVM)

This is an incremental implementation of the previous method. Here, training data is provided sequentially instead of the batch mode in which all examples are available at once. New training data is incorporated when it is available, without re-training from scratch. In Kivinen, Smola, and Williamson (2004), a simple and computationally efficient algorithm, based on the classical Stochastic Gradient Descent, was developed to update the hyper-plane incrementally parameters of the solution for online learning applications.

3.4.3. Online sequential extreme learning machine (OS-ELM)

It is an incremental implementation of the regular Extreme Learning Machine (ELM) problem. The ELM builds a Single Layer Feed-forward Network (SLFN) with $\widetilde{N}$ hidden nodes to approximate a set of N labelled training feature vectors such that:

$$f_{\widetilde{N}}\left(x_j\right) = \sum_{i=1}^{\widetilde{N}} \beta_i G\left(\mathbf{a}_i, b_i, \mathbf{x}_j\right) = y_j, \qquad j = 1, \ldots, N \tag{1}$$

where $\mathbf{a}_i$ and b_i are the parameters of the hidden nodes activation function G (additive or RBF); and β_i the weight that connects the i-th hidden node with the output. It is showed that (1) is satisfied for any randomly assigned values of the node parameters ($\mathbf{a}_i$ and b_i) by analytically computing the weight β_i, as long as $N \geqslant \widetilde{N}$.

In the specific case of OS-ELM, the approach is specifically adapted to

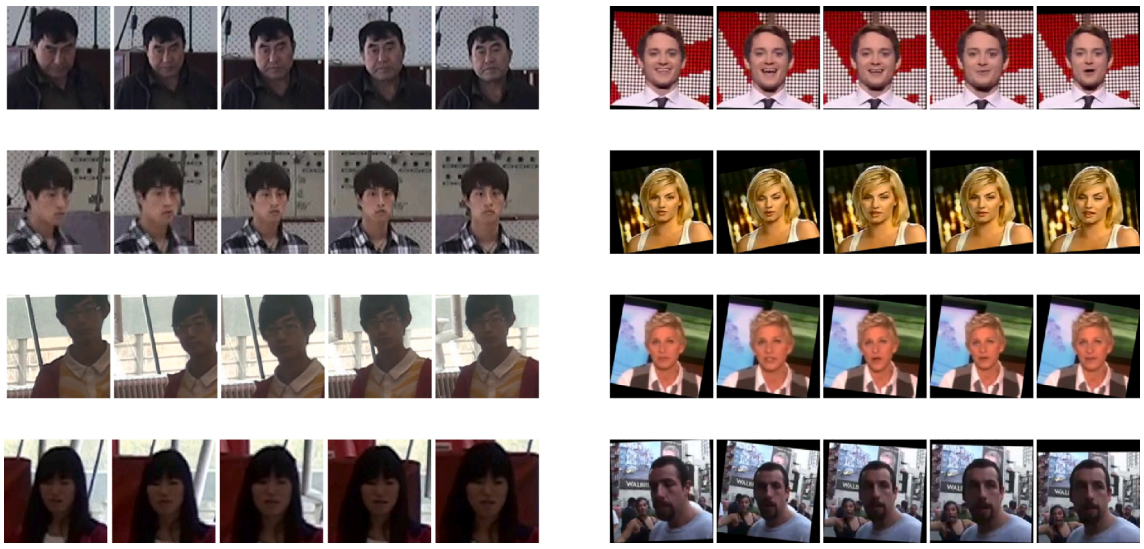


Fig. 2. Samples of both datasets, COX (left) and YTF (right).

compute and update the weight values sequentially as more data is becoming available ('chunk-by-chunk' or one-by-one) (Liang, Huang, Saratchandran, & Sundararajan, 2006). In our case, a sigmoid function is used as an activation function, and the number of hidden nodes is empirically fixed at $\widetilde{N} = 80$.

4. Methodology

4.1. Datasets

Two video datasets, with sequences of different users, have been used in the experiments (Fig. 2):

COX Face database (Huang et al., 2015) (COX). This dataset gathers video frames of 1000 identities. There are 3 video sequences captured by 3 different cameras (`cam1`, `cam2` and `cam3`). The subjects were asked to walk over an S-path, and their images were captured under variable lighting, pose, scale conditions, and a considerable amount of blur. Each camera recorded a part of the path, without temporal overlapping between them. In this dataset, the number of sequences of each identity is quite limited (3 sequences per subject). In order to mitigate this limitation, each video sequence was split in several of sub-sequences without alteration of the temporal order (Fig. 3).

YouTube Faces (Wolf, Hassner, & Maoz, 2011) (YTF). This dataset contains a total of 3425 videos downloaded from the YouTube platform of 1595 different identities. Each identity appears in between 1 and 6 different videos captured under completely different conditions. Table 2 contains the distribution of video frames per identity after face detection. As with the previous dataset, to augment the number of video sequences to query the system, each video sequence has been split into several sub-sequences while keeping the temporal coherence.

4.2. Face detection and feature extractor

A face detection technique is applied over every frame to discard the background part of images and for alignment purposes. The tool provided in the Dlib library (King, 2009) was selected for this task. After that, we will extract a feature vector of each face using a pre-trained ResNet-34 network (He, Zhang, Ren, & Sun, 2016) with just 29 convolution layers (RN29). The classifier layers have been removed from the network, as provided by Dlib (King, 2009), giving a feature vector of 128 dimensions. The network has been trained using a combination of the SCRUB dataset (Ng & Winkler, 2014) and the VGG-Face dataset (Parkhi, Vedaldi, & Zisserman, 2015). This implementation achieves an accuracy of 99.38% in the LFW dataset (which is comparable to the face verification state-of-the-art) and has shown quite desirable properties in terms of robustness to non-identity related variations (López-López et al., 2019).

4.3. Testing protocol

Using the protocol proposed by the COX database as an inspiration, each dataset was divided into three different subsets (See Table 3–5):

(a)The **train subset** is composed of the face images used as a negative set and as a validation set in the learning process. In the actual implementation, this negative set will be a random subset of 1000 samples from the whole train set.

- In the case of COX Face database, this subset is composed of the face images of 300 identities taken from each available camera.

Table 2
YouTube Faces distribution of the amount of videos per person after the face detection phase.

#videos	1	2	3	4	5	6
#people	588	472	305	167	51	8

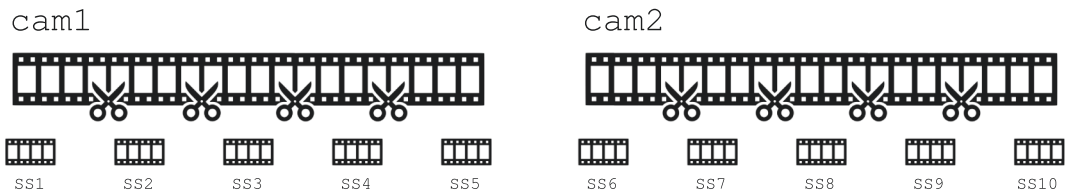


Fig. 3. Example of division of `cam1` and `cam2` in the COX Face database to generate the query sub-sequences (SS stands for sub-sequence).

Table 3
How the datasets' identities are divided in order to generate each subset defined in Section 4.3.

		Genuine				Impostor			
COX		still	cam1	cam2	cam3	still	cam1	cam2	cam3
	Train	0	0	0	0	300	300	300	300
	Gallery	0	700	700	0	0	0	0	0
	Probe	0	0	0	700	0	0	0	700
YTF		⩾4 videos		< 4 videos		⩾4 videos		< 4 videos	
	Train	0		0		0		1365	
	Gallery	226		0		0		0	
	Probe	226		0		226		0	

Table 4
COX: Study on FAR point of operational threshold De-SVM (template size = 5 frames).

Operational Threshold	TAR@FAR1		TAR		FAR	
	Initial	Final	Initial	Final	Initial	Final
1%	37.17±0.58	74.08±0.67	30.92±0.32	57.1±1.1	0.568±0.061	0.118±0.031
3%	37.17±0.58	82.61±0.68	48.41±0.21	79.08±0.52	2.407± 0.048	0.650±0.024
5%	37.17±0.58	85.45±0.25	59.01±0.37	88.14±0.24	5.01±0.13	1.714±0.052
7%	37.17±0.58	86.03±0.48	62.39±0.31	89.80±0.25	6.23±0.16	2.53±0.11
10%	37.17±0.58	86.69±0.40	68.90±0.22	93.01±0.43	9.34±0.10	4.90±0.10

- In the case of YouTube Faces database, this subset contains data from the identities that have less than 4 video-sequence per identity, giving a total of 1365 identities.

(b)The **gallery subset** is composed by the video-frame sequences used to create the initial template as well as the ones used to query the system (from both *genuine* and *impostor* identities).

- In the case of COX Face database, this set contains the 700 identities taken from `cam1` and `cam2`. Each video was divided into 5 sub-sequences to augment the number of possible queries, giving a total of 10 sub-sequences.
- In the case of YouTube Faces, this subset contains data from the identities that have equal or more than 4 video-sequences per identity, giving a total of 226 identities. It includes all but one video of each identity, which will create the following probe subset. The videos will be divided into a total of 10 sub-sequences without mixing different videos.

(c)The **probe subset** contains the video-frames sequences we draw to test the system in each step of the learning phase. The testing is performed after each query of the learning phase. This way, we can measure the evolution of the updating system.

- In the case of COX Face database, this set contains video sequences captured by `cam3` belonging to the 700 identities in the *gallery subset*. In this case, each video sequence was divided into 10 sub-sequences to have more sequences to test.
- In the case of YouTube Faces, this subset contains data from the identities that have equal or more than 4 video-sequences per identity, giving a total of 226 identities. It includes the remaining video after the creation of the previous gallery subset. As the other dataset, each sequence will be divided into 10 sub-sequences too.

It is important to remark that there are no common identities between the *train subset* and the other two subsets. Identities belonging to the *train subset* conform the Universal Model (UM) (la Torre et al., 2015; Li et al., 2005). On the other hand, *gallery* and *probe subsets* contain different sequences of shared identities. In the experiments, each of these identities will have associated its own Cohort Model (CM) (la Torre et al., 2015; Li et al., 2005). In practical terms, each identity CM will be conformed by itself and its 10 'most similar' (using SVM as metric (López-López et al., 2019)) impostors or hard-negatives. Consequently, this kind of testing is quite more demanding than regular random impostor testing.

4.4. SDR self-labelling: operational threshold

In each verification query, a short sequence of video frames is processed according to the SDR (see Section 3.2). The SDR gives a score to the video-sequence and decides based on a threshold, the so-called *operational threshold*. Its determination is crucial and especially tricky in an incremental learning context.

The strictness/gentleness on the *operational threshold* modulates the self-labelling process's confidence degree. Potentially, aspects as data quality, face characteristics, or the acquisition environment may affect its optimal determination (identity and time dependence). Nevertheless, we have opted to ignore these dependencies (both identity and time) when defining the determination procedure. It seems reasonable as a first step considering the enormous limitations in labelled data of the target context. Similar assumptions are commonly made in other works (Rattani, Marcialis, & Roli, 2013). A performs a further study of the implications of this assumption.

Thus, the *train subset* (which contains the identities of the UM) will be

Table 5
COX: Study on template size De-SVM at operational threshold 5%.

Template size	TAR@FAR1		TAR		FAR	
	Initial	Final	Initial	Final	Initial	Final
1	20.30±0.64	52.42±0.41	32.00±0.28	53.68±0.38	3.36±0.10	1.178±0.053
3	30.19±0.76	77.91±0.52	45.47±0.50	78.09±0.30	3.213±0.080	1.036±0.076
5	37.17±0.58	85.45±0.25	59.01±0.37	88.14±0.24	5.01±0.13	1.714±0.052
7	40.74±0.49	86.69±0.65	60.90±0.41	87.32±0.49	3.938±0.096	1.138±0.073
10	45.21±0.77	88.87±0.24	65.60±0.23	89.02±0.21	4.098±0.095	1.024±0.043

used as a validation set. Within this set, we will replicate the previous divisions (train, gallery and probe) to build and characterise (compute the ROC curve) sample models using what would be the initial template. As a convention, we have selected the threshold associated with 5% FAR point of the initial model as *operational threshold*. However, we test different operational thresholds (along with other templates sizes) on Section 5.2.2.

4.5. Metrics

The metrics used to evaluate are the measurement of TAR at a given 1% FAR (TAR@FAR1). We also provide the Transaction Level performance (TAR and FAR at the *operational threshold*).

Performance is assessed on the *probe subset*. After each query, the system es presented with 10 *genuine* sub-sequences and 1 *impostor* subsequence per *impostor* identity (10 in total). As it has been stated in Section 4.3, both querying and testing is performed using the CM of *genuine* identity. Finally, results are averaged over the total number of identities used as genuine in each verification process.

Finally, as mentioned in Section 4.3, the negative set used for training consists of a random subset of 1000 samples drawn from the *train subset*. This randomness adds uncertainty to the results and needs to be addressed. In order to deal whit it, experiments will be repeated 8 times and averaged.

5. Experiments and results

This section presents the experimental part of the paper. First, we establish some baselines (Section 5.1) to put into perspective the value of the results. From this point, we explore the ability to build a robust model with a minimum amount of labelling in an environment with both genuine and impostor queries (Section 5.2). Then, we test the robustness of the created model to repeated impostor *attacks* (Section 5.3). The complete results achieved in the experiments are recalled in Table 6, and the final discussion will be done in Section 5.4.

5.1. Baselines

As aforementioned (Section 1 and Section 2), the limited amount of previous literature framed in the specific of the conditions of our problem makes difficult to establish a direct comparison. In this regard, Malisiewicz et al. (2011) uses ensembles of very-specific SVM classifiers like De-SVM. Nevertheless, the fact that we are focusing on face related problems (in which their proposed calibration phase is problematic) prevent us from using their work as a fair and acceptable baseline. Thus, we opted for alternative ways of establishing comparison baselines, and they will be used to put our results into perspective for a proper analysis in Section 5.4.

5.1.1. State-of-the-art using available data

First of all, we need to consider that the newest deep feature representations are trained so as faces of the same identities are close to each other in the feature space (López-López et al., 2019). Thus, one can consider state-of-the-art performance the one obtained by training a classifier using just the available labelled data ($n = 5$ frames). This approach of using traditional classifiers (e.g. SVM) on top of deep learning feature representations is quite common in the literature and has proven to achieve remarkable performance (Crosswhite et al., 2018). The baseline is established using the ResNet-29 (RN29 + SVM) feature representation described in Section 4.2 and ResNet50-AF (RN50-AF + SVM) feature representation (Deng, Guo, Xue, & Zafeiriou, 2019) (which tops the state-of-the-art on LFW benchmark).

5.1.2. Supervised adaptation

Additionally, it would also be possible to perform the adaptation process done under supervised conditions. The performance obtained in this experiment represents an upper-bounds since they use entirely supervised labels (ideal and unrealistic case of perfect self-labelling). Then, the $n = 5$ available labelled frames are used as the initial template. This template was used to create the model M_0 and, after that, the system is queried (and the model consequently updated) with 10 different short video sequences (verification queries) from the genuine identity. Appendix B presents a comprehensive analysis of this scenario.

5.2. Unsupervised adaptation

The procedure in this experiment follows the philosophy of the supervised case (see Section 5.1.2). However, instead of having access to labels, the SDR must distinguish between genuine and impostor queries. Consequently, the first step was to generate M_0 from the initial template (5 frames). After that, the model was queried with 10 genuine sequences, G^s, and 10 different impostor sequences, I_k^s (where s stands for the sub-sequence number and k for the identity of the impostor). All of them belonging to the *gallery subset* (identities of the CM). After each

Table 6
Summary of TAR@FAR1% performances values obtained (values in %). Uncertainty is not represented in previous graphs for the sake of clarity. SU stands for self-updating.

COX				
Model	Initial	Superv. Adapt.	Unsuperv. Adapt.	Robustness
RN29 + SVM	37.19 ± 0.63	–	–	–
RN50-AF + SVM	**51.6 ± 1.0**	–	–	–
RN29 + SVM + SU	37.19 ± 0.63	88.89 ± 0.74	24.5 ± 1.0	10.04 ± 0.48
RN29 + I-SVM + SU	17.7 ± 3.6	79.8 ± 1.2	61.3 ± 1.5	5.51 ± 0.49
RN29 + OS-ELM + SU	10.27 ± 0.51	**92.35 ± 0.45**	75.0 ± 1.1	19.3 ± 2.7
RN29 + De-SVM (**Ours**)	37.17 ± 0.58	89.47 ± 0.24	**85.45 ± 0.25**	**64.4 ± 1.6**
YouTube Faces				
Model	Initial	Superv. Adapt.	Unsuperv. Adapt.	Robustness
RN29 + SVM	55.9 ± 1.3	–	–	–
RN50-AF + SVM	**81.33 ± 0.58**	–	–	–
RN29 + SVM + SU	55.9 ± 1.3	**88.68 ± 0.42**	34.7 ± 1.5	9.0 ± 1.1
RN29 + I-SVM + SU	42.0 ± 2.2	73.9 ± 2.2	65.5 ± 2.5	2.26 ± 0.67
RN29 + OS-ELM + SU	13.8 ± 1.8	76.8 ± 2.0	66.4 ± 1.9	32.2 ± 5.2
RN29 + De-SVM (**Ours**)	56.91 ± 0.59	76.26 ± 0.75	**75.5 ± 1.2**	**64.8 ± 1.1**

genuine query (odd query, $t = 1, 3, \ldots, 19$), an impostor query (even query, $t = 2, 4, 20$) was presented. The query order follows this pattern:

$$G^1 \rightarrow I_1^1 \rightarrow G^2 \rightarrow I_2^1 \rightarrow \ \ldots \ \rightarrow G^s \rightarrow I_k^1 \rightarrow \ \ldots \ \rightarrow G^{10} \rightarrow I_{10}^1$$

Performance measurements are done on the *probe set*, using samples of each identity's CM. As aforementioned, the CM consists of both the genuine identity and its 10 most similar impostors (see Section 4.3). Performance metrics (see Section 4.5) measurements are done after each query to the system.

Results (Figs. 4 and 5) for both COX and YTF, respectively, show the ability to improve the performance of the self-updating approach for every classification method apart from SVM. De-SVM is the one to achieve the best performance scoring at TAR@FAR1 of $86.03 \pm 0.48\%$ on COX and $75.5 \pm 1.2\%$ on YTF (Section 4.5 contains a detailed explanation about the uncertainty origin).

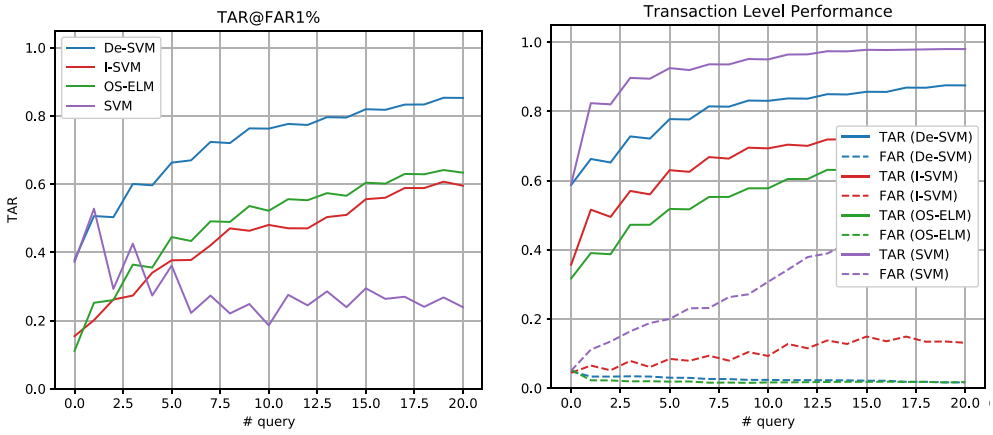
A comprehensive analysis of Figs. 4b and 5b allows to identify two different behaviours. On the one hand, De-SVM and OS-ELM can improve TAR while decreasing/maintaining FAR. Conversely, both SVM and I-SVM are unable to improve TAR without an unacceptable increase of FAR. To explain this behaviour, we need to recall a specific detail.

In Figs. 4b and 5b, initial and subsequent TAR measurements of I-SVM and SVM are quite high. This means that the model can incorporate much more genuine information (see Fig. 6). At the end of the experiment, we have a model that has acquired almost the same genuine information as the supervised case. Moreover, as counter-intuitive as it may seem, FAR seems to increase whenever a genuine sequence is presented, while FAR seems to decrease after an impostor query. This is especially noticeable when testing on COX. However, this behaviour is coherent with the one explored in the Appendix A. In that experiment, we show the importance of balance stability when using a constant threshold. Therefore, adding this amount of genuine information (high TAR) leans the classification problem to a more balanced one, making the initial threshold obsolete (FAR increases after genuine queries).

Finally, an essential detail to remark is the decreasing FAR observed for De-SVM and OS-ELM. This behaviour continues in the supervised scenario (Appendix B). While De-SVM presents a soft, monotonous decrease; OS-ELM presents a sharp decline at the beginning with a slight tendency change at the end.

5.2.1. *Relation of genuine/impostor trains*

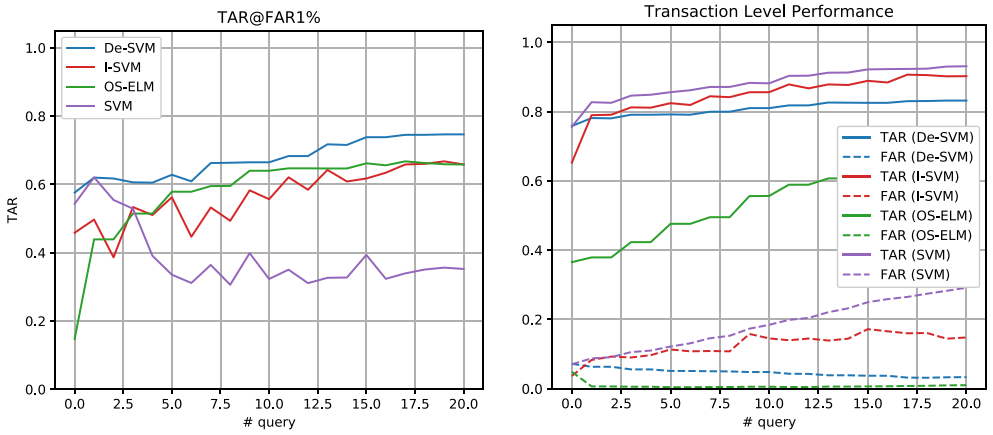
Paying attention to the relation of genuine and impostor update rate (over the total possible updates) we can extract important conclusions (Fig. 6). Firstly, I-SVM and SVM are quite more sensible to model corruption due to false acceptance updates. This is coherent with the behaviour observed using transaction-level performance. This corruption negatively affects in particular SVM. The behaviour is even more relevant if we remember that SVM is one of the best-performing methods



(a) TAR@FAR1 performance.

(b) TAR and FAR at the operational threshold.

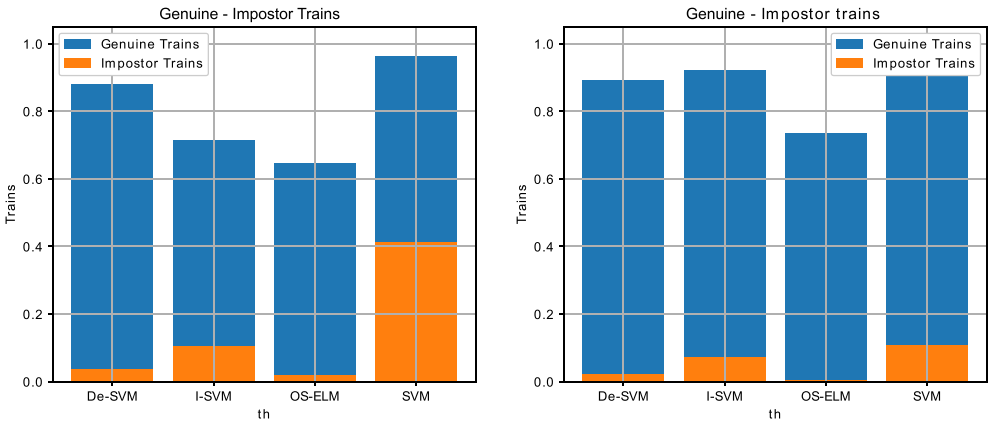
Fig. 4. YouTube Faces: Self-supervised experiment using an operational threshold of 5% initial FAR.



(a) TAR@FAR1 performance.

(b) TAR and FAR at the operational threshold.

Fig. 5. YouTube Faces: Self-supervised experiment using an operational threshold of 5% initial FAR.



(a) FACE COX database. (b) YouTube Faces database.

Fig. 6. Genuine and impostor updates performed by each classification technique in each database.

in the supervised experiments. A possible explanation for this behaviour is further studied on Appendix A.

Secondly, we can infer from the genuine/impostor updates that the most benefits are obtained from acquiring enough genuine information, even if it is at the cost of including some impostors. This justifies the assumption of setting the operational threshold to a 5% FAR.

5.2.2. *Effects on performance of different operational thresholds and template sizes*

In this section, we explore the effects of having different operational thresholds and templates sizes on De-SVM. To do that, we repeat the previous experiment on COX database varying these parameters. Results can be seen on Tables 4 and 5.

Analysing the operational threshold dependence (Table 4), we observe TAR@FAR1 increases as the operational threshold becomes less strict. However, the selected 5% FAR is an inflexion point from which the gain is more subtle. Looking at TAR and FAR performance, we observe that the increase of final TAR is done at FAR expenses. In this regard, we can assume then the 5% operational threshold as an acceptable compromise.

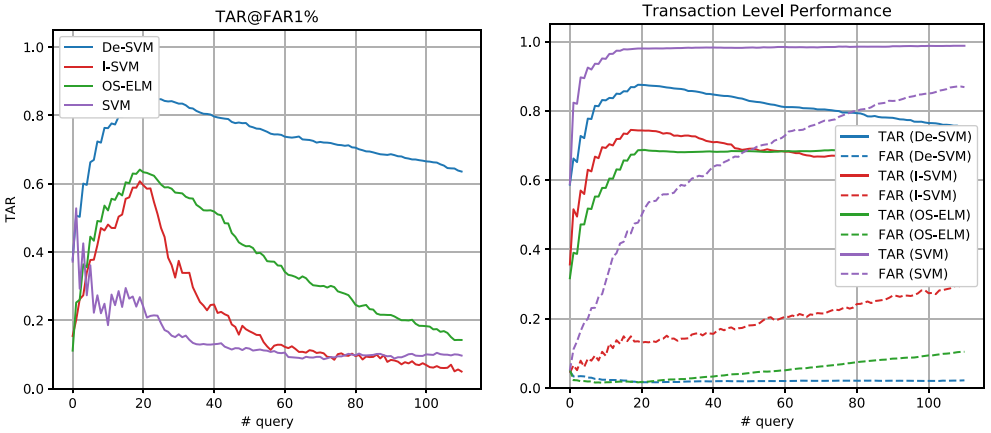
Turning now to the template size effect (Table 5), we observe that most of the performance improvement appears initially. Again, on the final performance, there is an inflexion point at the selected size of 5 frames. Since we want to address the problem of data scarceness, we want to extract the maximum power from the minimum amount of labelled information. Therefore, based on this behaviour and the one observed in Lopez-Lopez et al. (2019), the election of a 5 frames template seems quite reasonable.

5.3. *Post-robustness impostor testing*

In static classification scenarios, regular FAR measures robustness against impostors. However, the systems studied here are non-static making FAR non-static as well. Even more considering that the system's predictions are used as pseudo-labels during the updates, making false acceptances susceptible to feedback. This experiment intends to test each classification technique in the extreme scenario where the system is repeatedly queried with only impostor queries.

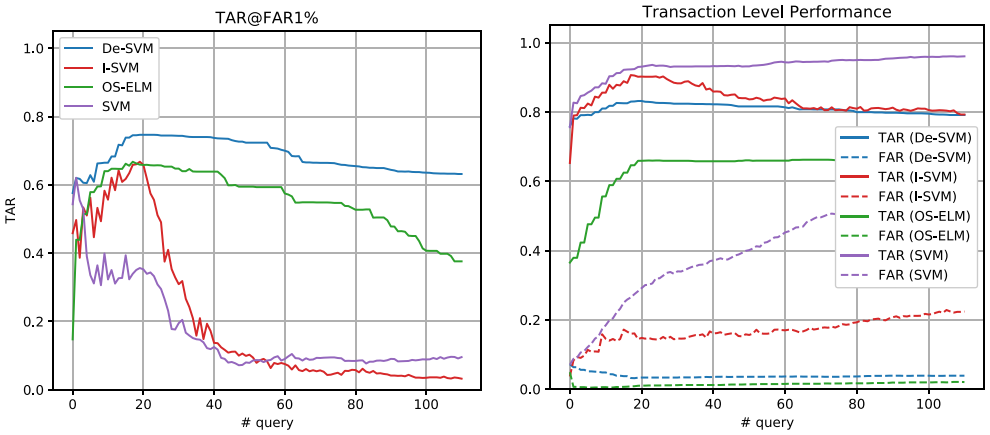
The idea is to part of the previous experiment. The initial template of 5 frames is maintained as well as the CM with the genuine identity and the 10 most similar impostors. So, after the 20^{th} query we will present a total of 90 impostor queries (every identity from the CM of the genuine target) with additional sub-sequences of the *gallery subset*. The fact that the CM is maintained (with the 10 most similar impostors) particularly augment the difficulty of the robustness testing. The pattern followed was:

$$\textit{Section}\,3.3\begin{cases}\textit{Section}\,5.2 & \{G^1\rightarrow I_1^1\rightarrow \ldots \rightarrow G^{10}\rightarrow I_{10}^1\rightarrow \\ & \{I_1^2\rightarrow \ldots \rightarrow I_{10}^2\rightarrow \quad \ldots \rightarrow I_1^{10}\rightarrow \ldots \rightarrow I_{10}^{10}\end{cases}$$



(a) TAR@FAR1 performance. (b) TAR and FAR at the operational threshold.

Fig. 7. FACE COX: Self-supervised performance comparison fixing the operational threshold at 5% initial FAR, testing robustness.



(a) TAR@FAR1 performance.

(b) TAR and FAR at the operational threshold.

Fig. 8. YouTube Faces: Self-supervised performance comparison fixing the operational threshold at 5% initial FAR, testing robustness.

The procedure to measure performance is the same as the previous experiment (using the *probe set* with identities of the CM), using the metrics described on Section 4.5. Measurements are performed after each of the queries. Results obtained can be seen in Figs. 7 and 8. Overall, the first thing we observe is that every method suffers a performance loss in this testing. Nevertheless, looking at Fig. 7a, we realise De-SVM's outstanding resistance compared to the other methods. TAR@FAR1 moves from $86.03 \pm 0.48\%$ and $75.5 \pm 1.2\%$ (COX and YTF, respectively) in query 20 to just below $64.4 \pm 1.6\%$ and $64.8 \pm 1.1\%$ in query 110. The following best performing technique is OS-ELM that goes from $75.0 \pm 1.1\%$ and $66.4 \pm 1.9\%$ to a final performance of $19.3 \pm 2.7\%$ and $32.2 \pm 5.2\%$ TAR@FAR1.

Looking again to transaction level performance (using the *operational threshold*) we find interesting behaviours again. Intuitively, one could think that a repeated impostor querying would make FAR out of control. Indeed, this behaviour is observed in 3 of the 4 classification techniques that have been tested (SVM and I-SVM). A softer behaviour is observed for OS-ELM. FAR values are maintained low, but the increasing trend (especially for COX database) is still observable. This is something which makes sense, given the fact that the only possible mistake is to accept an impostor query as genuine.

However, De-SVM presents just the opposite behaviour. Most damage comes from a decreasing TAR, instead of an increasing FAR. To understand this effect, we need to go deeper into ensembles' nature. Ensembles decisions are based on majorities. The majority of accepting genuine identity is built during the initial stages (query 0 to 20). Based on Fig. 6, this decision is supported by 9 out of 10 classifiers. It would be necessary to overcome this majority of 9 classifiers to confuse an impostor with a genuine persistently. This is something quite difficult given the fact that FAR is always below 5%.

In other words, impostor classifiers may agree on rejecting genuine identities (TAR decrease), but they cannot agree on the decision to accept another identity as genuine (FAR stability). This behaviour is quite interesting in fields like biometric identification, in which the main concern is to avoid impostors entering the system.

5.4. Summary and discussion

Finally, Table 6 represents a summary of the experiments. In this table, we present state-of-the-art with the available data and supervised adaptation baselines along with the rest of the experiments. The first set of baselines (Section 5.1.1), in the first two rows of the results for each dataset, shows the performance of two powerful learned feature representations used as feature extractors to train an SVM classifier using the available labelled data. The second set of baselines (Section 5.1.2), Superv. Adapt. column, shows the potential performance of each method in the unrealistic case of having perfect self-labelling. Both sets serve to put into perspective the results achieved in the rest of the experiments.

At first glance, it is noticeable that the self-updating method is an interesting approach to perform an unsupervised adaptation. It can improve initial performance in 3 of the 4 tested methods (including ours). This improvement is enough to overcome the state-of-the-art performance with the available data (Section 5.1.1) when using COX database. This is something quite remarkable given the short number of gallery samples (5 low-quality video frames). Nevertheless, in the case of YTF, AF + RN50 (Deng et al., 2019) stills beats the methods that include self-updating. A possible explanation for this is two-folded. First, unlike RN29, this feature representation is designed with YTF test in mind (Deng et al., 2019) acquiring state-of-the-art performance in this dataset too. Secondly, YTF is not database designed for the specific problem of video-surveillance. Consequently, most of the specific built-in characteristics (e.g. variable scale and light conditions, blurriness, etc.) are less present in video-frames (Fig. 2). This fact makes easier the transition between stills and videos.

Comparing to supervised adaptation (Section 5.1.2), every model experiences a predictable drop in performance when updating phase is done under unsupervised conditions. In this regard, De-SVM is the method that experiences the smallest of all. Thus, De-SVM is able to achieve comparable performance using less than a tenth of labels. On the other hand, SVM is the method that experiences the highest drop in this comparison.

Finally, in terms of robustness, De-SVM presents impressive characteristics in comparison to other classification methods. Its behaviour is even more remarkable, given that the only labelling used is the one to create the initial template. Besides, the fact that the performance damage is caused by lowering TAR and not increasing FAR represents a desirable and promising quality for any biometric application.

6. Conclusions

In this work, the problem of V2V-FV for face re-identification purposes, without a previous collaborative manual enrolment, is tackled. Based on state-of-the-art CNN features, self-updating can incorporate pseudo-labelled samples to perform incremental learning during the operational phase. Based on the obtained results, the self-updating approach arises as a promising strategy to take advantage of domain-specific samples incrementally retrieved.

Our De-SVM can incorporate new relevant information while maintaining a low false-positive rate in a completely unsupervised way. And

not only this, the behaviour becomes even more promising given that we are not using the full power of an ensemble-based approach yet. As the process of updating consist of adding a new classifier (*learn*), one could potentially be able to correct wrong updates in the future by removing this classifier (*forget*). The ability to *forget* is only possible with this kind of classification techniques.

As future work, we are considering to endow the method with the ability to forget, especially for life-long learning purposes. In these setups, incorporating a method to set an adaptive (variable) decision threshold could significantly impact the results. Besides, we need to consider that face verification only plays the role of a validation application. De-SVM could be translated to other video-related contexts as object detection from mobile robots, person Re-ID or other detection applications. Overall, the proposed method's benefits extend to any detection application in which good enough models cannot be generated offline. Thus, unsupervised adaptation could be a desirable capability.

CRediT authorship contribution statement

Eric Lopez-Lopez: Conceptualization, Methodology, Software, Validation, Formal analysis, Investigation, Writing - original draft, Visualization. **Carlos V. Regueiro:** Conceptualization, Methodology, Validation, Writing - original draft, Supervision. **Xosé M. Pardo:** Conceptualization, Methodology, Validation, Formal analysis, Resources, Writing - original draft, Supervision. **Annalisa Franco:** Conceptualization, Methodology, Writing - review & editing. **Alessandra Lumini:** Conceptualization, Methodology, Writing - review & editing.

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 TIN2017-90135-R MINECO (FEDER)), from The Consellaría de Cultura, Educación e Ordenación Universitaria (accreditations 2016–2019, EDG431G/01 and ED431G/08), and reference competitive groups (2017–2020, and ED431C 2017/04), 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).

Appendix A. Time independence of the operational threshold

The *operational threshold* is assumed to be neither time nor identity dependent (Section 4.4). Despite not being totally accurate, this assumption is forced by the label's scarceness of the operation's context. In fact, Fig. A.9 shows how, using a constant *operational threshold*, FAR increases after each update for a supervised batch SVM (described in Section 3.4.1) showing the assumption's inaccuracy. In this section, the aim is to explore the actual implications of the independence assumption (specifically regarding time).

The experiment conducted shows the temporal evolution of this threshold for five different points of the ROC curve (Fig. A.10) under supervised conditions (same experiment as in Section 5.1.2). According to the previous assumption, the ideal behaviour would correspond to steady (or, at least, decreasing) curves. Such behaviour would mean that the system can improve (to increase TAR) without increasing the probability of accepting impostors, potentially corrupting the model.

Results show different behaviours depending on the classification method used. Overall it is observed the explicit break of the threshold's time independence assumption. De-SVM is the method that presents the most steady behaviour among all the classification techniques during this experiment. For instance, for SVM, thresholds corresponding to 1% FAR (at the beginning) and 25% FAR (at the end) are the same. OS-ELM shows the opposite behaviour. The initial threshold associated with a FAR 5% decreases over time.

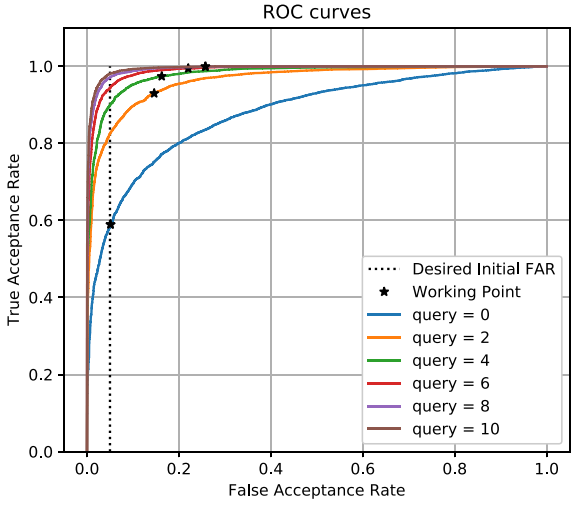


Fig. A.9. Evolution of the ROC curve and the ROC point associated to the *operational threshold* after each query for the supervised case using the SVM classification model.

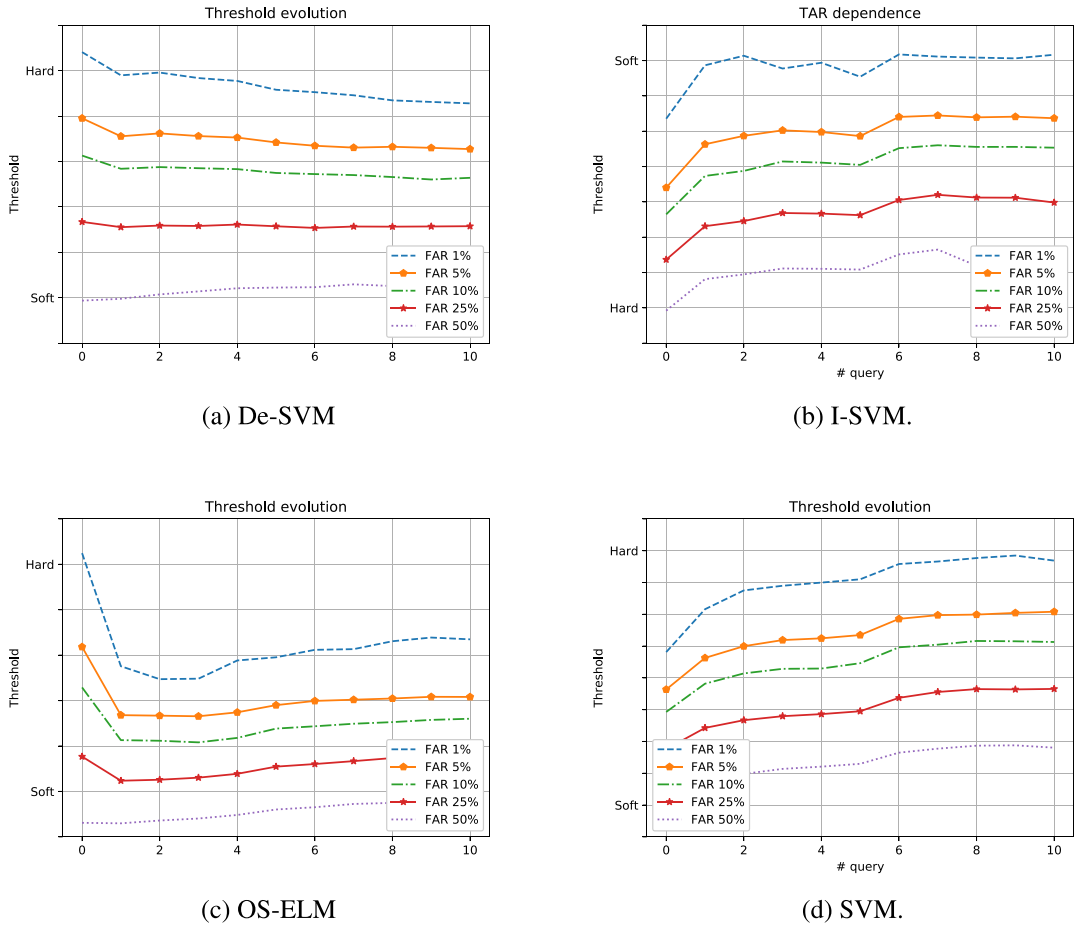


Fig. A.10. COX: Threshold evolution of a same FAR point of the ROC curve (Data for 1%, 5%, 10%, 25% and 50%).

This behaviour can be explained in terms of the sample balance (positive/negative) during the classifiers' creation. At the early stages, this balance is compromised. The limited number of positive samples contrasts with a large number of negative ones. The balance is recovered when the model adds more genuine queries. Subsequently, the *operational threshold* selected (at 5% FAR) using the initial model (unbalanced problem) leads to a decision boundary shifted towards the positive sample(s). This boundary does not correspond to the same ROC point of the final model (balanced problem).

In contrast, De-SVM uses an ensemble of 'unbalanced' SVM. However, each classifier is 'unbalanced' to the same degree, making the decision boundary much more stable during the queries. Besides, De-SVM presents a subtle decrease over time, indicating that the decision is becoming more strict. OS-ELM even show a more substantial decrease in the initial phases.

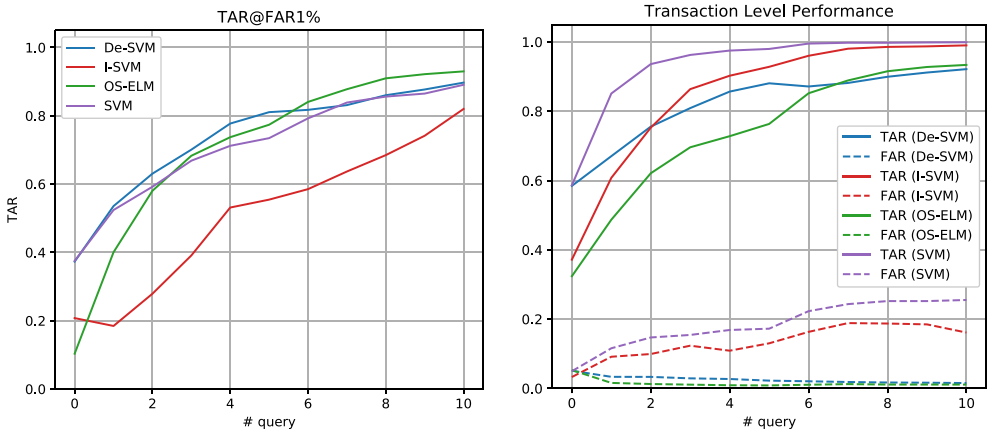
These experiments showed that, despite being inaccurate, the assumption is acceptable. Again, one can start to foreshadow a better performance of both De-SVM and OS-ELM in following experiments.

Appendix B. Detailed supervised results

This section showcases detailed performance behaviour under supervised conditions. Performance is measured on the *probe set*, using the samples of the CM of each identity. Measurements were done after the creation of the initial model ($t = 0$) and after each query ($t = 1, 2, \ldots, 10$). Results showcased in this experiment are directly comparable with the ones on Section 5.2, when the updates are done in the absence of labels.

Results (Figs. B.11 and B.12) show that every method can achieve a remarkable performance, especially when testing on COX Face database. In this sense, I-SVM is the method that experiences the hardest time during the experiment. Above all, results show the capability of building an acceptable model in conditions where there is plenty of available data. Differences in performance are more obvious at the *operational threshold*, as illustrated in Figs. B.11b and B.12b. De-SVM achieves the most modest performance in terms of TAR respect to the other methods. Nevertheless, it is important to remark that higher performances are obtained at the cost of higher (and increasing) values of FAR. Both OS-ELM and De-SVM present a decreasing FAR, which ends up below 5%. These curves suggest a more desirable behaviour when updates will be performed in an unsupervised manner. A high FAR value means that the probability of accepting impostors during the training could also be high, and so the high risk of corrupting the model.

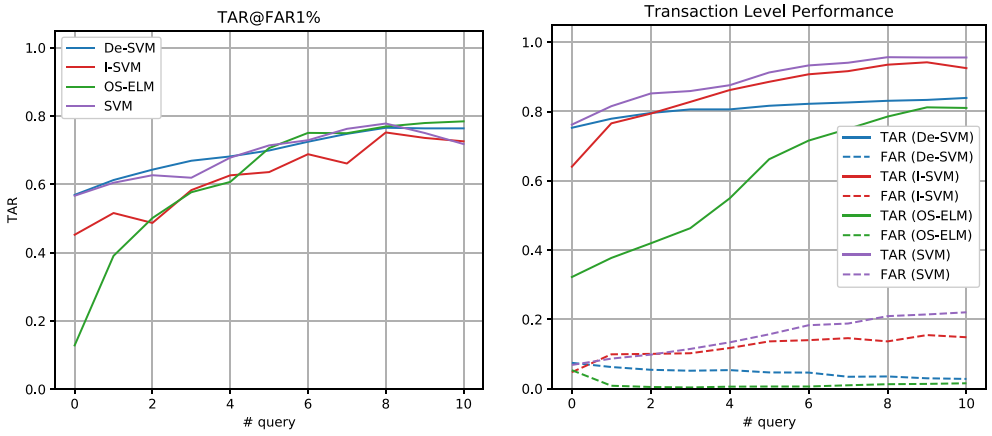
Finally, these last figures give an important clue about the relevance of the *operational threshold* and its crucial influence on the self-updating mechanism. This influence was comprehensively studied in the previous appendix (Appendix A).



(a) TAR@FAR1 performance.

(b) TAR and FAR at the operational threshold.

Fig. B.11. COX: Supervised updating performance comparison.



(a) TAR@FAR1 performance.

(b) TAR and FAR at the operational threshold.

Fig. B.12. YouTube Faces: Supervised updating performance comparison.

Appendix C. Using other central tendency measures in the FDR and SDR

In this section, we repeat the experiment performed on Section 5.2 using the mean as an alternative central tendency measure in the FDR and SDR. Results on Table 7 show that the differences in performance are insignificant. Thus, we have selected the median due to a mere theoretical convention. Conceptually, the median corresponds to performing a majority vote after obtaining a binary classifier response (using the operational threshold).

Table 7
COX: Comparison on the function used in the FDR and SDR.

	TAR@FAR1		TAR		FAR	
Function	Initial	Final	Initial	Final	Initial	Final
Median	37.17±0.58	85.45±0.25	59.01±0.37	88.14±0.24	5.01±0.13	1.714±0.052
Mean	36.56±0.56	84.27±0.48	59.00±0.26	87.39±0.32	5.01±0.12	1.681±0.040

References

Bashbaghi, S., Granger, E., Sabourin, R., & Bilodeau, G.-A. (2017). Dynamic ensembles of exemplar-SVMs for still-to-video face recognition. *Pattern Recognition, 69*, 61–81. https://doi.org/10.1016/j.patcog.2017.04.014

Bashbaghi, S., Granger, E., Sabourin, R., & Parchami, M. (2019). Deep learning architectures for face recognition in video surveillance. In X. Jiang, A. Hadid, Y. Pang, E. Granger, & X. Feng (Eds.), *Deep learning in object detection and recognition* (pp. 133–154). Singapore: Springer Singapore. https://doi.org/10.1007/978-981-10-5152-4_6.

Becattini, F., Seidenari, L., & Del Bimbo, A. (2017). Indexing quantized ensembles of exemplar-SVMs with rejecting taxonomies. *Multimedia Tools and Applications, 76*, 22647–22668. https://doi.org/10.1007/s11042-017-4794-7

Becker, S. (1999). Implicit learning in 3d object recognition: The importance of temporal context. *Neural Computation, 11*, 347–374. https://doi.org/10.1162/089976699300016683

Bianco, S. (2017). Large age-gap face verification by feature injection in deep networks. *Pattern Recognition Letters, 90*, 36–42. https://doi.org/10.1016/j.patrec.2017.03.006

Chefrour, A. (2019). Incremental supervised learning: Algorithms and applications in pattern recognition. *Evolutionary Intelligence, 12*, 97–112. https://doi.org/10.1007/s12065-019-00203-y

Chen, X., Wang, C., Xiao, B., & Cai, X. (2014). Scenario oriented discriminant analysis for still-to-video face recognition. In *IEEE international conference on image processing (ICIP)* (pp. 738–742). https://doi.org/10.1109/ICIP.2014.7025148

Cortes, C., & Vapnik, V. (1995). Support-vector networks. *Machine Learning, 20*, 273–297. https://doi.org/10.1007/BF00994018

Crosswhite, N., Byrne, J., Stauffer, C., Parkhi, O., Cao, Q., & Zisserman, A. (2018). Template adaptation for face verification and identification. *Image and Vision Computing, 79*, 35–48. https://doi.org/10.1016/j.imavis.2018.09.002

Deng, J., Guo, J., Xue, N., & Zafeiriou, S. (2019). Arcface: Additive angular margin loss for deep face recognition. In *2019 IEEE/CVF conference on computer vision and pattern recognition (CVPR)* (pp. 4685–4694). https://doi.org/10.1109/CVPR.2019.00482

Dewan, M. A. A., Granger, E., Marcialis, G.-L., Sabourin, R., & Roli, F. (2016). Adaptive appearance model tracking for still-to-video face recognition. *Pattern Recognition, 49*, 129–151. https://doi.org/10.1016/j.patcog.2015.08.002

Dhamecha, T. I., Noore, A., Singh, R., & Vatsa, M. (2019). Between-subclass piece-wise linear solutions in large scale kernel svm learning. *Pattern Recognition, 95*, 173–190. https://doi.org/10.1016/j.patcog.2019.04.012

Didaci, L., Marcialis, G. L., & Roli, F. (2014). Analysis of unsupervised template update in biometric recognition systems. *Pattern Recognition Letters, 37*, 151–160. https://doi.org/10.1016/j.patrec.2013.05.021

Ditzler, G., Roveri, M., Alippi, C., & Polikar, R. (2015). Learning in nonstationary environments: A survey. *IEEE Computational Intelligence Magazine, 10*, 12–25. https://doi.org/10.1109/MCI.2015.2471196

Fahad, A., Almalawi, A., Tari, Z., Alharthi, K., Qahtani, F. S. A., & Cheriet, M. (2019). Semtra: A semi-supervised approach to traffic flow labeling with minimal human effort. *Pattern Recognition, 91*, 1–12. https://doi.org/10.1016/j.patcog.2019.02.001

Li, Fayin, & Wechsler, H. (2005). Open set face recognition using transduction. *IEEE Transactions on Pattern Analysis and Machine Intelligence, 27*, 1686–1697. https://doi.org/10.1109/TPAMI.2005.224

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

Gepperth, A. & Hammer, B. (2016). Incremental learning algorithms and applications. In European symposium on artificial neural networks (ESANN) (pp. 27–29). Bruges, Belgium. https://doi.org/hal-01418129.

Grossberg, S. (1988). Nonlinear neural networks: Principles, mechanisms, and architectures. *Neural Networks, 1*, 17–61. https://doi.org/10.1016/0893-6080(88)90021-4

He, K., Zhang, X., Ren, S., & Sun, J. (2016). Deep residual learning for image recognition. In *IEEE conference on computer vision and pattern recognition (CVPR)* (pp. 770–778). https://doi.org/10.1109/CVPR.2016.90

Hoens, T. R., Polikar, R., & Chawla, N. V. (2012). Learning from streaming data with concept drift and imbalance: An overview. *Progress in Artificial Intelligence, 1*, 89–101. https://doi.org/10.1007/s13748-011-0008-0

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

Kemker, R., McClure, M., Abitino, A., Hayes, T. L., & Kanan, C. (2018). Measuring catastrophic forgetting in neural networks. In *AAAI conference on artificial intelligence* (pp. 3390–3398).

Kim, S., Choi, J., Kim, T., & Kim, C. (2019). Self-training and adversarial background regularization for unsupervised domain adaptive one-stage object detection. In *The IEEE international conference on computer vision (ICCV)*.

King, D. E. (2009). Dlib-ml: A machine learning toolkit. *Journal of Machine Learning Research, 10*, 1755–1758. https://doi.org/10.1145/1577069.1755843

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

Kivinen, J., Smola, A. J., & Williamson, R. C. (2004). Online learning with kernels. *IEEE Transactions on Signal Processing, 52*, 2165–2176. https://doi.org/10.1109/TSP.2004.830991

Krawczyk, B., Minku, L. L., Gama, J., Stefanowski, J., & Woźniak, M. (2017). Ensemble learning for data stream analysis: A survey. *Information Fusion, 37*, 132–156. https://doi.org/10.1016/j.inffus.2017.02.004

Li, Y., Yang, F., Liu, Y., Yeh, Y., Du, X., & Wang, Y. F. (2018). Adaptation and re-identification network: An unsupervised deep transfer learning approach to person re-identification. In *2018 IEEE/CVF conference on computer vision and pattern recognition workshops (CVPRW)* (pp. 285–2856). https://doi.org/10.1109/CVPRW.2018.00054

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

Lomonaco, V., & Maltoni, D. (2016). Comparing incremental learning strategies for convolutional neural networks. In *Artificial neural networks in pattern recognition (ANNPR)* (pp. 175–184). Springer International Publishing. https://doi.org/10.1007/978-3-319-46182-3_15.

Lopez-Lopez, E., Regueiro, C. V., Pardo, X. M., Franco, A., & Lumini, A. (2019). Incremental learning techniques within a self-updating approach for face verification in video-surveillance. In A. Morales, J. Fierrez, J. S. Sánchez, & B. Ribeiro (Eds.), *Pattern recognition and image analysis* (pp. 25–37). Cham: Springer International Publishing. https://doi.org/10.1007/978-3-030-31321-0_3.

López-López, E., Pardo, X. M., Regueiro, C. V., Iglesias, R., & Casado, F. E. (2019). Dataset bias exposed in face verification. *IET Biometrics, 8*, 249–258. https://doi.org/10.1049/iet-bmt.2018.5224

Malisiewicz, T., Gupta, A., & Efros, A. A. (2011). Ensemble of exemplar-SVMs for object detection and beyond. In *IEEE international conference on computer vision (ICCV)* (pp. 89–96). https://doi.org/10.1109/ICCV.2011.6126229

Masi, I., Wu, Y., Hassner, T., Natarajan, P., & del Rey, M. (2018). Deep face recognition: A survey. In *SIBGRAPI conference on graphics, patterns and images* (pp. 471–478). https://doi.org/10.1109/SIBGRAPI.2018.00067

Mian, A. (2011). Online learning from local features for video-based face recognition. *Pattern Recognition, 44*, 1068–1075. https://doi.org/10.1016/j.patcog.2010.12.001

Misra, I., Zitnick, C. L., & Hebert, M. (2016). Shuffle and learn: Unsupervised learning using temporal order verification. In *European conference on computer vision (ECCV)* (pp. 527–544). https://doi.org/10.1007/978-3-319-46448-0_32

Ng, H., & Winkler, S. (2014). A data-driven approach to cleaning large face datasets. In *2014 IEEE international conference on image processing (ICIP)* (pp. 343–347). https://doi.org/10.1109/ICIP.2014.7025068

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

Parkhi, O. M., Vedaldi, A., & Zisserman, A. (2015). Deep face recognition. In X. Xie, M. W. Jones, & G. K. L. Tam (Eds.), *Proceedings of the British machine vision conference (BMVC)* (pp. 41.1–41.12). BMVA Press. https://doi.org/10.5244/C.29.41.

Pernici, F., Bartoli, F., Bruni, M., & Del Bimbo, A. (2018). Memory based online learning of deep representations from video streams. In *IEEE conference on computer vision and pattern recognition (CVPR)* (pp. 2324–2334). https://doi.org/10.1109/CVPR.2018.00247

Pernici, F., & Bimbo, A. D. (2017). Unsupervised incremental learning of deep descriptors from video streams. In *IEEE international conference on multimedia expo workshops (ICMEW)* (pp. 477–482). https://doi.org/10.1109/ICMEW.2017.8026276

Pisani, P. H., Mhenni, A., Giot, R., Cherrier, E., Poh, N., Ferreira de Carvalho, A. C. P. d. L., Rosenberger, C., & Amara, N. E. B. (2019). Adaptive biometric systems: Review and perspectives. *ACM Computing Surveys, 52*, 102:1–102:38. https://doi.org/10.1145/3344255

Radosavovic, I., Dollár, P., Girshick, R., Gkioxari, G., & He, K. (2018). Data distillation: Towards omni-supervised learning. In *IEEE conference on computer vision and pattern recognition (CVPR)* (pp. 4119–4128). https://doi.org/10.1109/CVPR.2018.00433

Rattani, A., Marcialis, G. L., & Roli, F. (2013). Biometric system adaptation by self-update and graph-based techniques. *Journal of Visual Languages & Computing, 24*, 1–9. https://doi.org/10.1016/j.jvlc.2012.10.004

Redondo-Cabrera, C., & Lopez-Sastre, R. (2019). Unsupervised learning from videos using temporal coherency deep networks. *Computer Vision and Image Understanding, 179*, 79–89. https://doi.org/10.1016/j.cviu.2018.08.003

Roychowdhury, A., Chakrabarty, P., Singh, A., Jin, S., Jiang, H., Cao, L., & Learned-Miller, E. (2019). Automatic adaptation of object detectors to new domains using self-training. In *2019 IEEE/CVF conference on computer vision and pattern recognition (CVPR)* (pp. 780–790). https://doi.org/10.1109/CVPR.2019.00087

Sohn, K., Liu, S., Zhong, G., Yu, X., Yang, M., & Chandraker, M. (2017). Unsupervised domain adaptation for face recognition in unlabeled videos. In *IEEE international conference on computer vision (ICCV)* (pp. 5917–5925). https://doi.org/10.1109/ICCV.2017.630

Tommasi, T., Patricia, N., Caputo, B., & Tuytelaars, T. (2017). A deeper look at dataset bias. In G. Csurka (Ed.), *Domain adaptation in computer vision applications* (pp. 37–55). Cham: Springer International Publishing. https://doi.org/10.1007/978-3-319-58347-1_2.

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

Villamizar, M., Sanfeliu, A., & Moreno-Noguer, F. (2019). Online learning and detection of faces with low human supervision. *The Visual Computer, 35*, 349–370. https://doi.org/10.1007/s00371-018-01617-y

Wang, M., & Deng, W. (2018). Deep visual domain adaptation: A survey. *Neurocomputing, 312*, 135–153. https://doi.org/10.1016/j.neucom.2018.05.083

Wang, R., Shan, S., Chen, X., & Gao, W. (2008). Manifold-manifold distance with application to face recognition based on image set. In *IEEE conference on computer*

vision and pattern recognition (CVPR) (pp. 1–8). https://doi.org/10.1109/CVPR.2008.4587719

Wang, X., & Gupta, A. (2015). Unsupervised learning of visual representations using videos. In *IEEE international conference on computer vision (ICCV)* (pp. 2794–2802). https://doi.org/10.1109/ICCV.2015.320

Wolf, L., Hassner, T., & Maoz, I. (2011). Face recognition in unconstrained videos with matched background similarity. In *IEEE conference on computer vision and pattern recognition (CVPR)* (pp. 529–534). https://doi.org/10.1109/CVPR.2011.5995566

Wu, X., Zuo, W., Lin, L., Jia, W., & Zhang, D. (2018). F-svm: Combination of feature transformation and svm learning via convex relaxation. *IEEE Transactions on Neural Networks and Learning Systems, 29*, 5185–5199. https://doi.org/10.1109/TNNLS.2018.2791507

Yarowsky, D. (1995). Unsupervised word sense disambiguation rivaling supervised methods (pp. 189–196). doi: 10.3115/981658.981684.

Zhang, M., Liu, R., Nada, H., Uchida, H., Matsunami, T., & Abe, N. (2019). A pairwise learning strategy for video-based face recognition. In *The IEEE conference on computer vision and pattern recognition (CVPR) workshops*.

Zhang, X., Cao, J., Shen, C., & You, M. (2019). Self-training with progressive augmentation for unsupervised cross-domain person re-identification. In *The IEEE international conference on computer vision (ICCV)*.

Zhu, X. (2005). Semi-supervised learning literature survey. *Technical Report 1530 University of Wisconsin-Madison, Dept. Computer Science*.

Zou, Y., Yu, Z., Liu, X., Kumar, B. V., & Wang, J. (2019). Confidence regularized self-training. In *The IEEE international conference on computer vision (ICCV)*.