Self-supervised On-line Cumulative Learning from Video Streams

Federico Pernici[a], Matteo Bruni[a], Alberto Del Bimbo[a]

[a] *MICC, Media Integration and Communication Center, Dept. of information Engineering, University of Firenze, Italy.*

Abstract

We present a novel online self-supervised method for face identity learning from video streams. The method exploits deep face feature descriptors together with a memory based learning mechanism that takes advantage of the temporal coherence of visual data. Specifically, we introduce a discriminative descriptor matching solution based on Reverse Nearest Neighbour and a memory based cumulative learning strategy that discards redundant descriptors while time progresses. This allows building a comprehensive and cumulative representation of all the past visual information observed so far. It is shown that the proposed learning procedure is asymptotically stable and can be effectively used in relevant applications like multiple face identification and tracking from unconstrained video streams.

Experimental results show that the proposed method achieves comparable results in the task of multiple face tracking and better performance in face identification with offline approaches exploiting future information.

Keywords: Incremental Learning Cumulative Learning Multiple Object Tracking Face Recognition Lifelong Learning Long Term Object Tracking

1. Introduction

Supervised machine learning is a very successful learning paradigm in which a clear distinction is made between the training phase and the testing phase. Once a model is learned, it is no longer subjected to training and inference on novel unseen data tacitly assumes that the data distribution does not change over time. Once the learning phase is concluded no classes other than those used for learning can be predicted. Although, such hard division between training and testing and the availability of large corpus of annotated data have demonstrated exceptional achievements in learning the appearance of objects from images [44], they remain critical as linear improvements in performance require an exponential number of labeled examples [85]. In addition to this, efforts to collect large quantities of annotated images, such as ImageNet [21] and Microsoft COCO [51] don't have the necessary scalability and are hard to be extended, replicated or improved. These issues may also put a performance limit on models learned in this way.

Drawing inspiration from biological systems, a possible attractive alternative would be incrementally to learn the object appearance from never-ending video streams with no supervision, both exploiting the large quantity of unconstrained videos available in the Internet and the fact that adjacent video frames contain semantically similar information. This not only provides a variety of different viewing conditions in which objects can be observed but it also overcomes the restrictive barrier between the training and testing phase being each frame used for both training and testing. Specifically, each frame on a video stream can be used for learning, the following for testing and so on. Accordingly, never-ending tracking multiple subjects in the video could, at least in principle, support a sort of *self-supervised* incremental learning of their appearance. This would avoid or reduce the cost of annotation as time itself would provide a form of self-supervision which does not stop learning, but rather updates the learning model over time by accumulating knowledge without forgetting the past, reaching increasingly better accuracy and better data diversity as time advances.

However, this solution is not without problems. It is practically not possible to store all the data seen so far and re-learn a Deep Neural Network model periodically. Removing past data to adequately incorporate the new information without catastrophic forgetting, (i.e. performing Continual Learning [65]), is still an open challenge [50, 78, 73, 75], especially when new knowledge

Email addresses: federico.pernici@unifi.it (Federico Pernici), matteo.bruni@unifi.it (Matteo Bruni), alberto.delbimbo@unifi.it (Alberto Del Bimbo)

has to be incorporated in real time while tracking, without the availability of labels and with data coming from a stream which is often non-stationary [88, 2].

Single Object Tracking (SOT) [43] and Multiple Object Tracking (MOT) [46, 57] are closely related to the problem of learning from video streams but they have substantial differences and divergent goals from incremental and cumulative learning. While in SOT the object appearance is learned only for detecting the object in the next frame (the past information is gradually *forgotten* [35, 19]), cumulative learning from a video stream would require that *all* the past visual information of the object observed so far is collected in a *comprehensive* and *cumulative* representation. This not only requires tracking to be robust in the presence of very long term occlusions due intermittent (re)appearance of objects or other severe appearance changes, but that incremental learning is asymptotically stable so that it converges to an univocal cumulative representation. Moreover, modern SOT approaches based on Deep Learning are pre-trained on large video datasets [8, 86, 29, 48] and typically do not perform any learning at runtime or perform conservative updates [64]. Their extension to handle cumulative learning remains not trivial and however prone to catastrophic forgetting. *Long-term* SOT methods introduced in [37, 68] implement explicit target re-detection to reacquire the object after long term occlusion, despite of their successful performance on complex extended video sequences, their strategy for learning the appearance model does not substantially differ from those in SOT [42, 56].

Although MOT appears similar to the problem of learning the appearance of objects from video streams, major differences can be identified according to the following four criteria:

(1) Motion Continuity and Data Association. Most methods formulate MOT as a data association problem integrating several cues such as appearance, position, motion, and size into an affinity model to link track fragments (i.e. tracklets) into final trajectories. To be usefully exploited, this formulation implicitly requires that the objects are continuously detected and the camera is stationary, slowly moving or undergoing short-term rapid motions [98, 52, 20]. The motion continuity problem can be partially mitigated by the introduction of learned appearance model (i.e. features) trained on large corpus of data to perform short-term re-identification [97, 77, 7]. Instead in long-term re-identification after long occlusions, the continuity of motion is *no longer relevant* to the problem of data association [100]. When an object exits the field of view and re-enters after an *unknown* long period of time, re-association of the correct identity can be related *only* to the appearance of the object observed and the learned appearance model of all the objects observed so far. Video streams with many shot changes further reduce the relevance of motion continuity.

(2) Re-Id and Track deletion. MOT short-term re-identification is typically achieved by storing the appearance models of deactivated tracks for a fixed number frames such that an object can be either re-acquired or deleted (i.e. *forgotten*) [98, 7, 102]. However, simply setting a very large number frames after which to delete tracks, would require the explicit management of an undefined and large number of track identities with their corresponding appearance models undergoing cumulative learning. This issue has not been systematically investigated and therefore extending MOT approaches to include this long-term re-identification learning scenario it is not straightforward.

(3) MOT Datasets. MOT has been extensively studied with a prime focus on human body visual data where it is *not* reasonable to assume that clothes remain unchanged over very long-term periods of time as for face data. Consequently, relevant MOT datasets do not explicitly cover long-term re-acquisition and/or extended appearance variations [47, 60].

(4) Learning Setting. MOT methods have either offline or online processing mode [57] depending on whether observations from future frames are or are not utilized when handling the current frame. However, with the terms “incremental” and “cumulative” the reference here is to a learning setting rather then to a processing mode [88, 54, 25]. The term “online” alone typically used to characterize MOT methods does not reflect the concept of lifelong adaptation and cumulative learning in never-ending data streams that have changing statistics. In order to avoid confusion, we will refer to this learning setting as Multiple Object Cumulative Adaptation Learning (MOCAL).

Differently from SOT and MOT methods, in this paper we present a novel online self-supervised method that learns cumulative identity representations adapted to all the visual information observed so far. We evaluated our method on face visual data as it is more intrinsically pertinent to this learning setting. In order to focus on the aspects that distinguish our approach from MOT, we used datasets as in video face clustering [94, 87, 96, 100] that include large corpus of face objects with abrupt motions, extended appearance variations and very long-term occlusions.

Specifically, to achieve cumulative learning while handling the non-stationarity of the data stream, we update a representative dataset and use it as a memory of all the past visual information observed so far. To this aim, CNN face features [66, 13] are stored into a memory module and "distilled" based on their redundancy so that a compact and complete appearance representation of the individual identities is cumulative learned over time. The memory module consists of feature-identity pairs as recently introduced in [90, 36, 71]. Extracted face features from the current frame are used to both query and learn the memory model according to a Reverse Nearest Neighbor strategy [41]. The features returned by the memory are used to determine the final prediction of the identities. To avoid forgetting, no identity is explicitly deleted after a fixed number of frames has passed. As the memory increases, observed features are selectively removed only if there is subsequent information in their locality in representation space. This makes the representations of each identity more compact and discriminative since they are adapted to incorporate all the past data. When a memory budget is met, new information is written into the least used memory locations. It is further shown that the proposed incremental procedure for learning the memory module approximates asymptotically the case of infinite accumulation of feature data. A preliminary exploration of this work was presented in [69].

In the following, in Section 2, we cite the works that are related to our approach. In Section 3 contributions are provided. In Section 4 we expounded the approach in detail, in Section 5, experimental results are given and finally in Section 6 critical discussion and further experiments are provided.

2. Related Work

In this section the general and methodological issues raised in the introduction are examined in different bodies of existing literature. MOCAL setting is closely related to Continual Learning [15, 65] and Open-World learning [6]. We will describe each of them briefly highlighting the relationships/connections with our approach.

Continual Learning deals with the problem of sequentially learning a single model, preserving and reusing the previous knowledge while learning the new one. Instead Open-World learning deals with the problem of detecting new classes at test time (i.e. open-set) to avoid incorrect assignments to known classes. When new classes are incorporated in the model, then Continual Learning meets the problem of Open-World. In the open set evaluation protocol, learned face features [67, 13] with distance thresholding have shown to achieve reliable performance [40]. Our approach follows a similar strategy to detect novel identities also exploiting the fact that a single video frame eventually contains distinct face identities.

In Continual Learning, typically a sequence of tasks is learned one at a time, with label supervision, with all data of current task available and without revisiting past tasks. Task boundaries and class identities are therefore known at all times. This setting, is therefore not appropriate in applications that learn incrementally from unconstrained video streams. A recent and notable exception is provided by [2]. They learn face identities in a self supervised way. First they obtain face tracklets and then use this information to update the face representation. The tracks are then processed in chronological order so as to generate a non-i.i.d. stream of data. In our approach, representation is fixed and face-specific, but it is directly adapted from the data coming from a detector without requiring a multi-pass analysis of the video. According to this, our method allows learning in an online and cumulative fashion from an unconstrained video stream.

2.1. Multiple Object Tracking

An alternative approach that partially accomplishes the open-world and class-incremental learning (it does not perform cumulative learning) is Multiple Object Tracking (MOT) [46, 57, 17]. MOT exploits temporal self-supervision to automatically generate labels with data coming directly from the output of a detector. The major issue encountered by MOT when applied to cumulative learning is track management. Track identity creation and deletion are managed by two thresholds: a new identity is created when the object has been constantly detected for a certain number of frames while an identity is deleted if it is not associated for a duration of a predefined number of frames. The value of these thresholds typically depends on both the accuracy of detection models and the frame rate and are set in the order of few seconds [102]. Track deletion basically precludes MOT methods to perform long-term re-identification as required in MOCAL: objects that exit and re-enter the field of view after few seconds are managed as new identities. As a consequence, MOT methods cannot be directly applied to perform cumulative learning nor to handle unconstrained video streams.

In [100], video face clustering is exploited to adapt face appearance. Their method applies MOT in videos consisting of pre-segmented shots taken from different cameras. In order to take advantage of the continuity of the motion, each shot is processed independently to estimate tracklets. Face appearance adaptation learning is achieved with a further pass over the face image crops along the estimated tracklets by fine-tuning the CNN feature representation according to the triplet-loss [84]. This pass can be considered as addressing both adaptive and cumulative learning of the feature representation across the processed video. To overcomes the track deletion limit of MOT the fine-tuned features are then used in a final pass to cluster tracklets across multiple shots. The approach is similar to [2] except that the adaptation is not performed incrementally. Our approach follows the same intents of both [100] and [2] but formulates the problem as cumulative and online learning. Differently from [100] and [2] we can handle an infinite video stream. To this aim we leverage the success of recent tracking-by-detection approaches [39, 9, 99].

Tracking-by-detection has become the leading MOT paradigm exploiting both to the improved accuracy of CNN based object detectors [74, 53, 34, 14] and CNN feature representation [66, 13, 84, 91]. Performance with respect to earlier methods has been largely improved especially in the online processing modality. In [99], both Faster R-CNN detections and features learned using re-identification datasets [101] are combined obtaining a performance improvement by a margin of 30% with respect to the state of the art, showing that having higher-quality detections and feature representations reduces the need of complex association/tracking algorithms. Other similar tracking-by-detection methods have recently followed: [79, 93, 77, 7]. Specifically, [7] further simplifies the tracking-by-detection MOT paradigm by removing the optimal data-association step and the motion model. Our method exploits this simplified paradigm.

Among MOT methods operating online not based on tracking-by-detection, several interesting attempts has been proposed recently to favor short-term identity preserving (i.e. occlusion between objects) against the most favorable off-line methods exploiting future information. A few methods have exploited Single Object Tracking (SOT) to manage missing detections [102, 16, 95, 24]. Tracks deletion and appearance forgetting still limits the applicability of these methods in the MOCAL setting. In particular [24] addresses tracking multiple faces that exit and re-enter the field of view. The method exploits contextual relations (i.e. upper body appearance and relative camera poses) according to a graphical model to characterize the dependency between multiple objects. It consists of two phases: in the first phase the graphical model is learned off-line from some video sequences, in the second phase the appearance of face objects are learned online according to the SOT model described in [35]. However this method cannot handle an infinite video streams since it relies on pre-segmented shots to exploit motion continuity.

2.2. *Long-Term Single Object Tracking*

Another relevant research subject to our learning setting is long-term Single Object Tracking [37, 70, 12, 62, 89, 42, 56]. The aim of long-term SOT is to track a specific object over time and re-detect it when the object leaves and re-enters the scene. Only a few works on tracking have reported drift-free results on on very long video sequences [37, 23, 70, 33, 30] among the few, and only few of them have provided convincing evidence on the possibility of incremental appearance learning strategies that are asymptotically stable [37, 70]. However, all of these works perform incremental learning only to detect the object in the next frame and gradually forget the past information. In [100] authors evaluate a MOT baseline in which multiple TLD trackers [37] initialized with the ground-truth bounding box in the first frame are exploited. The baseline so defined can handle unconstrained videos avoiding to segment them into shots to exploit motion continuity.

2.3. *Learning With a Memory Module*

Inclusion of a memory mechanism in learning [45] is a key feature of our approach. On domains that have temporal coherence like Reinforcement Learning (RL), memory is used to store the past experience with some priority and to sample mini-batches to perform incremental/cumulative learning [61] [83]. This makes it possible to break the temporal correlations by mixing more and less recent experiences therefore handling the non-stationarity of data streams. More recently, Neural Turing Machine architectures have been proposed in [26, 27] and [81] that implement an augmented memory to quickly encode and retrieve new information. These architectures have the ability to rapidly bind never-before-seen information after a single presentation via an external memory module. However, in these cases, training data are still provided supervisedly and the methods are not primarily designed for handling video streams.

In [36] a memory module consisting of feature-value pairs to perform predictions based on past knowledge is proposed. Features are activations of the penultimate layer of a deep neural network (i.e. the internal feature representation), and values are the ground-truth targets. The output of the penultimate layer of the neural network is used as query to the memory module and the nearest neighbor returned by the memory is used as the final network prediction. As the memory increases it becomes more useful since it can give predictions that leverage on knowledge from past data with similar features. We use this basic strategy in which the feature-value pair consists in a face specific feature and its associated identity. One main limitation of [36] is in the lack of a mechanism to forget redundant observations to make rooms to novel fresh data. The work [80] suggests a memory based forgetting strategy based on the principle of spatio-temporal locality. We follow a similar principle in which observations are forgotten if there is subsequent information according to a distance ratio criterion between deep features.

3. Contributions

Our contributions can be summarized as follows:

1. We present a novel online method for the task of learning the appearance of face identities from unconstrained video streams. As video streams are infinitely long, this requires online accumulation and preservation of all the past visual knowledge observed so far.
2. We propose a memory module that achieves online cumulative learning in two different ways. (a) Avoiding the explicit deletion of object identities after a fixed number of frames has passed. (b) Selectively removing observed features depending on whether subsequent information in their locality is available in representation space.
3. The proposed method firstly addresses very long-term object re-acquisition in online MOT processing mode: when an object leaves the field-of-view and then reappears, it is not treated as an unseen object with a novel Id.
4. The proposed strategy is shown to be asymptotically stable. We argue this is a critical issue for any system that claims to operate lifelong.
5. The proposed method referred as IdOL (Identity Online Learning), performs comparably with offline approaches exploiting future information in the task of multiple face tracking in unconstrained videos while it achieves better performance in the face identification.

4. The proposed approach

The block diagram of the solution proposed is shown in Fig.1. We used the state of the art *Tiny Face Detector* [32] for detection and the *VGGFace* features [67] to represent faces. A memory module is used to collect the face features. In the ideal case (i.e. perfect invariance of the representation), observations of the same subject originate the same features. In the real case, we must expect that observations of the same subject under changes of pose or illumination or partial occlusions originate different (although correlated) features. The matching module is a discriminative classifier that associates each new observation to the most similar past observations already in the memory. The memory controller has the task of discarding redundant features: highly similar features of the same subject having comparable distance feature already in the memory module. Ideally, a new identity should be created whenever a new individual is observed that has not been observed before.

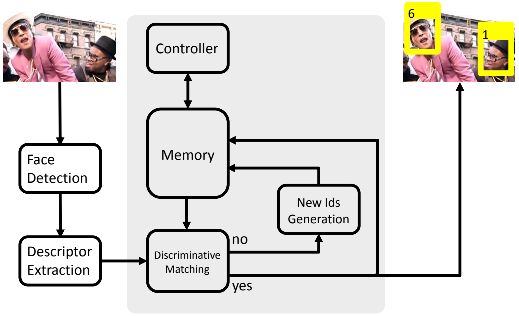


Figure 1: Block diagram of the incremental identity learning with basic workflow.

We loosely follow [36] and the memory module at time t is represented as:

$$\mathcal{M}(t) = \{(\mathbf{x}, \mathrm{Id}, e, a)_i\}_{i=1}^{N(t)} \quad (1)$$

where i is the index of a memory element and $\mathbf{x}$ is a deep feature, Id is the face identity associated, e is a value referred to as *eligibility* that accounts for the relevance of the item to be learned (discussed in the following), a is a value that tracks the age from the last match, and *N(t)* is the number of features in the memory at time t. We extend the feature-value pair (i.e. $\mathbf{x}$–Id) and the age in [36] by adding a further scalar quantity.

The mechanisms of identity matching, construction of the identity models, self-supervision using temporal coherence and the asymptotic behavior of the method are separately addressed in detail in the following subsections.

4.1. Reverse Nearest Neighbor Matching

The Nearest Neighbor distance ratio criterion [55] allows matching based on a discriminative rule between

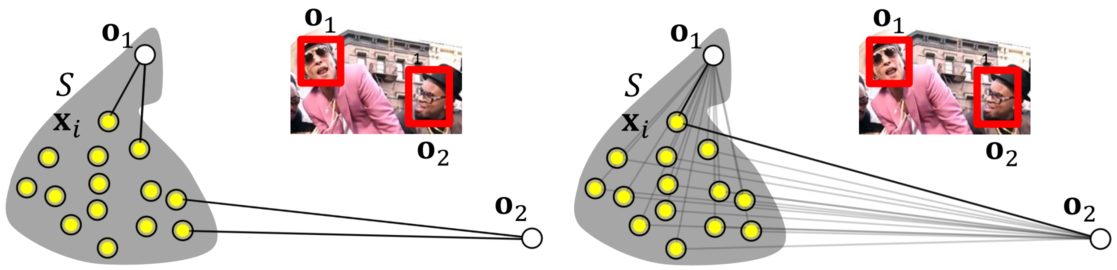


Figure 2: Nearest Neighbor (*left*) and Reverse Nearest Neighbor (*right*) matching with distances between the stored features and the observations. The stored features $\mathbf{x}_i$ in the grey area all have the same identity. ReNN can assess matching between $\mathbf{x}_i$ and $\mathbf{o}_1$ according to the distance ratio criterion.

the most and the second most similar sample. Unfortunately, in our learning scenario, we cannot exploit Nearest Neighbor with distance ratio criterion to assess matching. Since it is likely that detected faces of the same subject in consecutive frames have little differences from one frame to the following, similar features having comparable distances to the nearest and the second nearest will rapidly be stored in the memory module. As a consequence, the distance ratios of observations to the nearest and the second nearest feature in memory will be close to 1 and matchings are undecidable in most cases. To solve this problem, we propose to use Reverse Nearest Neighbor (ReNN) with the distance ratio criterion [41].

With ReNN, each feature in memory is NN-matched with the features of the observations in the incoming frame and distance ratio is used to assess matching. Fig. 2 explains this matching mechanism for a sample case. The features $\mathbf{o}_1$ has the same identity as the $\mathbf{x}_i$ in the memory while $\mathbf{o}_2$ has a different identity. Due to the fact that the $\mathbf{x}_i$ are close to each other, the NN-distance ratios of $\mathbf{o}_1$ and $\mathbf{o}_2$ to their nearest and second nearest $\mathbf{x}_i$ are are both close to 1 and both matchings result to be NN-undecidable (Fig. 2-*left*). Instead, with ReNN, the NN-distance ratios between the $\mathbf{x}_i$ and $\mathbf{o}_1$ and $\mathbf{o}_2$ clearly assess the matching with $\mathbf{o}_1$ (Fig. 2-*right*). The set of $\mathbf{x}_i$ that are ReNN matched to the observations at time t can be written as:

$$\mathcal{M}^+ = \left\{(\mathbf{x}, \mathrm{Id}, e, a)_i \in \mathcal{M}(t) \mid \frac{d_i^1}{d_i^2} < \bar{\rho}\right\} \tag{2}$$

where $\frac{d_i^1}{d_i^2}$ is the distance ratio between $\mathbf{x}_i$ and the nearest and second nearest face features in the frame at time t and $\bar{\rho}$ is the distance ratio threshold.

4.2. Learning the Memory Module

Collecting matched features indefinitely will soon accumulate features in the memory module and a large amount of redundant information will be included for each identity model. To avoid such redundancy, we associate to each i-th feature-identity pair a dimensionless quantity e_i referred to as *eligibility-to-be-learned* (shortly *eligibility*) that dynamically indicates the level of redundancy of the feature to be learned as representative of the identity. Eligibility is set to 1 when the feature is loaded into the memory and is decreased at each match with the observations according to:

$$e_i(t+1) = \eta_i \, e_i(t) \;\text{ with }\; \eta_i = \left[\frac{1}{\bar{\rho}} \frac{d_i^1}{d_i^2}\right]^{\alpha}, \tag{3}$$

where the matching threshold $\bar{\rho}$ of Eq. 2 is used for normalization and α to dilate the effect of the distance-ratio. When doing this, we also reset the feature age $a_i = 0$. As the eligibility e_i of a face feature $\mathbf{x}_i$ drops below a given threshold $\bar{e}$ (that happens after a number of matches), the feature is no more eligible to be learned as representative of the identity and is removed from the memory.

Eq. 3 down-weights eligibility as a function of the distance ratio at a rate proportional to the success of matching in consecutive frames. According to this, eligibility allows to take into account spatio temporal redundancy in a discriminative way. The equation it is a generalization of the Apollonious circle [1] to multiple dimensions. As shown in Fig. 3, features in regions close to $\mathbf{o}_1$ and far from $\mathbf{o}_2$ (dark red) have low η and therefore their eligibility is more down-weighted (they will have higher chance to be replaced in the future). Features in regions far from $\mathbf{o}_1$ and $\mathbf{o}_2$ (light red) have higher η and their eligibility is less down-weighted and their chance of not being discarded is higher. This asymmetry promotes diversity in the open space and defines a learning

[1] Apollonius of Perga (c. 262 BC - c. 190 BC) showed that a circle may also be defined as the set of points in a plane having a constant ratio of distances to two fixed foci.

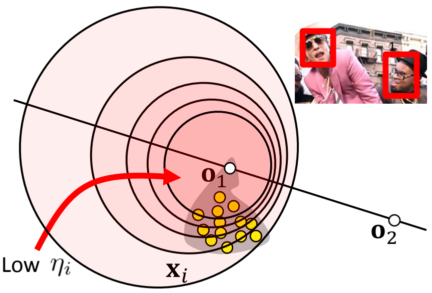


Figure 3: Learning the memory module. The 2D shape of the density function (shown by level curves) down-weighting the eligibility associated to each matched feature. Features $\mathbf{x}_i$ in proximity of the observed feature $\mathbf{o}_1$ have their eligibility decreased (low values of η) to reflect their redundancy. The asymmetric shape of the density encourages more diversity in the *open space* far from the identity $\mathbf{o}_2$ rather than close.

schema well suited for the *open world* face recognition scenario.

Our method operates on-line and does not require any prior information about how many identity classes will occur and can run for an unlimited amount of time. However, if the number of identities increases indefinitely the eligibility-based exemplar removal may not be sufficient to avoid memory overflow. Similarly to [82, 36], we remove from the memory the least recently matched exemplars (those with the highest value of the paramter a in Eq. 1), following the Least Recently Used Access (LRUA) strategy. This also allows to remove false positives by the detector that have not received other matches for a long time.

4.3. Self-supervision

The cumulative learning mechanism of the memory module breaks the temporal coherence of the data stream (i.e. *non-iid*) by mixing more and less recent recent observations. Nevertheless temporal coherence is used as form of self-supervision in the assignment of novel identities to limit their fragmentation and proliferation.

Assuming that faces of the same individual have similar features in consecutive frames, in the case in which an observation does not match the feature in memory, its feature is included in the memory with a new identity only if the same identity is assigned also in the following frames (two consecutive identity assignments and at least one matching in the following three frames was experimentally verified to provide good results). With this form of verification potential novel identities in the current frame are included in the memory only if at least one known identity is recognized. Since recognition is obtained according to the RNN with the distance ratio and since observations taken from a single frame derive from distinct identities, the unmatched identities in the current frame are known to be reasonably distant (i.e. different) from the recognized ones and are considered potentially novel.

In the case in which no observations match with features in memory, new identities are assigned to the non-matched observations only if the same situation persists for a time interval.

ReNN matching can determine ambiguous assignments when distinct face observations match with features of the same identity, or an observation matches with features of different identities in memory. In the first case we assign no identity to the observations (i.e. identities in the current frame are unique and therefore duplicated Ids in the same frame are not allowed). In the second case, we assign the most represented identity, i.e. that with the largest number of features in memory.

The complete algorithm of our IdOL (Identity On-line Learning) method for incremental identity learning, is reported in pseudocode in Algorithm 1. We indicate with $\mathcal{O}$ as the set of all the features extracted from the bounding boxes reported by the face detector in the current frame, and with $\mathcal{M}$ as the set of features in the memory module. Correspondences between $\mathcal{M}$ and $\mathcal{O}$ are computed in line 4 according to the Reverse Nearest Neighbour matching. The sets $\mathcal{M}^{+}$ and $\mathcal{O}^{+}$ indicate the elements that have established a direct correspondence. The set $\mathcal{I}_{curr}$ will contain (if any) the identity labels of novel subjects (not present in the memory model) detected in the current frame. It is initialized to the empty set for each novel frame (line 8). In line 9 all the matched observations $\mathcal{M}^{+}$ in the memory module are updated according to Eq. 3 and have their age reset to 0. Then, in line 10 potential identities $\mathcal{I}$ are predicted and subsequently intersected with those predicted in the previous frame $\mathcal{I}_{prev}$ to obtain the set $\mathcal{I}_{curr}$ of the identity labels estimated for the current frame (line 11). In line 12 the estimated label identities together with their observed features $\mathcal{O}_{curr}$ will be added in to the memory module as novel Ids with their eligibility and age values are set to 1 and 0 respectively. With an excess of notation we denote $\mathit{1}$ and $\mathit{0}$ as arrays of elements of value 1 and value 0 respectively. Their length is the same as the number of elements in $\mathcal{I}_{curr}$. In line 13 the potential identity labels of the previous frame $\mathcal{I}_{prev}$ are updated with those estimated in the current frame. In line 19 the t_{nc} counter is incremented when a frame has no matched correspondences. It get reset to 0 the first time a match occurs (line 14) or after that a number of frames $\hat{t}_{nc}$ are elapsed (line 16). In the latter case all the detected observations are declared as novel. In line 20 all observations from memory that are never matched after being included are removed if their age is greater than a given threshold $\bar{a}$.

Algorithm 1: IdOL - Identity On-line Learning

```
   Input: The video stream.
   Output: Assigned identities $\mathcal{I}_{curr}$ in the current frame.
1  repeat
2      Detect faces in the current frame;
3      Extract observations features $\mathcal{O}$;
4      Establish correspondences:
         $\mathcal{M}^+ \leftrightarrow \mathcal{O}^+ = \text{ReNN}(\mathcal{M}, \mathcal{O})$;
5      Identify non-matching memory elements:
         $\mathcal{M}^- = \mathcal{M} \setminus \mathcal{M}^+$;
6      Identify non-matched observations: $\mathcal{O}^- = \mathcal{O} \setminus \mathcal{O}^+$;
       // Case with matched observations
          (eligibility updating and temporal
          coherence verification)
7      if $|\mathcal{O}^+| > 0$ then
           // Initialize the set of identities to
              be included in the memory
8          $\mathcal{I}_{curr} = \emptyset$;
           // Update the eligibility with the
              matched observations
9          $\mathcal{M} = \{(\mathbf{x}, \text{Id}, \eta e, 0)_i \mid \forall (\mathbf{x}, \text{Id}, e, a)_i \in$
             $\mathcal{M}^+\} \cup \mathcal{M}^-$;
           // Assign known and novel identities to
              the observations
10         $\mathcal{I} = \mathit{MajorityId}(\mathcal{M}^+ \leftrightarrow \mathcal{O}^+) \bigcup \mathit{NewId}(\mathcal{O}^-)$;
           // Keep identities assigned in two
              consecutive frames
11         $\mathcal{I}_{curr} = \mathcal{I} \cap \mathcal{I}_{prev}$;
           // Include them in the memory module
              with their observations
12         $\mathcal{M} = \mathcal{M} \cup \{(\mathcal{O}_{curr}, \mathcal{I}_{curr}, 1, 0)\}$ ;
           // Keep the assigned identities for the
              next frame
13         $\mathcal{I}_{prev} = \mathcal{I}$;
14         $t_{nc} = 0$;
       // Case with no matched observations
          (novel identity assignment after a time
          interval has elapsed)
15     else if $t_{nc} > \bar{t}_{nc}$ then
16         $\mathcal{M} = \mathcal{M} \cup \{(\mathcal{O}, \mathit{NewId}(\mathcal{O}), 1, 0)\}$;
17         $t_{nc} = 0$;
18     else
19         $t_{nc} = t_{nc} + 1$
20     $\mathcal{M} = \mathcal{M} \setminus \{(\mathbf{x}, \text{Id}, e, a)_i \in \mathcal{M} \mid a_i > \bar{a}, e_i = 1\}$;
21     $\mathcal{M} = \mathcal{M} \setminus \{(\mathbf{x}, \text{Id}, e, a)_i \in \mathcal{M} \mid e_i < \bar{e}\}$;
22     $\mathcal{M} = \{(\mathbf{x}, \text{Id}, e, a+1)_i \mid \forall (\mathbf{x}, \text{Id}, e, a)_i \in \mathcal{M}\}$;
23     Apply LRUA;
24 until True;
```

4.4. Asymptotic Stability

The cumulative learning procedure described above stabilizes asymptotically around the probability density function of the features of each identity. This is guaranteed by the fact that the memory updating rule of Eq. 3 is a *contraction* that converges to its unique *fixed point*[2] according to the Contraction Mapping Theorem [4]:

> *Banach Contraction Mapping Theorem*
> Let (X, d) be a complete metric space and $M : X \mapsto X$ be a map (referred to as *contraction*) such that
>
> $$d(M(x), M(x')) \leq c \cdot d(x, x')$$
>
> for some $0 < c \leq 1$ and all x and $x' \in X$. Then M has a unique *fixed point* in X. Moreover, for any $x \in X$ the sequence of iterates x, $M(x)$, $M(M(x)), ..., M(...M(M(x)))$ converges to the *fixed point*.

In our case, the memory updating mechanism of Eq. 3:

$$e(t+1) = \eta\, e(t) \quad \text{with} \quad \eta = \left[\frac{1}{\bar{\rho}} \frac{d^1}{d^2}\right]^{\alpha}$$

being $\eta \in (0, 1]$, satisfies the conditions of the theorem above. It can be observed that the value $e = 0$ is the fixed point of this equation and corresponds to the case of an infinite accumulation of samples. In such a case, the Nearest Neighbor classifier error is bounded by twice the Bayes risk [18]. In our case, to have a finite number of samples, a threshold $\bar{e}$ close to 0 can be set that approximates with continuity the case of the infinite sample set.

5. Comparative evaluation

Our method is evaluated over publicly available datasets, namely *Music* and *Big Bang Theory* [5] and *QMUL multi-face dataset* [58]. We use the MOTA (Multiple Object Tracking Accuracy) performance metric defined as [60]:

$$\text{MOTA} = 1 - \frac{\sum_t (\text{FN}_t + \text{FP}_t + \text{IDS}_t)}{\sum_t \text{GT}_t} \tag{4}$$

where GT_t, FN_t, FP_t and IDS_t are respectively the number of ground truth objects, the number of false negatives, the number of false positives and the number of identity switches at each time t.

We compare our solution with the performance of the offline methods in [100]:

- mTLD running the TLD tracker in each shot [38]

[2] A fixed point of a function is an element of the function's domain that is mapped to itself by the function.

Table 1: IDS and MOTA comparative for the methods in [100] (*Music* dataset)

	Apink		BrunoMars		Darling		GirlsAloud	
Method	IDS ↓	MOTA ↑	IDS ↓	MOTA ↑	IDS ↓	MOTA ↑	IDS ↓	MOTA ↑
mTLD*	31	−2.2	35	−8.7	24	−22.0	9	−1.1
mTLD2*	173	77.4	278	52.6	278	59.8	322	46.7
ADMM*	179	72.4	428	50.6	412	53.0	487	46.6
IHTLS*	173	74.9	375	52.7	381	62.7	396	51.8
Siamese*	124	79.0	126	56.7	214	69.5	112	51.6
Triplet*	140	78.9	126	56.6	187	69.2	80	51.7
SymTriplet*	78	80.0	105	56.8	169	70.5	64	51.6
IdOL (VGGFace/VGG16-4096)	191	55.1	420	48.8	449	62.1	339	49.3
IdOL (VGGFace2/ResNet-2048)	178	61.4	375	59.4	432	63.0	315	55.0
IdOL (VGGFace2/SeNet-128)	177	62.6	367	60.1	427	64.2	306	55.8

	HelloBubble		PussycatDolls		Tara		Westlife	
Method	IDS ↓	MOTA ↑	IDS ↓	MOTA ↑	IDS ↓	MOTA ↑	IDS ↓	MOTA ↑
mTLD*	7	−3.5	24	3.1	130	1.4	20	−34.7
mTLD2*	139	52.6	296	68.3	251	56.0	177	58.1
ADMM*	115	47.6	287	63.2	251	29.4	223	62.4
IHTLS*	109	52.0	248	70.3	218	35.3	113	60.9
Siamese*	105	56.3	107	70.3	106	58.4	74	64.1
Triplet*	82	56.2	99	69.9	94	59.0	89	64.5
SymTriplet*	69	56.5	82	70.2	75	59.2	57	68.6
IdOL (VGGFace/VGG16-4096)	88	51.4	83	30.7	270	39.5	76	58.9
IdOL (VGGFace2/ResNet-2048)	92	49.1	80	33.7	257	42.3	70	64.1
IdOL (VGGFace2/SeNet-128)	85	51.5	77	35.2	254	42.5	68	63.9

* Values reported from [100]

- mTLD2 a modified versions of TLD that generates shot-level trajectories [100]
- ADMM [22];
- IHTLS [3];
- *Siamese*, *Triplet* and *SymTriplet* methods [100],

All these methods operate offline (so they exploit both past and future frames to learn identities). They apply the *Headhunter* version of DPM detector by [59] and detections are then linked into shot-level tracklets. Tracklets across shots are hence merged into trajectories using Hierarchical Agglomerative Clustering [92]. The *Siamese*, *Triplet* and *SymTriplet* methods [100] have a sophisticated refinement of identity assignment. Tracklets are used in pairs (in the *Siamese*) or triplets (in the *Triplet* and *SymTriplet*) to fine-tune an AlexNet-based CNN pretrained on the CASIA-WebFace and the descriptor of the fine-tuned CNN is finally used to link tracklets into shot-level tracklets. Comparison with these methods were made over the Music and Big Bang Theory datasets.

Tab. 1 provides a comparative overview of MOTA and IDS scores for the videos of the *Music* dataset. These are YouTube videos of live vocal concert recordings with very frequent shot changes (i.e. unconstrained videos), views from different cameras and special effects. There are a limited number of annotated characters in continuous fast movement. Faces have large variations of appearance due to rapid changes in pose, scale, makeup, illumination, camera motion and occlusions. In total, there are 117,598 face detections and 3,845 face tracks annotations. Our IdOL method has lower MOTA in most videos (although almost the same of ADMM and IHTLS). However, it has comparable IDS for HelloBubble, Apink PussycatsDolls and Westlife videos and lower IDS for Tara.

As feature representation is one of the main component of our method, Tab. 1 reports performance evaluated according to three different feature representations:

- 4096-dimensional feature learned from VGGFace with VGG16 architecture [66] (VGGFace/VGG16-4096)
- 2048-dimensional feature learned from MS-Celeb-1M and fine-tuned on VGGFace2 dataset [13] with SeNet [31] (VGGFace/SeNet-128)
- 128-dimensional feature learned from MS-Celeb-

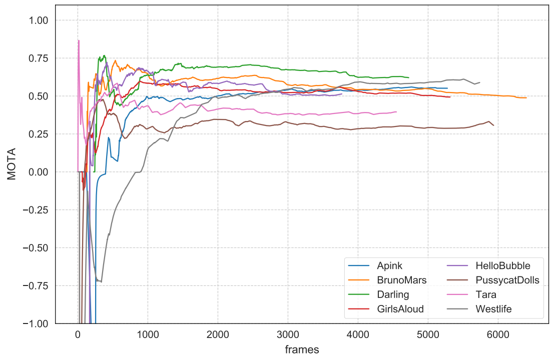


Figure 4: MOTA computed at each frame for the videos in the *Music* dataset

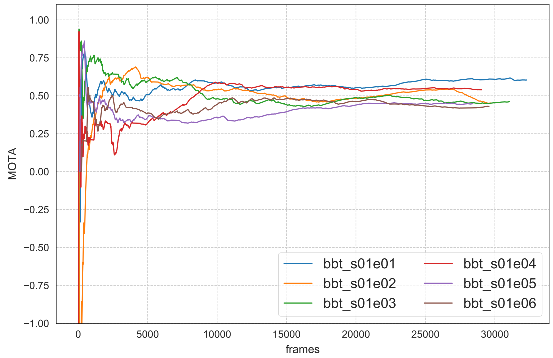


Figure 5: MOTA computed at each frame for the videos in the *Big Bang Theory* dataset

1M and fine-tuned on VGGFace2 dataset with ResNet [28] (VGGFace2/ResNet-2048) [3]

As it can be noticed, performance follows the increasing quality of the different feature representation evaluated. This is due to the expressive power of more competitive CNN architectures and the exploitation of richer training datasets.

Fig.4 shows the plot of MOTA of our method computed at each frame for the VGGFace/VGG16-4096 features. Due to the incremental learning mechanism, at the beginning there is not sufficient information available so the identity models are largely incomplete and a large number of errors may occur. As more and more observations are received that contain different views and conditions of the faces, MOTA stabilizes. In the *Music* dataset, asymptotic values of MOTA were reached approximately after 1000 frames for all the videos, despite of the different editing and contents.

In order to assess our method on longer video sequences also in conditions similar to surveillance contexts, we compared performance also on the *Big Bang Theory* dataset. This dataset collects six episodes of *Big Bang Theory* TV Sitcom, Season 1. These are much longer videos (approx 20' each) with indoor ordinary scenes under a variety of settings and illumination conditions. They contain a much larger number of identities with crowded scenes. Also in this case, faces have large variations of appearance due to rapid changes in pose, scale, makeup, illumination, camera motion and occlusions.In total, there are 373,392 face detections and 4,986 face tracks annotations.

Table 2 reports MOTA and IDS scores for the videos of the *Big Bang Theory* dataset and Fig. 5 shows the plots of MOTA computed at each frame. It can be noticed that considerations similar to those drawn for the *Music* dataset hold also in this case, despite of the differences between the two datasets. The presence of less frequent cuts and less extreme conditions due to editing effects and camera takes than in the *Music* dataset determines sensibly lower Identity Switch values and closer MOTA values in almost all the videos. MOTA plots have earlier convergence to their asymptotic values. They all share similar behavior due to the uniform style of the series.

We further compare our solution with performance of methods reported in [24]:

- Tracking-Clustering [11]
- Min-Cost Flow [72]
- CRF [24]
- M^3 Networks [24]

Specifically Tracking-Clustering is a modified approach of [11] in which the Haar cascade face detector is replaced with a DPM detector and Min-Cost Flow performs offline optimal data association. Comparison with these methods were made over the *QMUL multi-face dataset*. This dataset consists of three single shot video sequences with four subjects entering and exiting the field of view, namely FRONTAL, FAST and TURNING. Although captured by a static camera, all three video sequences contain intense face motions and occlusions. In addition, subjects change their face poses frequently in the TURNING sequence and perform fast movements in the FAST sequence.

Tab. 3 provides a comparative overview of MOTA and IDS scores for the videos of the QMUL Multiple Face Dataset. As can be noticed our approach consistently outperforms the other four compared methods on

[3] Pretrained models are available at `https://github.com/ox-vgg/vgg_face2`

Table 2: IDS and MOTA comparative for the methods in [100] (*Big Bang Theory* dataset)

	BBT_s01E01		BBT_s01E02		BBT_s01E03		BBT_s01E04		BBT_s01E05		BBT_s01E06	
Method	IDS ↓	MOTA ↑	IDS ↓	MOTA ↑	IDS ↓	MOTA ↑	IDS ↓	MOTA ↑	IDS ↓	MOTA ↑	IDS ↓	MOTA ↑
mTLD*	1	−16.3	1	−7.6	5	−2.1	0	−15.9	1	−15.5	0	−3.9
mTLD2*	223	58.4	174	43.6	142	38.0	103	11.6	169	46.4	192	37.7
ADMM*	323	42.5	395	41.3	370	30.8	298	9.7	380	37.4	527	47.5
IHTLS*	312	45.7	394	42.4	376	33.5	295	13.3	360	33.8	515	43.2
Siamese*	144	69.0	116	60.4	109	52.6	85	23.0	128	60.7	156	46.2
Triplet*	164	69.3	143	60.2	121	50.7	103	18.0	118	60.5	185	45.4
SymTriplet*	156	72.2	102	61.6	126	51.9	77	19.5	90	60.9	196	47.6
IdOL	26	60.4	55	45.2	14	46.1	75	53.9	35	44.7	204	43.0

* Values reported from [100]

Table 3: IDS and MOTA comparative for the methods in [24] (*QMUL multi-face dataset*)

	Average	
Method	IDS	MOTA
Tracking-Clustering*	23	61.8
Min-Cost Flow*	29	53.7
CRF*	20	65.2
M^3 Networks*	17	68.8
IdOL (VGGFace)	10.0	81.5
IdOL (VGGFace2/SeNet)	**2.3**	**87.5**
IdOL (VGGFace2/ResNet)	2.3	87.2

* Values reported from [24]

both MOTA and IDS. Since the dataset is composed by single shot videos, [24] can operate online as our method (i.e. the offline shot segmentation pass is not required).

6. Critical discussion and additional experiments

The MOTA score is a largely accepted metrics for Multiple Object Tracking. However it has clear limitations to assess the performance of cumulative learning as in the MOCAL learning setting. In the following, we discuss such limitations and perform additional evaluations.

6.1. Influence of detection

MOTA produces a cumulative score considering False Positives, False Negatives and Identity Switches. Typically, the number of False Positives and False Negatives are much larger than Identity Switches (see Table 5 f.e.). False Positives of MOTA are essentially determined by false positive detections, while False Negatives are in part due to missed detections and in part to the case in which no identity is assigned to a face observation. From the above it descends that MOTA score can be largely influenced by the performance of the detector.

While it can be presumed that a more effective detector has little influence on the performance of the MOT methods reported in the comparison (these methods operate off-line and most tracklets due to erroneous detections can be removed using the future information), the effectiveness of the detector largely influences the performance of online incremental identity learning, since future information cannot be exploited in this case. A key requirement for our task is therefore that the detector has as few False Positives and False Negatives as possible. The *Headhunter* detector was verified being clearly inadequate to this end. Table 4 shows the performance gap of the IdOL method with the *Headhunter* and the *Tiny Face Detector*, for the T-ARA video of the *Music* dataset (the detections of the *Headhunter* were released only for this video by the authors). However, since the *Tiny Face detector* is capable to detect also very small-sized faces (say less than 40 pixels), given the fact that most of these faces are not ground-truth annotated in the *Music* and *Big Bang Theory* datasets (see Figure 6 as an example), it happens that False Positives are counted whenever the detector detects a non annotated face. Tables 5 and 6 show the increase of performance achievable by the IdOL method in the ideal condition of no False Positive detections, considering the ground truth bounding boxes (left side).

False Positives of the detector have the additional drawback of increasing identity switching and the number of wrong new identities. In the IdOL method, the mechanism of Temporal Coherence verification limits the proliferation of such new identities. The effects of the Temporal Coherence verification are shown in Table 7 for the *Music* dataset using the ground truth bounding boxes as detections. While it may increase False Negatives, it avoids the increase of Identity Switches.

Table 4: Influence of detection: FP FN MOTA for different detectors (*TARA* video of *Music* dataset)

	Tara		
Method	FP ↓	FN ↓	MOTA ↑
IdOL *HeadHunter* detector	1939	8592	25.6
IdOL *Tiny-face* detector	259	8259	39.5

Table 5: Influence of detection: IDS FN FP MOTA of IdOL using ground-truth bounding boxes versus Tiny Detector bounding boxes (*Music* dataset)

	Ground Truth Bounding Box				Tiny Detector			
Video	IDS ↓	FN ↓	FP ↓	MOTA ↑	IDS ↓	FN ↓	FP ↓	MOTA ↑
Apink	130	2105	0	69.3	191	2627	446	55.1
BrunoMars	391	3644	0	75.8	420	4178	3950	48.8
Darling	361	2620	0	68.7	449	2278	887	62.1
GirlsAloud	469	4837	0	67.6	339	6691	1272	49.3
HelloBubble	160	707	0	83.4	88	2150	301	51.4
PussycatDolls	316	4697	0	64.9	83	7050	2764	30.7
Tara	542	5210	0	60.4	270	8259	259	39.5
Westlife	138	1433	0	86.2	76	3403	1198	58.9

6.2. *Cluster Purity*

The MOTA score computes a cumulative performance score until the time of evaluation based on instantaneous measures of False Negatives, False Positives and Identity Switches. According to this, if in a sequence an Identity Switch occurs at a frame and the new (incorrect) identity is confirmed in the following frames, MOTA counts one Identity Switch only, at the frame at which it occurred. Instead in the case above, to assess the quality of online learning we should count as many Identity Switches as the times the original identity has been mismatched. According to this, it appears that MOTA is not fully adequate to measure the performance of on-line identity learning. A better metrics is the Weighted Cluster Purity (WCP), defined as [100]:

$$WCP = \frac{1}{M}\sum_{c} m_c p_c \tag{5}$$

where M is the number of identities detected in the video, c the index of the cluster, m_c the number of identity instances in the cluster and p_c the cluster purity, measured as the ratio between the most occurred identity in the cluster and m_c.

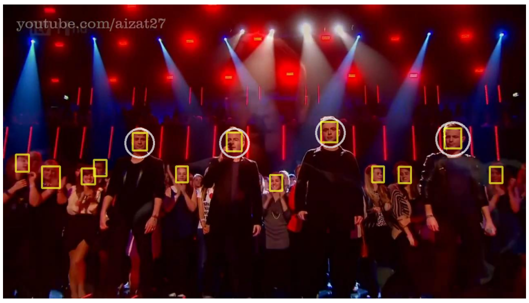


Figure 6: Detections of the *Tiny-face* detector for a sample frame (*Music* dataset). Most of the small bounding boxes are not annotated as faces in the ground-truth. Ground-truth annotated faces are circled.

Tables 8 and 9 present the WCP scores for the *Music* and *Big Bang Theory* datasets, and compare the IdOL method with respect to a few MOT methods of [100]. In almost all the cases the IdOL method method has largely better WCP scores. The lower score in the Apink video with respect to *Siamese*, *Triplet* and *Symtriplet* methods can be ascribed both to the ethnicity bias of the *VGGFace* features and to the effect of fine tuning on these methods.

Fig. 7 and 8 show the WCP plots for the videos in the *Music* and the *Big Bang Theory* datasets. At each frame WCP is calculated from the beginning up to that frame. It can be noticed that for the *Big Bang Theory* plots rapidly converge to the asymptotic values (each cluster contains a sufficiently complete description of the identity and features of different identities have been discarded). In some videos of the *Music* dataset, the presence of very frequent discontinuities and extreme conditions makes less effective the mechanism for keeping identity switches low.

Table 6: Influence of detection: IDS FN FP MOTA of IDoL using ground-truth bounding boxes versus Tiny Detector bounding boxes (*Big Bang Theory* dataset)

Video	Ground Truth Bounding Box				Tiny Detector			
	IDS ↓	FN ↓	FP ↓	MOTA ↑	IDS ↓	FN ↓	FP ↓	MOTA ↑
bbt_s01e01	218	1361	0	96.0	26	4631	10 875	60.3
bbt_s01e02	178	3947	0	86.3	55	5525	10 914	45.2
bbt_s01e03	191	6999	0	79.6	14	9668	9285	46.1
bbt_s01e04	341	2345	0	92.1	75	8615	7054	53.9
bbt_s01e05	381	3130	0	89.8	35	9919	9009	44.7
bbt_s01e06	559	5449	0	87.4	204	15 872	11 103	43.0

Table 7: Influence of Temporal Coherence: IDS FN MOTA of IDoL using Temporal Coherence versus IdOL without Temporal Coherence (*Music* dataset)

Video	IdOL with Temporal Coherence			IdOL without Temporal Coherence		
	IDS ↓	FN ↓	MOTA ↑	IDS ↓	FN ↓	MOTA ↑
Apink	130	2105	69.3	265	1558	74.9
BrunoMars	391	3644	75.8	741	2301	81.8
Darling	361	2620	68.7	655	2017	72.0
GirlsAloud	469	4837	67.6	897	3927	70.6
HelloBubble	160	707	83.4	295	358	87.5
PussycatDolls	316	4697	64.9	884	3787	67.3
Tara	542	5210	60.4	1025	4436	62.4
Westlife	138	1433	86.2	274	1105	87.9

Fig. 14 shows sample frames of the *Big Bang Theory* and *Music* videos with the detected faces and their assigned identities. For each video two frames are shown where the same persons are taken in different conditions, with large appearance variations due to partial occlusions (Figs. 14b, 14c, 14h, 14g), pose changes (Figs. 14a, 14c, 14e, 14f), aspect change (Figs. 14d, 14e) and in-plane rotations (Figs. 14g). It can be noticed that the learning mechanism is able to distinguish the same identity also in the presence of such large variations. Fig. 14i evidences the effect of the ethnicity bias of the *VGGFace* features. In this case, the method is not able to predict unique identities for the faces and does not make any identity assignments.

The plots in Fig. 9 show the number of identities learned by the IdOL method at each frame in comparison with the ground truth for the videos of the *Big Bang Theory* dataset. For the sake of comparison we also

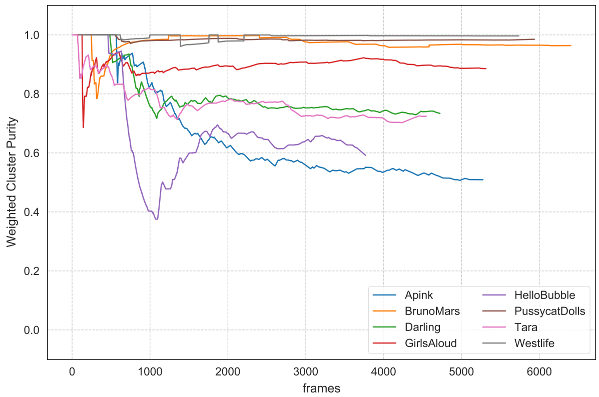


Figure 7: Weighted Cluster Purity computed at each frame for the videos in the *Music* dataset

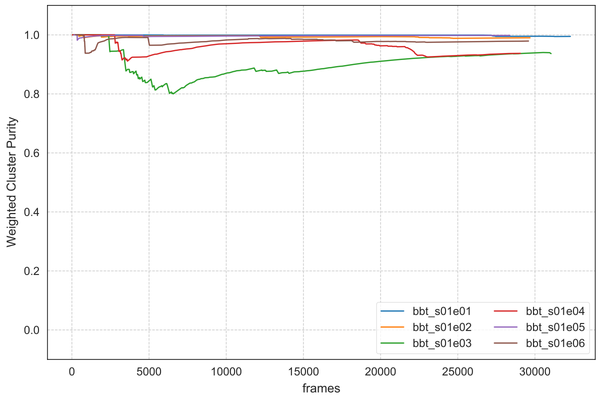


Figure 8: Weighted Cluster Purity computed at each frame for the videos in the *Big Bang Theory* dataset

Table 8: Weighted Cluster Purity score comparative. *Music* dataset.

Method	Apink	Bruno Mars	Darling	Girls Aloud	Hello Bubble	Pussycat Dolls	T-ara	Westlife
VGG-Face*	0.24	0.44	0.20	0.31	0.29	0.46	0.23	0.27
Siamese*	0.48	0.88	0.46	0.67	0.54	0.77	0.69	0.54
Triplet*	0.60	0.83	0.49	0.67	0.60	0.77	0.68	0.52
SymTriplet*	**0.72**	0.90	0.70	0.69	**0.64**	0.78	0.69	0.56
IdOL (VGGFace/VGG16-4096)	0.51	**0.96**	0.73	0.89	0.59	**0.98**	0.72	**0.99**
IdOL (VGGFace2/ResNet-2048)	0.72	0.92	0.78	0.92	0.47	0.97	0.74	0.99
IdOL (VGGFace2/SeNet-128)	**0.73**	0.91	**0.79**	**0.93**	0.48	0.98	**0.78**	0.95

* Values reported from [100]

Table 9: Weighted Cluster Purity score comparative. *Big Bang Theory* dataset.

Method	bbt_s01e01	bbt_s01e02	bbt_s01e03	bbt_s01e04	bbt_s01e05	bbt_s01e06
VGG-Face*	0.91	0.85	0.83	0.54	0.65	0.46
Siamese*	0.94	0.95	0.87	0.74	0.70	0.70
Triplet*	0.94	0.95	0.92	0.74	0.68	0.70
SymTriplet*	0.94	0.95	0.92	0.78	0.85	0.75
IdOL	**0.99**	**0.99**	**0.94**	**0.94**	**0.99**	**0.97**

* Values reported from [100]

show the plot of a *Baseline* approach that performs 1-Nearest Neighbor matching (i.e. forward) with thresholding of the distance value. For this baseline a maximal number of memory elements for each identity is used and elements are removed randomly when the identity budget is met (the best result of experiments with different distance thresholds and max memory elements is reported). As can be noticed, the number of identities estimated by the *Baseline* method increases with time while our approach closely follows the number of identities of the ground-truth. The results confirm the effectiveness of the IdOL method to learn an unknown number of identities.

6.3. Scaling and Asymptotic Stability

As the number of identities grow, online learning of identities becomes more challenging. In order to verify the behavior of the IdOL method in this case, we concatenated the six sequences of the *Big Bang Theory* dataset to form a single longer sequence of about two hours, and manually annotated all the subjects up to a total number of 99 different identities.

Fig. 10a and Fig. 10b respectively show the number of features in memory at each time instant and the number of ground-truth identities that showed up until then. It can be noticed that the number of features in memory follows the same trend as the number of identities. Fig. 10c shows the MOTA plot on the whole sequence (dark bold line) and the MOTA plots calculated for each video segment (colored lines). As the observations are accumulated, all the identities are progressively learned and MOTA keeps stable despite that the number of identities has been increased of one order of magnitude with respect to the individual sequences. The MOTA fluctuations due to the insufficient information that were observed at the beginning of each sequence are no more present and the MOTA score is higher than the MOTA of the individual video segments in most cases. In Episode 4 (*e04*), a high number of new identities joins and MOTA is temporarily lower due to the increased complexity of learning.

6.4. Ablation Study

To demonstrate the effectiveness of the solution we conduct an ablation study in comparison with a baseline in which identities are never explicitly deleted when they exit the field of view. The Baseline uses memory and Reverse Nearest Neighbour as IdOL,and randomly forgets features when a memory budget is met (three different budget values $|\mathcal{M}|$ are evaluated: 100, 500, 1000 and 1500 elements). The Baseline does not use the learning mechanism of Eq.3. Since features are randomly deleted in the baseline to maintain the memory budget, performance values are averaged over 100 tests.

We also consider the effect of different representations, using the features discussed in section 5. The *QMUL multi-face dataset* is used for the ablation as it provides specific types of appearance variation of the four subjects. Effects of viewing conditions are considered by evaluating performance for the three videos of the dataset, separately and concatenated in different

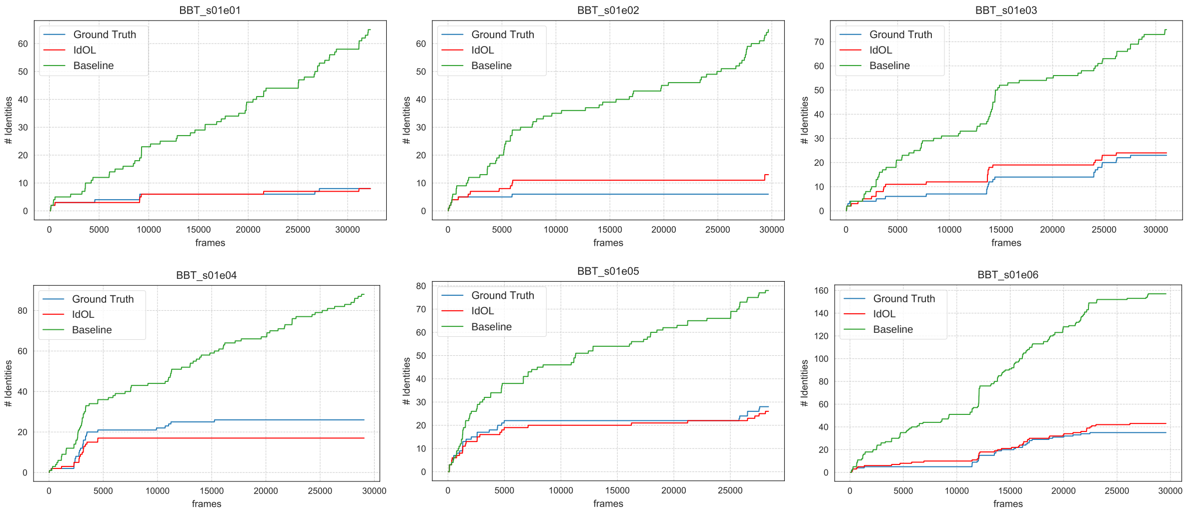
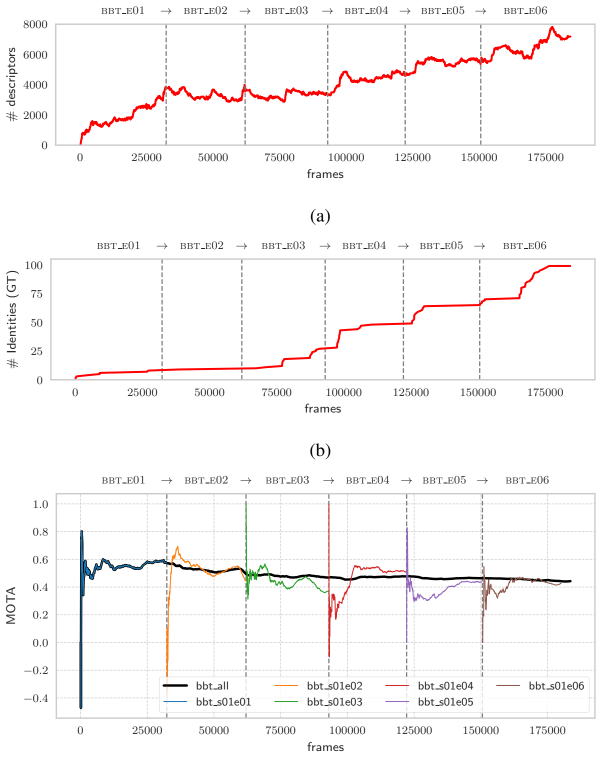


Figure 9: Number of identities learned at each frame for the videos in the *Big Bang Theory* dataset: Ground-truth, IdOL and Baseline.



(a)

(b)

(c)

Figure 10: *Big Bang Theory* 2 hours sequence (184298 frames) obtained by linking the videos of the 6 Episodes: *(a)* IdOL number of features in memory; *(b)* ground-truth number of identities (cumulative); *(c)* IdOL MOTA computed at each frame for the whole sequence (black bold line) and each Episode (colored lines).

order. This latter allows to evaluate cumulative learning with curricula of different complexity. The average performance values are also reported in both cases. Results for the three videos of the dataset, separately evaluated are shown in Tab. 10a. For the Baseline, the performance increases with the size of the memory and with the quality of the feature representation. The IdOL method scores the best results overall with a substantially lower number of features stored in the memory. The difference is particularly evident in the case of the VGGFace2/SeNet-128 feature representation. In this case the number of features in memory is one order of magnitude smaller than with VGGFace/VGG16-4096. VGGFace2/SeNet-128 and VGGFace2/ResNet score the best performance. This depends on the CNN architecture used, the VGGFace2 training dataset and on the lower feature dimension. Since the IdOL matching procedure is based on distance ratio, under reasonable assumptions of feature distribution, the distance ratio between the nearest and the farthest neighbors to a given features in high dimensional space is almost 1 [10, 1] so making the criterion less discriminative than it is in a lower dimensional space.

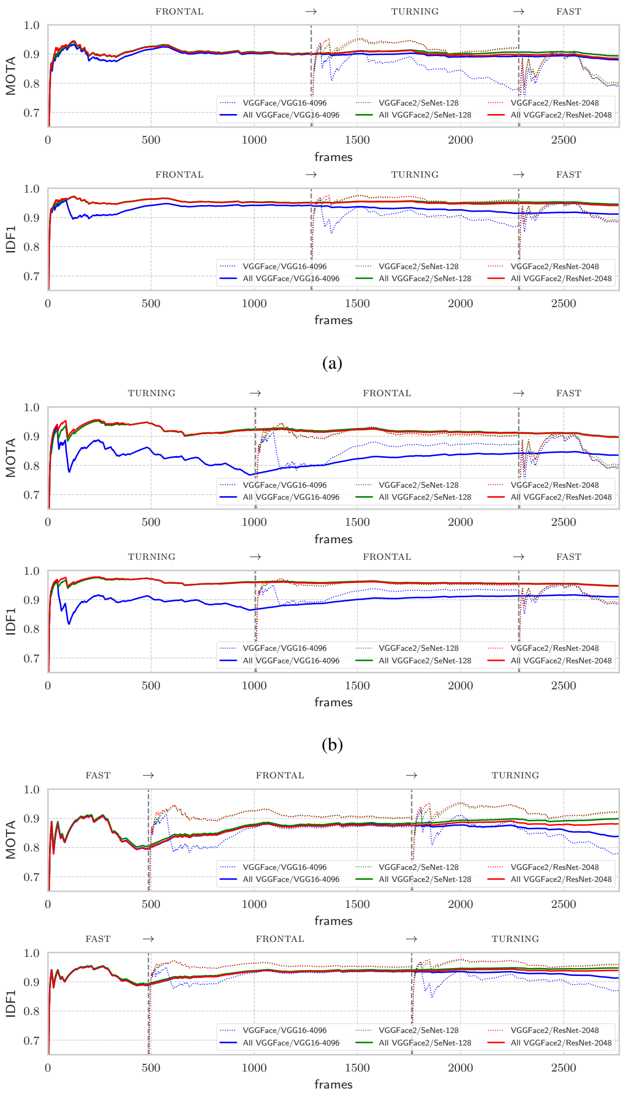
Results for learning curricula of different complexity obtained by concatenating the three videos of the *QMUL multi-face dataset* (Fast → Frontal → Turning, Frontal → Turning → Fast and Turning → Frontal → Fast), are shown in Tab. 10b and Fig. 11. Results in Tab. 10b confirm the conclusions derived for Tab. 10a. Fig. 11 compares IDF1[4] and MOTA of the

[4] IDF1 is the ratio of correctly identified detections over the aver-

three curricula with IDF1 and MOTA of the individual sequences (bold and dotted lines, respectively). It clearly appears the increase of both MOTA and IDF1 by cumulative learning from the previous sequences. Features learned from the VGGFace2 dataset (green and red) show better performance than features trained with the VGGFace one (blue). With gradual increasing of complexity of the curriculum Frontal → Turning → Fast in Fig. 11a, IdOL is however able to improve the performance of the Turning sequence also on the weakest VGGFace/VGG16-4096 representation. On the contrary, as shown in Fig. 11b, starting with Turning does not improve the performance on the Frontal. As shown in Fig. 11c, similar behaviour of Fig. 11a is observed when the Fast video sequence is moved to the the beginning.

Finally Fig. 12 shows a direct comparison between the Baseline with $|\mathcal{M}| = 500$ and IdOL both using the VGGFace2/SeNet-128 feature representation. IdOL and the Baseline concluded the processing with 392 and 500 elements, respectively showing that with a comparable number of features, our method does not change substantially MOTA and IDF1 over time.

The performance evaluations we reported use parameters values that correspond to typical conditions observed in most real sequences. We use the following values: $\bar{\rho} = 1/1.6 = 0.625$. This setting has the following interpretation: in order to asses the match, the second nearest neighbour must be 1.6 times more distant than the first nearest neighbor. The eligibility threshold $\bar{e}$, used to delete a feature, is set according to the length of the processed video as the length have a direct impact on feature diversity on the memory module. We set: $\bar{e} = 0.5$ for all the videos in the *Music* dataset and *QMUL* multiface-dataset, $\bar{e} = 0.9$ for each single *BBT* video, $\bar{e} = 0.99$ for the concatenated *BBT*. The value α in Eq. 3 is set to $\alpha = 0.001$ for all the datasets.



(a)

(b)

(c)

Figure 11: Effects of cumulative learning: performance plots of MOTA and IDF1 over the three curricula (bold) and the individual sequences (dotted) of the *QMUL* multi-face dataset. Plots are shown for different feature representations: VGGface/VGG16 (blue), VGGFace2/SeNet (green) VGGFace2/ResNet (red). Learning performance shows some dependency on initial learning: in the end, learning is clearly improved with curricula starting with the Frontal and Turning sequence; does not improve when curricula starts with the Fast. Effects of the quality of features are clearly evident.

6.5. *Computational Issues*

An evident drawback of ReNN is that in practice a huge number of features (those accumulated in the memory module) is matched against a relatively small set of features (those observed in the current image). Due to that, deciding matching by sorting is prohibitively expensive and tree-based data structures [63] cannot be used effectively since the number of the prototypes in the image is orders of magnitude smaller than the number of prototypes in the memory. However, this drawback can be easily solved by computing the first and second minimum distance through consecutive applications of linear search with GPU implementation. In this way, we can exploit the very efficient CUDA matrix multiplication kernel for the computation of the squared distance matrix and GPU parallelism [49]. Fig. 13 shows scaling of performance on Intel

age number of ground-truth and computed detections [76].

Table 10: Ablation study: Baselines with fixed memory budget (100, 500, 1000, 1500 features) versus the IdOL method. MOTA IDF1 and IDS are evaluated for different feature representations (VGGFace/VGG16, VGGFace2/SeNet, VGGFace2/ResNet). (a) performance values over the three sequences of the *QMUL multi-face dataset*; (b) Performance values evaluated over the three curricula. Average performance values of the three cases are also presented.

Feature Representation		Fast				Frontal				Turning				Average		
Dataset/Architecture	Dim	MOTA	IDF1	IDS	$\lvert\mathcal{M}\rvert$	MOTA	IDF1	IDS	$\lvert\mathcal{M}\rvert$	MOTA	IDF1	IDS	$\lvert\mathcal{M}\rvert$	MOTA	IDF1	IDS
BASELINES (Tiny Face Detector)																
VGGFace/VGG16	4096	67.526	0.683	10	100	66.261	0.727	16	100	41.502	0.509	11	100	58.430	0.640	12.3
VGGFace2/SeNet	128	76.005	0.750	6	100	63.996	0.683	22	100	43.165	0.513	14	100	61.055	0.649	14.0
VGGFace2/ResNet	2048	76.401	0.745	9	100	64.310	0.697	18	100	50.227	0.618	14	100	63.646	0.687	13.7
VGGFace/VGG16	4096	77.572	0.709	13	500	84.046	0.849	26	500	70.758	0.729	36	500	77.459	0.762	25.0
VGGFace2/SeNet	128	77.258	0.759	11	500	83.073	0.850	38	500	74.979	0.775	32	500	78.436	0.795	27.0
VGGFace2/ResNet	2048	76.518	0.751	17	500	83.987	0.854	38	500	75.839	0.779	43	500	78.781	0.794	32.7
VGGFace/VGG16	4096	78.053	0.706	12	1000	84.670	0.859	25	1000	72.358	0.737	36	1000	78.361	0.767	24.3
VGGFace2/SeNet	128	77.328	0.765	14	1000	86.261	0.901	54	1000	87.659	0.878	68	1000	83.749	0.848	45.3
VGGFace2/ResNet	2048	76.466	0.753	22	1000	86.375	0.887	49	1000	86.892	0.847	66	1000	83.244	0.829	45.7
VGGFace/VGG16	4096	77.989	0.706	12	1123	87.925	0.881	28	1500	73.752	0.743	35	1500	79.889	0.777	25.0
VGGFace2/SeNet	128	77.328	0.765	14	1123	86.292	0.907	55	1500	87.859	0.882	66	1500	83.826	0.851	45.0
VGGFace2/ResNet	2048	76.466	0.753	22	1123	86.389	0.891	54	1500	88.575	0.859	68	1500	83.810	0.834	48.0
IdOL (Tiny Face Detector)																
VGGFace/VGG16	4096	79.138	0.886	4	785	87.152	0.933	6	1501	78.079	0.871	20	1705	81.457	0.897	10.0
VGGFace2/SeNet	128	**80.259**	**0.892**	**0**	203	**90.104**	**0.951**	**2**	321	**92.323**	**0.960**	**5**	323	**87.562**	**0.934**	**2.3**
VGGFace2/ResNet	2048	79.483	0.888	**0**	305	90.007	0.950	**2**	464	92.032	0.959	**5**	576	87.174	0.932	**2.3**

(a)

Feature Representation		Fast→Frontal →Turning				Frontal →Turning →Fast				Turning →Frontal →Fast				Average		
Dataset/Architecture	Dim	MOTA	IDF1	IDS	$\lvert\mathcal{M}\rvert$	MOTA	IDF1	IDS	$\lvert\mathcal{M}\rvert$	MOTA	IDF1	IDS	$\lvert\mathcal{M}\rvert$	MOTA	IDF1	IDS
BASELINES (Tiny Face Detector)																
VGGFace/VGG16	4096	29.218	0.402	15	100	42.929	0.504	23	100	24.886	0.322	15	100	32.344	0.409	17.7
VGGFace2/SeNet	128	30.981	0.397	16	100	43.926	0.493	32	100	31.822	0.381	24	100	35.576	0.424	24.0
VGGFace2/ResNet	2048	19.069	0.265	14	100	44.947	0.510	27	100	39.055	0.483	29	100	34.357	0.419	23.3
VGGFace/VGG16	4096	57.070	0.598	50	500	66.803	0.719	53	500	60.892	0.653	64	500	61.588	0.657	55.7
VGGFace2/SeNet	128	45.376	0.485	47	500	67.436	0.729	63	500	63.587	0.662	61	500	58.800	0.625	57.0
VGGFace2/ResNet	2048	37.857	0.418	52	500	69.080	0.739	73	500	62.747	0.680	74	500	56.561	0.612	66.3
VGGFace/VGG16	4096	47.313	0.496	46	1000	69.795	0.741	61	1000	61.828	0.665	64	1000	59.645	0.634	57.0
VGGFace2/SeNet	128	45.610	0.532	83	1000	80.956	0.835	132	1000	83.916	0.842	152	1000	70.161	0.736	122.3
VGGFace2/ResNet	2048	40.397	0.469	75	1000	79.600	0.804	115	1000	76.103	0.775	130	1000	65.367	0.683	106.7
VGGFace/VGG16	4096	49.792	0.522	52	1500	77.317	0.781	86	1500	61.483	0.660	68	1500	62.864	0.654	68.7
VGGFace2/SeNet	128	48.073	0.557	90	1500	85.377	0.876	146	1500	84.866	0.855	160	1500	72.772	0.763	132.0
VGGFace2/ResNet	2048	46.373	0.540	93	1500	85.626	0.843	144	1500	85.096	0.838	167	1500	72.365	0.740	134.7
IdOL (Tiny Face Detector)																
VGGFace/VGG16	4096	83.828	0.913	26	2436	88.049	0.912	78	2376	83.473	0.909	26	2465	85.117	0.911	43.3
VGGFace2/SeNet	128	**89.850**	**0.947**	14	381	**89.380**	**0.946**	**7**	367	**89.792**	**0.948**	**6**	392	**89.674**	**0.947**	9.0
VGGFace2/ResNet	2048	88.118	0.939	**12**	694	88.531	0.941	8	713	89.632	0.947	**6**	660	88.760	0.942	**8.7**

(b)

i7-2600K 3.40GHz and Nvidia Geforce Titan X as a function of the number of features in the memory module. It is evident that, using GPU, performance keeps almost constant as the number of features in memory increases. With such hardware support, the full system operates on-line at 8 frames per second with 800x600 video frame resolution.

7. Conclusions

In this paper, we have presented a novel solution for cumulative learning face identities in unconstrained video streams based on face appearance. We discussed the substantial differences between our learning setting (referred as MOCAL, Multiple Object Cumulative Adaptation Learning), Multiple Object Tracking and Continual Learning when applied to video streams. Our solution updates a representative dataset and use it as a memory of all the past visual information observed so far. This strategy enables the accumulation and preservation of essential knowledge and at the same time allows to handle the non-stationarity of the data stream. We have shown that the proposed method is theoretically sound, asymptotically stable and operates on-line. Its effectiveness has been demonstrated in comparison with Multiple Object Tracking methods over pub-

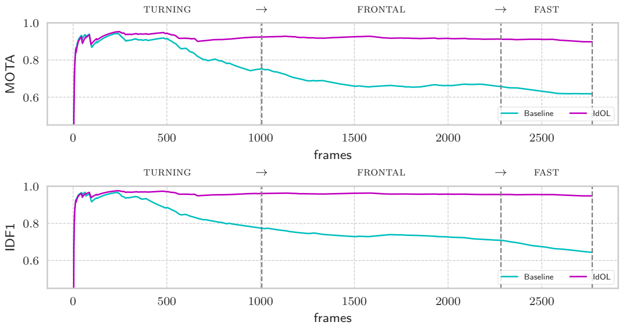


Figure 12: Effects of cumulative learning: performance plots of MOTA and IDF1 over the Turning → Frontal → Fast curriculum for the Baseline with 500 features memory budget (light green) versus the IdOL method (392 features). In the IdOL method MOTA and IDF1 soon reach the maximum value and keep stable over time, despite of the lower number of features in memory.

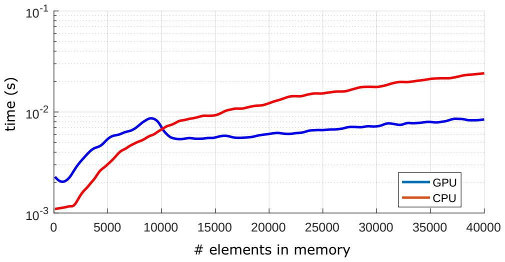


Figure 13: Average processing time of Reverse Nearest Neighbor as a function of the number of features in memory: on Intel i7-2600K 3.40GHz and NVIDIA Geforce Titan X GPU.

lic datasets. We showed that the method is capable of cumulative learning effectively over long unconstrained video sequences. The method can be applied in principle to any other context for which a detector/feature combination is available (i.e. vehicle, person, boat, traffic sign).

Acknowledgment

This research was supported by the project Ideha "Innovazioni per L'elaborazione dei Dati nel Settore del Patrimonio Culturale" (MIUR - PON ARS01-00421) and Leonardo SpA, Italy.

References

[1] Aggarwal, C.C., Hinneburg, A., Keim, D.A., 2001. On the surprising behavior of distance metrics in high dimensional space, in: International conference on database theory, Springer. pp. 420–434.

[2] Aljundi, R., Kelchtermans, K., Tuytelaars, T., 2019. Task-free continual learning, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 11254–11263.

[3] Ayazoglu, M., Sznaier, M., Camps, O.I., 2012. Fast algorithms for structured robust principal component analysis, in: Computer Vision and Pattern Recognition (CVPR), 2012 IEEE Conference on, IEEE. pp. 1704–1711.

[4] Banach, S., 1922. Sur les opérations dans les ensembles abstraits et leur application aux équations intégrales. Fundamenta Mathematicae 3, 133–181.

[5] Bäuml, M., Tapaswi, M., Stiefelhagen, R., 2013. Semi-supervised Learning with Constraints for Person Identification in Multimedia Data, in: IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp. 3602–3609.

[6] Bendale, A., Boult, T., 2015. Towards open world recognition, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 1893–1902.

[7] Bergmann, P., Meinhardt, T., Leal-Taixe, L., 2019. Tracking without bells and whistles, in: The IEEE International Conference on Computer Vision (ICCV).

[8] Bertinetto, L., Valmadre, J., Henriques, J.F., Vedaldi, A., Torr, P.H., 2016. Fully-convolutional siamese networks for object tracking, in: European conference on computer vision, Springer. pp. 850–865.

[9] Bewley, A., Ge, Z., Ott, L., Ramos, F., Upcroft, B., 2016. Simple online and realtime tracking, in: 2016 IEEE International Conference on Image Processing (ICIP), IEEE. pp. 3464–3468.

[10] Beyer, K., Goldstein, J., Ramakrishnan, R., Shaft, U., 1999. When is "nearest neighbor" meaningful?, in: International conference on database theory, Springer. pp. 217–235.

[11] Bojanowski, P., Bach, F., Laptev, I., Ponce, J., Schmid, C., Sivic, J., 2013. Finding actors and actions in movies, in: Proceedings of the IEEE international conference on computer vision, pp. 2280–2287.

[12] Camps, O., Cucchiara, R., Del Bimbo, A., Matas, J., Pernici, F., Sclaroff, S., 2014. Long term detection and tracking workshop., in: Conjunction With The IEEE Conference on Computer Vision and Pattern Recognition (CVPR) Workshops (LTDT2014).

[13] Cao, Q., Shen, L., Xie, W., Parkhi, O.M., Zisserman, A., 2018. Vggface2: A dataset for recognising faces across pose and age, in: International Conference on Automatic Face and Gesture Recognition.

[14] Cao, Z., Simon, T., Wei, S.E., Sheikh, Y., 2017. Realtime multi-person 2d pose estimation using part affinity fields, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 7291–7299.

[15] Chen, Z., Liu, B., 2018. Lifelong machine learning. Synthesis Lectures on Artificial Intelligence and Machine Learning 12, 1–207.

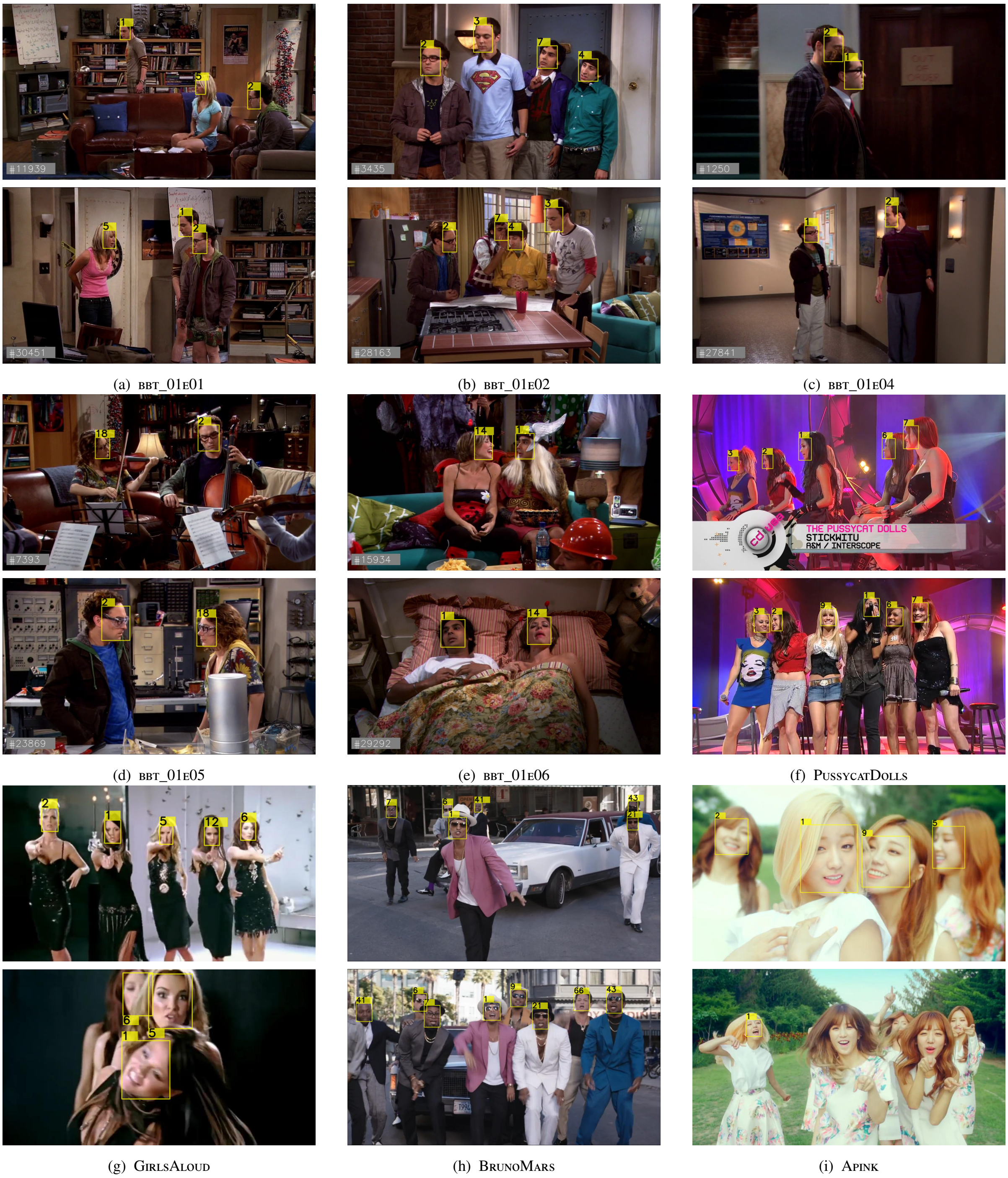
[16] Chu, Q., Ouyang, W., Li, H., Wang, X., Liu, B., Yu, N., 2017. Online multi-object tracking using cnn-based single object tracker with spatial-temporal attention mechanism, in: Proceedings of the IEEE International Conference on Computer Vision, pp. 4836–4845.

[17] Ciaparrone, G., Sánchez, F.L., Tabik, S., Troiano, L., Tagliaferri, R., Herrera, F., 2019. Deep learning in video multi-object tracking: A survey. Neurocomputing .

[18] Cover, T., Hart, P., 1967. Nearest neighbor pattern classification. IEEE transactions on information theory 13, 21–27.

[19] Danelljan, M., Bhat, G., Shahbaz Khan, F., Felsberg, M., 2017. Eco: Efficient convolution operators for tracking, in: Proceedings of the IEEE conference on computer vision and pattern recognition, pp. 6638–6646.

[20] Del Bimbo, A., Lisanti, G., Masi, I., Pernici, F., 2010. Device-tagged feature-based localization and mapping of wide areas with a ptz camera, in: 2010 IEEE Computer Society Confer-



(a) bbt_01e01 (b) bbt_01e02 (c) bbt_01e04

(d) bbt_01e05 (e) bbt_01e06 (f) PussycatDolls

(g) GirlsAloud (h) BrunoMars (i) Apink

Figure 14: Sample pairs of frames from the *Music* and *Big Bang Theory* datasets showing takes of the same individuals in different conditions and the identity label assigned by the IdOL method

ence on Computer Vision and Pattern Recognition-Workshops, IEEE. pp. 39–44.

[21] Deng, J., Dong, W., Socher, R., Li, L.J., Li, K., Fei-Fei, L., 2009. Imagenet: A large-scale hierarchical image database, in: Computer Vision and Pattern Recognition, 2009. CVPR 2009. IEEE Conference on, IEEE. pp. 248–255.

[22] Dicle, C., Camps, O.I., Sznaier, M., 2013. The way they move: Tracking multiple targets with similar appearance, in: Proceedings of the IEEE International Conference on Computer Vision, pp. 2304–2311.

[23] Dinh, T.B., Vo, N., Medioni, G., 2011. Context tracker: Exploring supporters and distracters in unconstrained environments, in: CVPR.

[24] Du, M., Chellappa, R., 2015. Face association for videos using conditional random fields and max-margin markov networks. IEEE transactions on pattern analysis and machine intelligence 38, 1762–1773.

[25] Gepperth, A., Hammer, B., 2016. Incremental learning algorithms and applications .

[26] Graves, A., Wayne, G., Danihelka, I., 2014. Neural turing machines. arXiv preprint arXiv:1410.5401 .

[27] Graves, A., Wayne, G., Reynolds, M., Harley, T., Danihelka, I., Grabska-Barwińska, A., Colmenarejo, S.G., Grefenstette, E., Ramalho, T., Agapiou, J., et al., 2016. Hybrid computing using a neural network with dynamic external memory. Nature 538, 471–476.

[28] He, K., Zhang, X., Ren, S., Sun, J., 2016. Deep residual learning for image recognition, in: Proceedings of the IEEE conference on computer vision and pattern recognition, pp. 770–778.

[29] Held, D., Thrun, S., Savarese, S., 2016. Learning to track at 100 fps with deep regression networks, in: European Conference on Computer Vision, Springer. pp. 749–765.

[30] Hong, Z., Chen, Z., Wang, C., Mei, X., Prokhorov, D., Tao, D., 2015. Multi-store tracker (muster): A cognitive psychology inspired approach to object tracking.

[31] Hu, J., Shen, L., Sun, G., 2018. Squeeze-and-excitation networks, in: Proceedings of the IEEE conference on computer vision and pattern recognition, pp. 7132–7141.

[32] Hu, P., Ramanan, D., 2017. Finding tiny faces, in: The IEEE Conference on Computer Vision and Pattern Recognition (CVPR).

[33] Hua, Y., Alahari, K., Schmid, C., 2014. Occlusion and motion reasoning for long-term tracking, in: Computer Vision–ECCV 2014. Springer, pp. 172–187.

[34] Insafutdinov, E., Pishchulin, L., Andres, B., Andriluka, M., Schiele, B., 2016. Deepercut: A deeper, stronger, and faster multi-person pose estimation model, in: European Conference on Computer Vision, Springer. pp. 34–50.

[35] Jepson, A.D., Fleet, D.J., El-Maraghi, T.F., 2003. Robust online appearance models for visual tracking. IEEE transactions on pattern analysis and machine intelligence 25, 1296–1311.

[36] Kaiser, L., Nachum, O., Roy, A., Bengio, S., 2017. Learning to remember rare events. ICLR .

[37] Kalal, Z., Matas, J., Mikolajczyk, K., 2010. P-n learning: Bootstrapping binary classifiers by structural constraints, in: CVPR.

[38] Kalal, Z., Mikolajczyk, K., Matas, J., 2012. Tracking-learning-detection. IEEE transactions on pattern analysis and machine intelligence 34, 1409–1422.

[39] Kim, C., Li, F., Ciptadi, A., Rehg, J.M., 2015. Multiple hypothesis tracking revisited, in: Proceedings of the IEEE International Conference on Computer Vision, pp. 4696–4704.

[40] Klare, B.F., Klein, B., Taborsky, E., Blanton, A., Cheney, J., Allen, K., Grother, P., Mah, A., Jain, A.K., 2015. Pushing the frontiers of unconstrained face detection and recognition: Iarpa janus benchmark a, in: Proceedings of the IEEE conference on computer vision and pattern recognition, pp. 1931–1939.

[41] Korn, F., Muthukrishnan, S., 2000. Influence sets based on reverse nearest neighbor queries, in: Proceedings of the 2000 ACM SIGMOD International Conference on Management of Data, ACM, New York, NY, USA. pp. 201–212.

[42] Kristan, M., Matas, J., Leonardis, A., Felsberg, M., Pflugfelder, R., Kamarainen, J.K., Cehovin Zajc, L., Drbohlav, O., Lukezic, A., Berg, A., et al., 2019. The seventh visual object tracking vot2019 challenge results, in: Proceedings of the IEEE International Conference on Computer Vision Workshops, pp. 0–0.

[43] Kristan, M., Matas, J., Leonardis, A., Vojir, T., Pflugfelder, R., Fernandez, G., Nebehay, G., Porikli, F., Čehovin, L., 2016. A novel performance evaluation methodology for single-target trackers. IEEE Transactions on Pattern Analysis and Machine Intelligence 38, 2137–2155.

[44] Krizhevsky, A., Sutskever, I., Hinton, G.E., 2012. Imagenet classification with deep convolutional neural networks., in: NIPS, p. 4.

[45] Kumaran, D., Hassabis, D., McClelland, J.L., 2016. What learning systems do intelligent agents need? complementary learning systems theory updated. Trends in cognitive sciences 20, 512–534.

[46] Leal-Taixé, L., Milan, A., Schindler, K., Cremers, D., Reid, I., Roth, S., 2017. Tracking the trackers: An analysis of the state of the art in multiple object tracking. arXiv preprint arXiv:1704.02781 .

[47] Lee, K.W., Sankaran, N., Setlur, S., Napp, N., Govindaraju, V., 2018. Wardrobe model for long term re-identification and appearance prediction, in: 2018 15th IEEE International Conference on Advanced Video and Signal Based Surveillance (AVSS), IEEE. pp. 1–6.

[48] Li, B., Yan, J., Wu, W., Zhu, Z., Hu, X., 2018. High performance visual tracking with siamese region proposal network, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 8971–8980.

[49] Li, S., Amenta, N., 2015. Brute-force k-nearest neighbors search on the gpu, in: International Conference on Similarity Search and Applications, Springer. pp. 259–270.

[50] Li, Z., Hoiem, D., 2016. Learning without forgetting, in: European Conference on Computer Vision, Springer. pp. 614–629.

[51] Lin, T.Y., Maire, M., Belongie, S., Hays, J., Perona, P., Ramanan, D., Dollár, P., Zitnick, C.L., 2014. Microsoft coco: Common objects in context, in: European Conference on Computer Vision, Springer. pp. 740–755.

[52] Lisanti, G., Masi, I., Pernici, F., Del Bimbo, A., 2016. Continuous localization and mapping of a pan–tilt–zoom camera for wide area tracking. Machine Vision and Applications 27, 1071–1085.

[53] Liu, W., Anguelov, D., Erhan, D., Szegedy, C., Reed, S., Fu, C.Y., Berg, A.C., 2016. Ssd: Single shot multibox detector, in: European conference on computer vision, Springer. pp. 21–37.

[54] Losing, V., Hammer, B., Wersing, H., 2018. Incremental online learning: A review and comparison of state of the art algorithms. Neurocomputing 275, 1261–1274.

[55] Lowe, D.G., 2004. Distinctive image features from scale-invariant keypoints. IJCV 60, 91–110.

[56] Lukežič, A., Zajc, L.Č., Vojíř, T., Matas, J., Kristan, M., 2019. Performance evaluation methodology for long-term visual object tracking. arXiv preprint arXiv:1906.08675 .

[57] Luo, W., Zhao, X., Kim, T., 2017. Multiple object tracking: A review. CoRR abs/1409.7618.

[58] Maggio, E., Piccardo, E., Regazzoni, C., Cavallaro, A., 2007. Particle phd filtering for multi-target visual tracking, in: 2007

IEEE International Conference on Acoustics, Speech and Signal Processing-ICASSP'07, IEEE. pp. I–1101.

[59] Mathias, M., Benenson, R., Pedersoli, M., Van Gool, L., 2014. Face detection without bells and whistles, in: European Conference on Computer Vision, Springer. pp. 720–735.

[60] Milan, A., Leal-Taixé, L., Reid, I., Roth, S., Schindler, K., 2016. Mot16: A benchmark for multi-object tracking. arXiv preprint arXiv:1603.00831 .

[61] Mnih, V., Kavukcuoglu, K., Silver, D., Rusu, A.A., Veness, J., Bellemare, M.G., Graves, A., Riedmiller, M., Fidjeland, A.K., Ostrovski, G., et al., 2015. Human-level control through deep reinforcement learning. Nature 518, 529–533.

[62] Moudgil, A., Gandhi, V., 2017. Long-term visual object tracking benchmark. arXiv preprint arXiv:1712.01358 .

[63] Muja, M., Lowe, D.G., 2009. Fast approximate nearest neighbors with automatic algorithm configuration, in: International Conference on Computer Vision Theory and Application VISSAPP'09), INSTICC Press. pp. 331–340.

[64] Nam, H., Han, B., 2016. Learning multi-domain convolutional neural networks for visual tracking, in: Proceedings of the IEEE conference on computer vision and pattern recognition, pp. 4293–4302.

[65] Parisi, G.I., Kemker, R., Part, J.L., Kanan, C., Wermter, S., 2019. Continual lifelong learning with neural networks: A review. Neural Networks .

[66] Parkhi, O.M., Vedaldi, A., Zisserman, A., 2015a. Deep face recognition, in: British Machine Vision Conference.

[67] Parkhi, O.M., Vedaldi, A., Zisserman, A., 2015b. Deep face recognition, in: Proceedings of the British Machine Vision Conference 2015, BMVC 2015, Swansea, UK, September 7-10, 2015, pp. 41.1–41.12.

[68] Pernici, F., 2012. Facehugger: The alien tracker applied to faces, in: Computer Vision–ECCV 2012. Workshops and Demonstrations, Springer. pp. 597–601.

[69] Pernici, F., Bartoli, F., Bruni, M., Bimbo, A.D., 2018. Memory based online learning of deep representations from video streams, in: IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2018 Salt Lake City, Utah, USA, June 18-22.

[70] Pernici, F., Del Bimbo, A., 2013. Object tracking by oversampling local features. IEEE transactions on pattern analysis and machine intelligence 36, 2538–2551.

[71] Pernici, F., Del Bimbo, A., 2017. Unsupervised incremental learning of deep descriptors from video streams, in: 2017 IEEE International Conference on Multimedia & Expo Workshops (ICMEW), IEEE. pp. 477–482.

[72] Pirsiavash, H., Ramanan, D., Fowlkes, C.C., 2011. Globally-optimal greedy algorithms for tracking a variable number of objects, in: CVPR 2011, IEEE. pp. 1201–1208.

[73] Rebuffi, S.A., Kolesnikov, A., Sperl, G., Lampert, C.H., 2017. icarl: Incremental classifier and representation learning, in: The IEEE Conference on Computer Vision and Pattern Recognition (CVPR).

[74] Ren, S., He, K., Girshick, R., Sun, J., 2015. Faster r-cnn: Towards real-time object detection with region proposal networks, in: Advances in neural information processing systems, pp. 91–99.

[75] Riemer, M., Cases, I., Ajemian, R., Liu, M., Rish, I., Tu, Y., Tesauro, G., 2019. Learning to learn without forgetting by maximizing transfer and minimizing interference. International Conference on Learning Representations (ICLR) .

[76] Ristani, E., Solera, F., Zou, R., Cucchiara, R., Tomasi, C., 2016. Performance measures and a data set for multi-target, multi-camera tracking, in: European Conference on Computer Vision, Springer. pp. 17–35.

[77] Ristani, E., Tomasi, C., 2018. Features for multi-target multi-camera tracking and re-identification, in: Proceedings of the IEEE conference on computer vision and pattern recognition, pp. 6036–6046.

[78] Rusu, A.A., Rabinowitz, N.C., Desjardins, G., Soyer, H., Kirkpatrick, J., Kavukcuoglu, K., Pascanu, R., Hadsell, R., 2016. Progressive neural networks. arXiv preprint arXiv:1606.04671 .

[79] Sadeghian, A., Alahi, A., Savarese, S., 2017. Tracking the untrackable: Learning to track multiple cues with long-term dependencies. arXiv preprint arXiv:1701.01909 .

[80] Salganicoff, M., 1993. Density-adaptive learning and forgetting, in: Proceedings of the Tenth International Conference on International Conference on Machine Learning, Morgan Kaufmann Publishers Inc.. pp. 276–283.

[81] Santoro, A., Bartunov, S., Botvinick, M., Wierstra, D., Lillicrap, T., 2016a. Meta-learning with memory-augmented neural networks, in: International conference on machine learning, pp. 1842–1850.

[82] Santoro, A., Bartunov, S., Botvinick, M., Wierstra, D., Lillicrap, T., 2016b. One-shot learning with memory-augmented neural networks. arXiv preprint arXiv:1605.06065 .

[83] Schaul, T., Quan, J., Antonoglou, I., Silver, D., 2016. Prioritized experience replay, in: International Conference on Learning Representations, Puerto Rico.

[84] Schroff, F., Kalenichenko, D., Philbin, J., 2015. Facenet: A unified embedding for face recognition and clustering. CoRR abs/1503.03832. URL: http://arxiv.org/abs/1503.03832, arXiv:1503.03832.

[85] Sun, C., Shrivastava, A., Singh, S., Gupta, A., 2017. Revisiting unreasonable effectiveness of data in deep learning era, in: Proceedings of the IEEE international conference on computer vision, pp. 843–852.

[86] Tao, R., Gavves, E., Smeulders, A.W., 2016. Siamese instance search for tracking, in: Proceedings of the IEEE conference on computer vision and pattern recognition, pp. 1420–1429.

[87] Tapaswi, M., Parkhi, O.M., Rahtu, E., Sommerlade, E., Stiefelhagen, R., Zisserman, A., 2014. Total cluster: A person agnostic clustering method for broadcast videos, in: Proceedings of the 2014 Indian Conference on Computer Vision Graphics and Image Processing, ACM. p. 7.

[88] Thórisson, K.R., Bieger, J., Li, X., Wang, P., 2019. Cumulative learning, in: International Conference on Artificial General Intelligence, Springer. pp. 198–208.

[89] Valmadre, J., Bertinetto, L., Henriques, J.F., Tao, R., Vedaldi, A., Smeulders, A.W., Torr, P.H., Gavves, E., 2018. Long-term tracking in the wild: A benchmark, in: Proceedings of the European Conference on Computer Vision (ECCV), pp. 670–685.

[90] Vinyals, O., Blundell, C., Lillicrap, T., Wierstra, D., et al., 2016. Matching networks for one shot learning, in: Advances in neural information processing systems, pp. 3630–3638.

[91] Wang, G., Lai, J., Huang, P., Xie, X., 2019. Spatial-temporal person re-identification, in: Proceedings of the AAAI Conference on Artificial Intelligence, pp. 8933–8940.

[92] Ward, J.H., 1963. Hierarchical grouping to optimize an objective function. Journal of the American Statistical Association 58, 236–244.

[93] Wojke, N., Bewley, A., Paulus, D., 2017. Simple online and realtime tracking with a deep association metric, in: 2017 IEEE International Conference on Image Processing (ICIP), pp. 3645–3649.

[94] Wu, B., Zhang, Y., Hu, B.G., Ji, Q., 2013. Constrained clustering and its application to face clustering in videos, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 3507–3514.

[95] Xiang, Y., Alahi, A., Savarese, S., 2015. Learning to track: Online multi-object tracking by decision making, in: Proceedings of the IEEE international conference on computer vision, pp. 4705–4713.

[96] Xiao, S., Tan, M., Xu, D., 2014. Weighted block-sparse low rank representation for face clustering in videos, in: European Conference on Computer Vision, Springer. pp. 123–138.

[97] Yang, B., Nevatia, R., 2012. An online learned crf model for multi-target tracking, in: 2012 IEEE Conference on Computer Vision and Pattern Recognition, IEEE. pp. 2034–2041.

[98] Yoon, J.H., Lee, C.R., Yang, M.H., Yoon, K.J., 2019. Structural constraint data association for online multi-object tracking. International Journal of Computer Vision 127, 1–21.

[99] Yu, F., Li, W., Li, Q., Liu, Y., Shi, X., Yan, J., 2016. Poi: Multiple object tracking with high performance detection and appearance feature, in: European Conference on Computer Vision, Springer. pp. 36–42.

[100] Zhang, S., Gong, Y., Huang, J.B., Lim, J., Wang, J., Ahuja, N., Yang, M.H., 2016. Tracking persons-of-interest via adaptive discriminative features, in: European Conference on Computer Vision, Springer. pp. 415–433.

[101] Zheng, L., Zhang, H., Sun, S., Chandraker, M., Yang, Y., Tian, Q., 2017. Person re-identification in the wild, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 1367–1376.

[102] Zhu, J., Yang, H., Liu, N., Kim, M., Zhang, W., Yang, M.H., 2018. Online multi-object tracking with dual matching attention networks, in: Proceedings of the European Conference on Computer Vision (ECCV), pp. 366–382.