Memory Based Online Learning of Deep Representations from Video Streams

Federico Pernici, Federico Bartoli, Matteo Bruni and Alberto Del Bimbo
MICC – Media Integration and Communcation Center
University Of Florence – Italy
{federico.pernici, federico.bartoli, matteo.bruni, alberto.delbimbo}@unifi.it

Abstract

We present a novel online unsupervised method for face identity learning from video streams. The method exploits deep face 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 forgetting strategy that detect redundant descriptors and discard them appropriately while time progresses. 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. Code will be publicly available.

1. Introduction

Visual data is massive and is growing faster than our ability to store and index it, nurtured by the diffusion and widespread use of social platforms. Their fundamental role in advancing object representation, object recognition and scene classification research have been undoubtedly assessed by the achievements of Deep Learning [1]. However, the cost of supervision remains the most critical factor for the applicability of such learning methods as linear improvements in performance require an exponential number of labeled examples [2]. Efforts to collect large quantities of annotated images, such as ImageNet [3] and Microsoft coco [4], while having an important role in advancing object recognition, don't have the necessary scalability and are hard to be extended, replicated or improved. They may also impose a ceiling on the performance of systems trained in this manner. Semi or unsupervised Deep Learning from image data still remains hard to achieve.

An attracting alternative would be to learn the object appearance from video streams with no supervision, both exploiting the large quantity of video available in the Internet and the fact that adjacent video frames contain semantically similar information. This provides a variety of conditions in which an object can be framed, and therefore a comprehensive representation of its appearance can be obtained. Accordingly, tracking a subject in the video could, at least in principle, support a sort of unsupervised incremental learning of its appearance. This would avoid or reduce the cost of annotation as time itself would provide a form of self-supervision. However, this solution is not exempt of problems [5]. On the one hand, parameter re-learning of Deep Networks, to adequately incorporate the new information without catastrophic interference, is still an open challenge [6, 7], especially when re-learning should be done in real time while tracking, without the availability of labels and with data coming from a stream which is often non-stationary. On the other hand, classic object tracking [8] has substantially divergent goals from continuous incremental learning. While in tracking the object appearance is learned only for detecting the object in the next frame (the past information is gradually *forgotten*), continuous incremental learning would require that *all* the past visual information of the object observed so far is collected in a comprehensive and cumulative representation. This requires that tracking does not drift in the presence of occlusions or appearance changes and that incremental learning should be asymptotically stable in order to converge to an univocal representa-

Figure 1. Memory based appearance representation. *Left*: Each element in the memory consists of a descriptor with an associated identity (indicated by box color) and an associated scalar value reflecting the degree of redundancy (indicated by gray box area) with respect to the current representation. *Right*: The shaded regions show the original appearance representation (i.e. VGGface). The descriptors outside those regions are learned from the video and extend the original appearance representation.

tion.

In this paper, we present a novel online unsupervised method for face identity learning from unconstrained video streams. The method exploits CNN based face detectors and descriptors together with a novel incremental memory based learning mechanism that collects descriptors and distills them based on their redundancy with respect to the current representation. This allows building a sufficiently compact and complete appearance representation of the individual identities as time advances (Fig. 1).

While we obtained comparable results with offline approaches exploiting future information in the task of multiple face tracking, our model is able to achieve better performance in face identification from unconstrained video. In addition to this, it is shown that the proposed learning procedure is asymptotically stable and the experimental evaluation confirms the theoretical result. In the following, in Section 2, we cite a few works that have been of inspiration for our work. In Section 3 we highlight our contributions, in Section 4 we expounded the approach in detail and finally, in Section 5, experimental results are given.

2. Related Work

Memory Based Learning: Inclusion of a memory mechanism in learning [9] 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 learning [10] [11]. This makes it possible to break the temporal correlations by mixing more and less recent experiences. More recently, Neural Turing Machine architectures have been proposed in [12, 13] and [14] 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 don't scale with massive video streams.

Open Set: In addition to the incremental learning procedure, the system needs to have the capability to discriminate between already known and unknown classes (*open set*) [15]. The open set classification is a problem of balancing known space (specialization) and unknown open space (generalization) according to the class rejection option. Formalization for open space risk is considered as the relative measure of open space compared to the overall measure space [15, 16, 17, 18]. The underlying assumption in these approaches is that data is I.I.D. which allows sampling the overall space uniformly. However, in a continuously data stream context, as in this paper, data is no longer independent and identically distributed, therefore balancing the known space vs the unknown space is more difficult since space with absence of data may be misinterpreted for open space. Storing data in a memory module can limit these effects [19, 20].

Open World: The other fundamental problem is incorporating the identified novel classes into the learning system (*open world*) [21]. This requirement favors non-parametric methods, since they can quickly learn never seen before information by simply storing examples. The Nearest Class Mean (NCM) classifier proposed in [22], has been shown to work well and be more robust than standard parametric classifiers in an incremental learning setting [22] [23] [24]. NCM's main shortcoming is that nonlinear data representation and/or non I.I.D. data streams limit the effectiveness of using the mean. We adopt from NCM the idea of prototype-based classification. However, the prototypes we use are not the average features vectors but we keep a representative non redundant discriminative subset.

Multiple Object Tracking: All the methods we described so far make use of ground truth labels and typically address the categorization problem in which data is manually cropped around the object boundary. An alternative approach that in principle accomplishes the class-incremental learning criteria expounded above (i.e. *open set* and *open world*) but with the addition of unknown labels and with data coming from the output of a detector (i.e. no manual cropped data) is Multiple Object Tracking (MOT) [25, 26]. Recent MOT algorithms typically adopt appearance and motion cues into an affinity model to estimate and link detections to form tracklets which are afterwards combined into final trajectories [27, 28, 29, 30, 31, 32]. Most existing MOT methods are applied to pedestrian tracking and either use simple color histogram features [28, 33, 34, 35, 36] or hand-crafted features [37, 38, 39, 40] as the appearance representation of objects and have simple appearance update mechanisms. Few exceptions can operate online and use deep features [41, 42, 43, 44] but they still assume continuous motion and do not update the appearance. MOT methods are not suited to abrupt changes across different shots or scenes since the assumptions of continuous motion no longer hold. Abrupt changes across different shots are typically handled offline by exploiting tracklets into predetermined non-overlapping shots as in clustering face descriptors [45] [46] [47] [48].

Long Term Object Tracking: Finally, another relevant research subject to our learning setting is long-term object tracking [49]. The aim of long-term object tracking 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 ([50, 51, 52, 53, 54] among the few), and only few of them have provided convincing evidence on the possibility of incremental appearance learning strategies that are asymptotically stable [50][52]. However, all of these works perform incremental learning only to detect

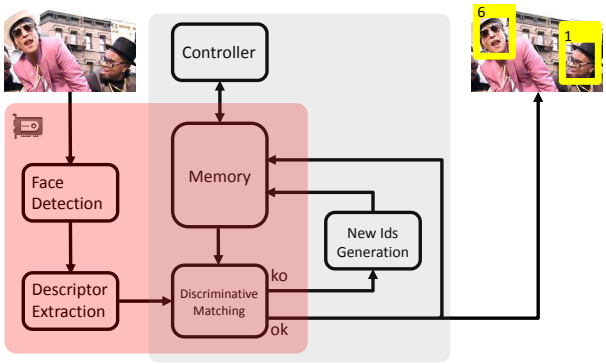

Figure 2. Block diagram presenting the major work flow and functional components in the proposed method. The gray shaded region highlights the components discussed in this paper. The memory module and the matching strategy run on the GPU.

the object in the next frame and gradually forget the past information.

3. Contributions

1. We firstly combine in a principled manner Multiple Object Tracking in an online *Open World* learning system in which the learning strategy is shown to be asymptotically stable.

2. The proposed method performs very well with respect to offline clustering methods which exploits future information.

3. Different from several existing approaches, our proposed method operates online and and hence have a wider range of applications particularly face recognition with auto-enrollment of unrecognized subjects.

4. The proposed approach

In our system, deep face descriptors are computed on face regions detected by a face detector and stored in a memory module as:

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

where $\mathbf{x}_i$ is the deep descriptor, Id_i is the object identity (an incremental number), e_i is the eligibility factor (discussed in the following), a_i tracks the age of items stored in memory and $N(t)$ is the number of descriptors at time t in the memory module.

The block diagram of the proposed system is shown in Fig. 2. As video frames are observed, new faces are detected and their descriptors are matched with those already in the memory. Each newly observed descriptor will be assigned with the object identity of its closest neighbour according to a discriminative strategy based on reverse nearest neighbor described in the next section. Unmatched descriptors of the faces in the incoming frame are stored in the memory module with a new Id. They ideally represent hypothesys of new identities that have not been observed yet and will eventually appear in the following frames. In order to learn a cumulative and comprehensive identity representation of each observed subject, two distincts problems are addressed. They are concerned with matching in consecutive frames and control of the memory module. These are separately addressed in the following subsections respectively.

4.1. Reverse Nearest Neighbour Matching

While tracking in consecutive frames, it is likely that the face of the same individual will have little differences from one frame to the following. In this case, highly similar descriptors will be stored in the memory and quickly a new face descriptor of the same individual will have comparable distances to the nearest and the second nearest descriptor already in the memory. In this case, a discriminative classifier like the Nearest Neighbor (NN) based on the distance-ratio criterion [55] does not work properly and matching cannot be assessed. We solved this problem by performing descriptor matching according to Reverse Nearest Neighbour (ReNN) [56]:

$$\mathcal{M}^\star = \left\{ (\mathbf{x}_i, \mathrm{Id}_i, e_i, a_i) \in \mathcal{M}(t) \mid \frac{||\mathbf{x}_i - 1\mathrm{NN}_{I_t}(\mathbf{x}_i)||}{||\mathbf{x}_i - 2\mathrm{NN}_{I_t}(\mathbf{x}_i)||} < \bar{\rho}, \right\} \quad (2)$$

where $\bar{\rho}$ is the distance ratio threshold for accepting a match, $\mathbf{x}_i$ is a deep face descriptor in the memory module and $1\mathrm{NN}_{I_t}(\mathbf{x}_i)$ and $2\mathrm{NN}_{I_t}(\mathbf{x}_i)$ are respectively its nearest and second nearest neighbor deep face descriptor in the incoming frame I_t.

Fig. 3 shows the effects of this change of perspective: here two new observations are detected (two distinct faces, respectively marked as $\mathbf{o}_1$ and $\mathbf{o}_2$). They both have distance ratio close to 1 to the nearest $\mathbf{x}_i$s in the memory (the dots inside the grey region S). Therefore both their matchings are undecidable. Differently from NN, ReNN is able to correctly determine the nearest descriptor for each new descriptor in the incoming frame. In fact, with ReNN, the

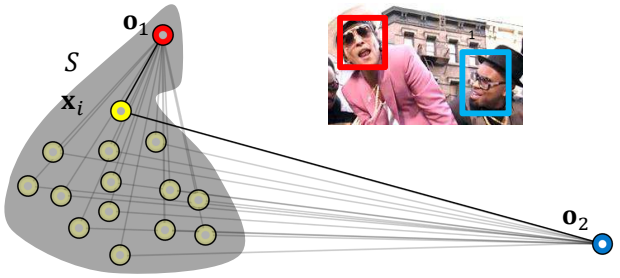

Figure 3. Reverse Nearest Neighbor for a repeated temporal visual structure (S) with the distance ratio criterion. All elements $\mathbf{x}_i$ match with $\mathbf{o}_1$, for clarity only one of them is highlighted to show the distances (thick black lines).

roles of $\mathbf{x}_i$ and $\mathbf{o}_i$ are exchanged and the distance ratio is computed between each $\mathbf{x}_i$ and the $\mathbf{o}_i$ as shown in figure for one of the $\mathbf{x}_i$s (the yellow dot is associated to the newly observed red dot). Due to the fact that with ReNN a large number of descriptors (those accumulated in the memory module) is matched against a relatively small set of descriptors (those observed in the current image), calculation of the ratio between distances could be computationally expensive. However, minimum distances can be efficiently obtained by performing twice a brute force search with parallel implementation on GPU [57]. This technique not only leverages the very efficient CUDA matrix multiplication kernel for computing the squared distance matrix but it also exploits the GPU parallelism since each query is independent. GPU limited bandwidth is not an issue being the memory incrementally populated.

The other important advantage of using ReNN is that all the descriptors $\mathbf{x}_i$ of the shown repeated structure S of Fig. 3 match with the descriptor $\mathbf{o}_1$ resulting in a one to many correspondence: $\{\mathbf{o}_1\} \leftrightarrow \{\mathbf{x}_i\}$. This capability provides a simple and sound method in the selection of those redundant descriptors that need to be condensed into a more compact representation. The feature $\mathbf{o}_1$ will be used, as described in the next section, to influence the other matched (redundant) features $\mathbf{x}_i$ regarding the fact that they belong to the same repeated structure. Therefore not only ReNN restores the discriminative matching capability under the distance ratio criterion but it also creates the foundation for the development of memory control strategies to correctly forget the redundant feature information.

4.2. Memory Control

Descriptors that have been matched according to ReNN ideally represent different appearances of a same subject face. However, collecting these descriptors indefinitely could quickly determine memory overload. To detect redundant descriptors and discard them appropriately, we defined a dimensionless quantity e_i referred to as *eligibility*. This is set to $e_i = 1$ as a descriptor is entered in the memory module and hence decreased at each match with a newly observed descriptor proportionally to the distance ratio:

$$e_i(t+1) = \eta_i \, e_i(t). \tag{3}$$

When doing this, we also re-set the age: $a_i = 0$. Eligibility allows to take into account the discriminative spatial redundancy at a rate proportional to the success of matching in consecutive frames. In fact, as the eligibility e_i of a face descriptor $\mathbf{x}_i$ in the memory drops below a given threshold $\bar{e}$ (that happens after a number of matches), that descriptor with its associated identity, age and relative eligibility is removed from the memory module:

$$\text{if}\,(e_i < \bar{e})\,\text{then}\,\mathcal{M}(t+1) = \mathcal{M}(t) \setminus \{(\mathbf{x}_i, \mathrm{Id}_i, e_i, a_i)\}. \tag{4}$$

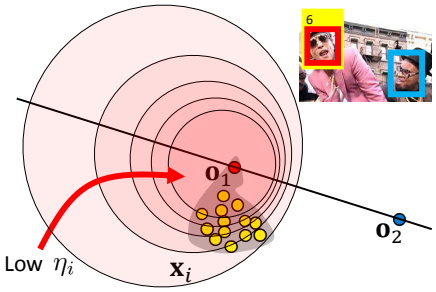

Figure 4. The shape of the density (here in 2D) down-weighting the eligibility associated to each matched descriptor in the memory. Features $\mathbf{x}_i$ in proximity of the observed descriptor $\mathbf{o}_1$ have their eligibility decreased to encourage their redundancy. The asymmetric shape of the density encourages more diversity in the *open space* far from the identity $\mathbf{o}_2$ rather than close.

The value η_i is computed according to:

$$\eta_i = \left[\frac{1}{\bar{\rho}} \frac{d_i^1}{d_i^2}\right]^{\alpha}, \tag{5}$$

where d_i^1 and d_i^2 are respectively the distances between $\mathbf{x}_i$ and its first and second nearest neighbour $\mathbf{o}_i$, the value $\bar{\rho}$ is the distance-ratio threshold of Eq. 2 here used to normalize η_i in the unit interval. The value of α emphasizes the effect of the distance-ratio. With every memory update we also increment the age a_i of all non-matched elements by 1. Eq. 5 defines a density that weights more the eligibility around the matched features and less the eligibility far apart from their second nearest neighbor. This definition is similar to discriminative distance metric learning in which the features belonging to two different classes are expected to be separated as much as possible in the feature space. The density defined by Eq. 5 can be visualized in Fig. 4 for some values of the distance ratio below the matching threshold $\bar{\rho}$. Each 2D circle in the figure visually represents the density weighting the eligibility of the matching descriptors. The geometric shape of the density is a generalization to multiple dimensions of the Apollonious circle[1]. In particular, the asymmetric shape of the density induced by the distance ratio encourages learning feature diversity in the open space. Therefore not only the matching is discriminative and indicated for rejecting hypotheses (*Open Set*) but also well suited for learning in an *Open World*.

4.3. Temporal Coherence in Image Space

The memory model previously described breaks the temporal correlations by mixing more and less recent observations. However correlated observations remain useful as they provide a natural supervisory signal to label novel iden-

[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.

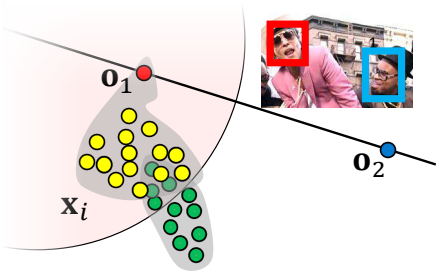

Figure 5. Matching with multiple identities. The identity $\mathbf{o}_1$ matches with two identities (yellow and green). The ambiguity is resolved by assigning $\mathbf{o}_1$ with the Id having the largest number of matched descriptors (i.e. the yellow identity).

tities. According to this the following constraints are introduced:

1. Id novelty: Potential novel identities in the current frame are included in the memory only if at least one known identity is recognized in the current frame. This allows introducing novel identity information which is known to be reasonably different from the recognized ones.

2. Id temporal coherence: An identity is assigned and included in the memory only if has been assigned in two consecutive frames. After the assignment (i.e. memory inclusion) it must match at least once in the following 3 frames, otherwise it is discarded.

3. Id uniqueness: Duplicated Ids in the current frame are not considered.

4. Id ambiguity: A subject may match with multiple identities. This ambiguity is resolved by assigning all the matched descriptors with the Id having the largest number of matched descriptors as shown in Fig. 5.

Bounding box overlap, typically used in multiple object tracking, is not exploited since not effective in unconstrained video with abrupt motions. Video temporal coherence in the image space is explicitly enforced by the 2nd constraint.

4.4. Memory Overflow Control

Our method, operating online, does not require any prior information about how many identity classes will occur, and can run for an unlimited amount of time. However, since the memory requirement is dominated by the size of the stored exemplars, if the number of identities increases indefinitely the exemplar removal based on Eq. 4 may not be sufficient in handling redundancy and the system may overflow its bounded memory. In this condition the system is forced to remove some stored exemplars by the memory limitations. To overcome this issue we follow a strategy similar to [14, 58] that involves the use of a policy based on removing from the memory the Least Recently Used Access (LRUA) exemplars. This is achieved by finding memory items with maximum age a_i in the memory, and write to one of those. Therefore the system preserves recently encoded information according to the Eligibility strategy, or writes to the last used location according to the LRUA strategy. The latter can function as an update of the memory with newer, possibly more relevant information by avoiding the deletion of rare but useful descriptors. A benefit of the LRUA strategy is that of handling those features collected in the memory that will never obtain matches. This effect is largely due to scene occluders or with descriptors extracted from bounding boxes computed from false positives of the face detector. In the long run such features may waste critical space in the memory buffer.

4.5. Asymptotic stability

Under the assumption that descriptors are sufficiently distinctive (as in the case of deep face descriptors), the incremental learning procedure described above stabilizes asymptotically around the probability density function of the descriptors of each individual subject face. This can be proved by studying the discrete dynamic system of Eq. 3 relating $e(t+1)$ to $e(t)$ by the map $T : X \mapsto X$ as $e(t+1) = T(e(t))$. A fixed point of T corresponds to an equilibrium state of the discrete dynamical system. In particular if T is a contraction there is a unique equilibrium state and the system approaches this state as time goes to infinity starting from any initial state. In this case the fixed point is globally asymptotically stable. More formally:

Theorem (Contraction Mapping) 1 *Let (X, d) be a complete metric space and $T : X \mapsto X$ be the map of Eq. 3 such that $d(T(e), T(e')) \leq c \cdot d(e, e')$ for some $0 < c \leq 1$ and all e and $e' \in X$. Then T has a unique fixed point in X. Moreover, for any $e(0) \in X$ the sequence $e(n)$ defined as $e(n+1) = T(e(n))$, converges to the fixed point of T.*

The key element that guarantees such theoretical asymptotic stability is that the ReNN distance ratio is always below 1. In fact, it is easy to demonstrate that the updating rule of Eq. 3 is a contraction and converges to its unique fixed point 0 according to the Contraction Mapping theorem (Banach fixed-point theorem).

The asymptotic stability of the method is illustrated in Fig. 6 with a simple one-dimensional case. Two patterns of synthetic descriptors, respectively modeling the case of a distinctive identity (red curve) and a non distinctive identity (black curve) are generated by two distinct 1D Gaussian distributions. The learning method was ran for 1000 iterations for three different configurations of the two distributions. The configurations reflect the limit case in which the distinctiveness assumption of the deep descriptors no longer holds. Mismatches might therefore corrupt the distinctive identity. The blue points represent the eligibility of the distinctive identity. The histogram in yellow represents the distribution of the distinctive identity as incrementally learned by the system. The three figures represent distinct cases in

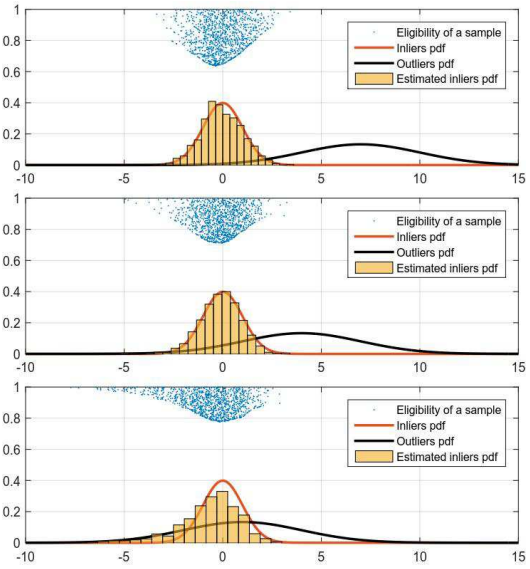

Figure 6. Asymptotic stability of incremental learning of a face identity in a sample sequence

which the non distinctive identity is progressively overlapping the distinctive one. The ReNN matching mechanism and the memory control mechanism still keep the learned distinctive identity close to its ground truth pdf.

5. Quantitative Experiments

We focus on tracking/identifying multiple faces according to their unknown identities in unconstrained videos consisting of many shots typically taken from different cameras. We used the *Music*-dataset in [48] which includes 8 music videos downloaded from YouTube with annotations of 3,845 face tracks and 117,598 face detections. We also add the first 6 episodes from Season 1 of the Big Bang Theory TV Sitcom (referred as *BBT*01-06) [36]. Each video is about more than 20 minutes long with 5-13 people and is taken mostly indoors. The main difficulty lies in identifying faces of the same subject from a long video footage.

The two algorithm parameters in Eq. 5 are set empirically to: $\bar{\rho} = 1.6$ and $\alpha = 0.01$. Deep face descriptor are extracted according to [62]. We firstly show the capability of the proposed method to perform online learning without drifting using the long sequences of the *BBT* dataset. This consists on monitoring the performance degradation of the system as time advances. A decrease in performance may eventually hinder learning being the system in a condition from which is not possible to recover. In order to build a picture of the performance over time we evaluate the method with the metric set commonly used in multiple object tracking [63]. In particular we report the MOTA: The Multiple Object Tracking Accuracy that takes into account false positives, wrongly rejected identities and identity switches as:

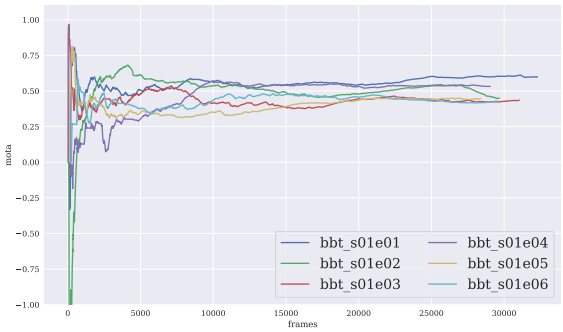

Figure 7. MOTA for each video sequence in the *BBT* dataset.

$\text{MOTA} = 1 - \frac{\sum_t (\text{FN}_t + \text{FP}_t + \text{IDS}_t)}{\sum_t \text{GT}_t}$ 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 time t. The identity switches are defined as the total number of times that a tracked trajectory changes its matched GT identity. Fig. 7 shows the MOTA curves as time progresses for each video sequence of the *BBT* dataset for about 30000 frames. Each individual frame is used to test the model before it is used for training by the incremental learning procedure [64]. As can be seen from the figure the curves reveal the stability of the learning mechanism confirming the theoretical result of Sec. 4.5. The initial fluctuations typically vary from sequence to sequence and reflect the approximate invariance of the original representation. That is, the few descriptors entering in the memory at the beginning of each sequence do not provide substantial improvement with respect to the original representation. However, as time advances, the reduction of fluctuations reveal that the proposed method is able to learn by collecting all the non-redundant descriptors it can from the video stream until no more improvement is possible.

We further compare the proposed algorithm with other state-of-the-art MOT trackers, including modified versions of TLD [65], ADMM [60], IHTLS [61]. We specifically compare with two multi-face tracking methods using the TLD method implemented as described in [48]. The first method, called mTLD, runs multiple TLD trackers for all faces, and each TLD tracker is initialized with the ground truth bounding box in the first frame. The second method, referred as mTLD2, is used to generate shot-level trajectories within each shot initializing TLD trackers with untracked detections, and link the detections in the following frames according to their overlap scores with TLD outputs.

The methods indicated as Pre-trained, SymTriplet, Triplet and Siamese refers to the four alternatives methods proposed in [48]. In these methods including ADMM, mTLD, mTLD2 and IHTLS, shot changes are detected and the video is divided into non-overlapping shots. Within each shot, a face detector is applied and adjacent detections are linked into tracklets. The recovered collection of

Table 1. Quantitative comparison with other state-of-the-art multi-object tracking methods on the *Music* video dataset

APINK

Method	Mode	IDS ↓	MOTA ↑	MOTP ↑
mTLD [59]	Offline	31	-2.2	71.2
ADMM [60]	Offline	179	72.4	76.1
IHTLS [61]	Offline	173	74.9	76.1
Pre-Trained [48]	Offline	100	54.0	75.5
mTLD2 [59]	Offline	173	77.4	76.3
Siamese [48]	Offline	124	79.0	76.3
Triplet [48]	Offline	140	78.9	76.3
SymTriplet [48]	Offline	78	80.0	76.3
Ours-dpm	**Online**	121	21.8	61
Ours-tiny	**Online**	191	55.1	65.4

BRUNOMARS

Method	Mode	IDS ↓	MOTA ↑	MOTP ↑
mTLD	Offline	35	-8.7	65.3
ADMM	Offline	428	50.6	85.7
IHTLS	Offline	375	52.7	85.8
Pre-Trained	Offline	151	48.3	88.0
mTLD2	Offline	278	52.6	87.9
Siamese	Offline	126	56.7	87.8
Triplet	Offline	126	56.6	87.8
SymTriplet	Offline	105	56.8	87.8
Ours-dpm	**Online**	78	4.5	61
Ours-tiny	**Online**	420	48.8	65.5

DARLING

Method	Mode	IDS ↓	MOTA ↑	MOTP ↑
mTLD	Offline	24	-22.0	69.9
ADMM	Offline	412	53.0	88.4
IHTLS	Offline	381	62.7	88.4
Pre-Trained	Offline	115	42.7	88.5
mTLD2	Offline	278	59.8	89.3
Siamese	Offline	214	69.5	88.9
Triplet	Offline	187	69.2	88.9
SymTriplet	Offline	169	70.5	88.9
Ours-dpm	**Online**	64	2.2	63.7
Ours-tiny	**Online**	449	62.1	66.0

GIRLSALOUD

Method	Mode	IDS ↓	MOTA ↑	MOTP ↑
mTLD	Offline	9	-1.1	71.0
ADMM	Offline	487	46.6	87.1
IHTLS	Offline	396	51.8	87.2
Pre-Trained	Offline	138	42.7	87.7
mTLD2	Offline	322	46.7	88.2
Siamese	Offline	112	51.6	87.8
Triplet	Offline	80	51.7	87.8
SymTriplet	Offline	64	51.6	87.8
Ours-dpm	**Online**	51	-2.7	61
Ours-tiny	**Online**	339	49.3	66.1

HELLOBUBBLE

Method	Mode	IDS ↓	MOTA ↑	MOTP ↑
mTLD	Offline	7	-3.5	66.5
ADMM	Offline	115	47.6	69.9
IHTLS	Offline	109	52.0	69.9
Pre-Trained	Offline	71	36.6	68.5
mTLD2	Offline	139	52.6	70.5
Siamese	Offline	105	56.3	70.6
Triplet	Offline	82	56.2	70.5
SymTriplet	Offline	69	56.5	70.5
Ours-dpm	**Online**	170	4.0	59.0
Ours-tiny	**Online**	88	51.4	69.9

PUSSYCATDOLLS

Method	Mode	IDS ↓	MOTA ↑	MOTP ↑
mTLD	Offline	24	3.1	71.3
ADMM	Offline	287	63.2	63.5
IHTLS	Offline	248	70.3	63.5
Pre-Trained	Offline	128	65.1	64.9
mTLD2	Offline	296	68.3	64.9
Siamese	Offline	107	70.3	64.9
Triplet	Offline	99	69.9	64.9
SymTriplet	Offline	82	70.2	64.9
Ours-dpm	**Online**	55	-13.5	61.1
Ours-tiny	**Online**	83	30.7	62.7

TARA

Method	Mode	IDS ↓	MOTA ↑	MOTP ↑
mTLD	Offline	130	1.4	67.9
ADMM	Offline	251	29.4	63.8
IHTLS	Offline	218	35.3	63.8
Pre-Trained	Offline	143	57.3	72.4
mTLD2	Offline	251	56.0	72.6
Siamese	Offline	106	58.4	72.5
Triplet	Offline	94	59.0	72.5
SymTriplet	Offline	75	59.2	72.4
Ours-dpm	**Online**	124	15	68
Ours-tiny	**Online**	270	39.5	76.4

WESTLIFE

Method	Mode	IDS ↓	MOTA ↑	MOTP ↑
mTLD	Offline	20	-34.7	56.9
ADMM	Offline	223	62.4	87.5
IHTLS	Offline	113	60.9	87.5
Pre-Trained	Offline	85	57.0	88.2
mTLD2	Offline	177	58.1	88.1
Siamese	Offline	74	64.1	88.0
Triplet	Offline	89	64.5	88.0
SymTriplet	Offline	57	68.6	88.1
Ours-dpm	**Online**	47	-0.2	61.5
Ours-tiny	**Online**	76	58.9	66.1

tracklets are used as face pairs (Siamese) or face triplets (Triplet and SymTriplet) to fine-tune a CNN initial face feature descriptor based on the AlexNet architecture trained on the CASIA-WebFace (Pre-trained). Then, appearance of each detection is represented with the fine-tuned feature descriptors and tracklets within each shot are linked into shot-level tracklets according to a global multiple object tracking [34, 66]. Finally tracklets across shots are subsequently merged across multiple shots into final trajectories according to the Hierarchical Agglomerative Clustering.

We reported two alternative versions using the (Deformable Part Model) DPM [67] and the Tiny [68] face detectors. For such comparisons we also include the metric MOTP: The Multiple Object Tracking Precision. MOTP is the average dissimilarity between all true positives and their corresponding ground truth objects. MOTP is a measure of localization precision. Given the quite different nature between offline and online this comparison is to be considered a proof-of-concept. However, given the good performance of the offline methods we compare to, it is certainly non-trivial for our online method to do any better. Table 1 shows that our online tracking algorithm does reasonably well with respect to offline algorithms, although there are some exceptions. In HELLOBUBBLE, BRUNOMARS, DARLING, TARA and WESTLIFE our best performing method has the MOTA score similar to the ADMM and IHTLS methods with little less identity switches. Despite the on par performance, our method achieves the results without exploiting future information. Performance are still good in APINK, the identity switches are still comparable despite a decrease in MOTA. Excluding Siamese, Triplet and SymTriplet that use a refined descriptor specifically tailored to the clustered identities extracted with the multiple passes over the sequence, our method is on par with the other offline methods. Our main observation is that with modern CNN based face detector and descriptor, the state-of-the-art offline trackers do not have expected advantages over the simpler online ones. Advantages further thin when processing long video sequences that do not fit into memory.

Results are confirmed in the *BBT* dataset as shown in Table 2. As in the previous comparison on the *Music* dataset, except for the Siamese, Triplet and SymTriplet the overall performance are very good. In the Episode four we achieved better results. Considering that CNN descriptor fine-tuning takes around 1 hour per sequence on a modern GPU, our method perform favorably in those applications operating on real time streaming data. Currently our approach runs at 10 frame per second on 800x600 video frame resolution on a Nvidia GeForce GTX TITAN X (Maxwell).

MOTA, while largely used to evaluate performance in multiple object tracking, it is not fully appropriate to evaluate the performance of identification in a *open world* scenario. In fact, it does not explicitly handle object re-identification. Different identities assigned to the same individual in two distinct scenes are not accounted as an identity switch. This effect has particular impact with videos

Table 2. Quantitative comparison with other state-of-the-art multi-object tracking methods on the *BBT* dataset.

BBT_S01E01				
Method	Mode	IDS ↓	MOTA ↑	MOTP ↑
mTLD [59]	Offline	1	-16.3	74.8
ADMM [60]	Offline	323	42.5	64.0
IHTLS [61]	Offline	312	45.7	64.0
Pre-Trained [48]	Offline	171	41.9	73.3
mTLD2 [59]	Offline	223	58.4	73.8
Siamese [48]	Offline	144	69.0	73.7
Triplet [48]	Offline	164	69.3	73.6
SymTriplet [48]	Offline	156	72.2	73.7
Ours-tiny	**Online**	24	59.9	70.3

BBT_S01E02				
Method	Mode	IDS ↓	MOTA ↑	MOTP ↑
mTLD	Offline	1	-7.6	82.8
ADMM	Offline	395	41.3	71.3
IHTLS	Offline	394	42.4	71.4
Pre-Trained	Offline	130	27.4	74.5
mTLD2	Offline	174	43.6	75.9
Siamese	Offline	116	60.4	75.8
Triplet	Offline	143	60.2	75.7
SymTriplet	Offline	102	61.6	75.7
Ours-tiny	**Online**	57	45.1	68.8

BBT_S01E03				
Method	Mode	IDS ↓	MOTA ↑	MOTP ↑
mTLD	Offline	5	-2.1	69.4
ADMM	Offline	370	30.8	68.1
IHTLS	Offline	376	33.5	68.0
Pre-Trained	Offline	110	17.8	67.5
mTLD2	Offline	142	38.0	67.9
Siamese	Offline	109	52.6	67.9
Triplet	Offline	121	50.7	67.8
SymTriplet	Offline	126	51.9	67.8
Ours-tiny	**Online**	14	43.6	68.4

BBT_S01E04				
Method	Mode	IDS ↓	MOTA ↑	MOTP ↑
mTLD	Offline	0	-15.9	76.8
ADMM	Offline	298	9.7	65.8
IHTLS	Offline	295	13.3	65.8
Pre-Trained	Offline	46	0.1	66.3
mTLD2	Offline	103	11.6	66.3
Siamese	Offline	85	23.0	66.4
Triplet	Offline	103	18.0	66.4
SymTriplet	Offline	77	19.5	66.4
Ours-tiny	**Online**	84	53.2	69.6

BBT_S01E05				
Method	Mode	IDS ↓	MOTA ↑	MOTP ↑
mTLD	Offline	1	-15.5	76.9
ADMM	Offline	380	37.4	68.2
IHTLS	Offline	360	33.8	68.2
Pre-Trained	Offline	98	32.3	75.0
mTLD2	Offline	169	46.4	74.9
Siamese	Offline	128	60.7	75.0
Triplet	Offline	118	60.5	74.9
SymTriplet	Offline	90	60.9	74.9
Ours-tiny	**Online**	36	44.5	69.3

BBT_S01E06				
Method	Mode	IDS ↓	MOTA ↑	MOTP ↑
mTLD	Offline	0	-3.9	89.3
ADMM	Offline	527	47.5	97.6
IHTLS	Offline	515	43.2	97.7
Pre-Trained	Offline	191	27.8	98.2
mTLD2	Offline	192	37.7	97.8
Siamese	Offline	156	46.2	97.9
Triplet	Offline	185	45.4	98.0
SymTriplet	Offline	196	47.6	98.0
Ours-tiny	**Online**	222	42.9	69.2

obtained from multiple cameras or with many shots. In order to take into account this case, for each sequence we also report the weighted cluster purity, defined as: $W = \frac{1}{M}\sum_c m_c p_c$, where c is the identity cluster, m_c the number of assigned identities, p_c the ratio between the most frequently occurred identity and m_c, and M denotes the total number of identity detections in the video. Table 3 and 4 show the quantitative results of the comparison with the *Music* and the *BBT* datasets. HOG, AlexNet and VGGface indicate the method [48] using alternative descriptors. HOG uses a conventional hand-crafted feature with 4356 dimensions, AlexNet uses a generic feature representation with 4096 dimensions. Our proposed approach achieves the best performance in six out of eight videos in the *Music* dataset and it achieves state of the art in all the *BBT* video sequences.

Finally, Fig. 8 shows four frames of the of the BRUNO MARS sequence with the learned identities superimposed. Faces appear sensibly diverse (see f.e. individual number

Table 3. Clustering results on *Music* Dataset. Weighted purity of each video is measured on ideal number of clusters.

MUSIC DATASET								
Videos	Apink	B. Mars	Darling	Girls A.	Hello B.	P. Dolls	T-ara	Westlife
HOG	0.20	0.36	0.19	0.29	0.35	0.28	0.22	0.27
AlexNet	0.22	0.36	0.18	0.30	0.31	0.31	0.25	0.37
Pre-trained	0.29	0.50	0.24	0.33	0.34	0.31	0.31	0.37
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
Ours-tiny	0.51	**0.96**	**0.73**	**0.89**	0.59	**0.97**	**0.72**	**0.98**

Table 4. Clustering results on Big Bang Theory Dataset. Weighted purity of each video is measured on ideal number of clusters.

BIG BANG THEORY						
Episodes	BBT01	BBT02	BBT03	BBT04	BBT05	BBT06
HOG	0.37	0.32	0.38	0.35	0.29	0.26
AlexNet	0.47	0.32	0.45	0.35	0.29	0.26
Pre-trained	0.62	0.72	0.73	0.57	0.52	0.52
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
Ours-tiny	**0.98**	**0.98**	**0.98**	**0.85**	**0.98**	**0.94**

Figure 8. Four frames from the BRUNOMARS video sequence with the superimposed estimated identities are shown.

1), nonetheless it can be observed that the learning mechanism is capable to extend the original representation to preserve identities under large pose variations including face profiles not included in the original representation.

6. Conclusion

In this paper we exploited deep CNN based face detection and descriptors coupled with a novel memory based learning mechanism that learns face identities from video sequences unsupervisedly, exploiting the temporal coherence of video frames. Particularly, all the past observed information is learned in a comprehensive representation. We demonstrate the effectiveness of the proposed method with respect to multiple face tracking on the *Music* and *BBT* datasets. The proposed method is simple, theoretically sound, asymptotically stable and follows the cumulative and convergent nature of human learning. It can be applied in principle to any other context for which a detector-descriptor combination is available (i.e. car, person, boat, traffic sign).

Acknowledgment

This research has been partially supported by Leonardo S.p.A, Italy.

References

[1] Alex Krizhevsky, Ilya Sutskever, and Geoffrey E Hinton. Imagenet classification with deep convolutional neural networks. In *NIPS*, volume 1, page 4, 2012. 1

[2] Chen Sun, Abhinav Shrivastava, Saurabh Singh, and Abhinav Gupta. Revisiting unreasonable effectiveness of data in deep learning era. In *The IEEE International Conference on Computer Vision (ICCV)*, Oct 2017. 1

[3] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *Computer Vision and Pattern Recognition, 2009. CVPR 2009. IEEE Conference on*, pages 248–255. IEEE, 2009. 1

[4] Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick. Microsoft coco: Common objects in context. In *European Conference on Computer Vision*, pages 740–755. Springer, 2014. 1

[5] Greenberg Andy. Watch a 10-year-old's face unlock his mom's iphone x. https://www.wired.com/story/10-year-old-face-id-unlocks-mothers-iphone-x/, November 2017. [Online; posted 14-November-2017]. 1

[6] Zhizhong Li and Derek Hoiem. Learning without forgetting. In *European Conference on Computer Vision*, pages 614–629. Springer, 2016. 1

[7] Andrei A Rusu, Neil C Rabinowitz, Guillaume Desjardins, Hubert Soyer, James Kirkpatrick, Koray Kavukcuoglu, Razvan Pascanu, and Raia Hadsell. Progressive neural networks. *arXiv preprint arXiv:1606.04671*, 2016. 1

[8] Matej Kristan, Ales Leonardis, Jiri Matas, Michael Felsberg, Roman Pflugfelder, Luka Cehovin Zajc, Tomas Vojir, Gustav Hager, Alan Lukezic, Abdelrahman Eldesokey, and Gustavo Fernandez. The visual object tracking vot2017 challenge results. In *The IEEE International Conference on Computer Vision (ICCV)*, Oct 2017. 1

[9] Dharshan Kumaran, Demis Hassabis, and James L McClelland. What learning systems do intelligent agents need? complementary learning systems theory updated. *Trends in cognitive sciences*, 20(7):512–534, 2016. 2

[10] Volodymyr Mnih, Koray Kavukcuoglu, David Silver, Andrei A Rusu, Joel Veness, Marc G Bellemare, Alex Graves, Martin Riedmiller, Andreas K Fidjeland, Georg Ostrovski, et al. Human-level control through deep reinforcement learning. *Nature*, 518(7540):529–533, 2015. 2

[11] Tom Schaul, John Quan, Ioannis Antonoglou, and David Silver. Prioritized experience replay. In *International Conference on Learning Representations*, Puerto Rico, 2016. 2

[12] Alex Graves, Greg Wayne, and Ivo Danihelka. Neural turing machines. *arXiv preprint arXiv:1410.5401*, 2014. 2

[13] Alex Graves, Greg Wayne, Malcolm Reynolds, Tim Harley, Ivo Danihelka, Agnieszka Grabska-Barwińska, Sergio Gómez Colmenarejo, Edward Grefenstette, Tiago Ramalho, John Agapiou, et al. Hybrid computing using a neural network with dynamic external memory. *Nature*, 538(7626):471–476, 2016. 2

[14] Adam Santoro, Sergey Bartunov, Matthew Botvinick, Daan Wierstra, and Timothy Lillicrap. Meta-learning with memory-augmented neural networks. In *International conference on machine learning*, pages 1842–1850, 2016. 2, 5

[15] Walter J Scheirer, Anderson de Rezende Rocha, Archana Sapkota, and Terrance E Boult. Toward open set recognition. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 35(7):1757–1772, 2013. 2

[16] Walter J. Scheirer, Lalit P. Jain, and Terrance E. Boult. Probability models for open set recognition. *IEEE Transactions on Pattern Analysis and Machine Intelligence (T-PAMI)*, 36, November 2014. 2

[17] Abhijit Bendale and Terrance E. Boult. Towards open set deep networks. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2016. 2

[18] Ethan M Rudd, Lalit P Jain, Walter J Scheirer, and Terrance E Boult. The extreme value machine. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2017. 2

[19] Antoine Cornuéjols. Machine Learning: The Necessity of Order (is order in order ?). In E. Lehtinen & T. O'Shea (Eds.) F. Ritter, J. Nerb, editor, *In order to learn: How the sequences of topics affect learning*. Oxford University Press, 2006. 2

[20] Antoine Cornuéjols. On-line learning: where are we so far? In *Ubiquitous knowledge discovery*, pages 129–147. Springer, 2010. 2

[21] Abhijit Bendale and Terrance Boult. Towards open world recognition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 1893–1902, 2015. 2

[22] Thomas Mensink, Jakob Verbeek, Florent Perronnin, and Gabriela Csurka. Distance-based image classification: Generalizing to new classes at near-zero cost. *IEEE transactions on pattern analysis and machine intelligence*, 35(11):2624–2637, 2013. 2

[23] Thomas Mensink, Jakob Verbeek, Florent Perronnin, and Gabriela Csurka. Metric learning for large scale image classification: Generalizing to new classes at near-zero cost. *Computer Vision–ECCV 2012*, pages 488–501, 2012. 2

[24] Marko Ristin, Matthieu Guillaumin, Juergen Gall, and Luc Van Gool. Incremental learning of ncm forests for large-scale image classification. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 3654–3661, 2014. 2

[25] Laura Leal-Taixé, Anton Milan, Konrad Schindler, Daniel Cremers, Ian Reid, and Stefan Roth. Tracking the trackers: An analysis of the state of the art in multiple object tracking. *arXiv preprint arXiv:1704.02781*, 2017. 2

[26] Wenhan Luo, Xiaowei Zhao, and Tae-Kyun Kim. Multiple object tracking: A review. *CoRR*, abs/1409.7618, 2014. 2

[27] William Brendel, Mohamed Amer, and Sinisa Todorovic. Multiobject tracking as maximum weight independent set. In *Computer Vision and Pattern Recognition (CVPR), 2011 IEEE Conference on*, pages 1273–1280. IEEE, 2011. 2

[28] Li Zhang, Yuan Li, and Ramakant Nevatia. Global data association for multi-object tracking using network flows. In *Computer Vision and Pattern Recognition, 2008. CVPR 2008. IEEE Conference on*, pages 1–8. IEEE, 2008. 2

[29] Anton Andriyenko, Konrad Schindler, and Stefan Roth. Discrete-continuous optimization for multi-target tracking. In *Computer Vision and Pattern Recognition (CVPR), 2012 IEEE Conference on*, pages 1926–1933. IEEE, 2012. 2

[30] Yuan Li, Haizhou Ai, Chang Huang, and Shihong Lao. Robust head tracking with particles based on multiple cues fusion. In *European Conference on Computer Vision*, pages 29–39. Springer, 2006. 2

[31] Yuan Li, Haizhou Ai, Takayoshi Yamashita, Shihong Lao, and Masato Kawade. Tracking in low frame rate video: A cascade particle filter with discriminative observers of different life spans. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 30(10):1728–1740, 2008. 2

[32] Seung-Hwan Bae and Kuk-Jin Yoon. Robust online multi-object tracking based on tracklet confidence and online discriminative appearance learning. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1218–1225, 2014. 2

[33] Horesh Ben Shitrit, Jerome Berclaz, Francois Fleuret, and Pascal Fua. Tracking multiple people under global appearance constraints. In *Computer Vision (ICCV), 2011 IEEE International Conference on*, pages 137–144. IEEE, 2011. 2

[34] Chang Huang, Bo Wu, and Ramakant Nevatia. Robust object tracking by hierarchical association of detection responses. In *European Conference on Computer Vision*, pages 788–801. Springer, 2008. 2, 7

[35] Yuan Li, Chang Huang, and Ram Nevatia. Learning to associate: Hybridboosted multi-target tracker for crowded scene. In *Computer Vision and Pattern Recognition, 2009. CVPR 2009. IEEE Conference on*, pages 2953–2960. IEEE, 2009. 2

[36] Baoyuan Wu, Siwei Lyu, Bao-Gang Hu, and Qiang Ji. Simultaneous clustering and tracklet linking for multi-face tracking in videos. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 2856–2863, 2013. 2, 6

[37] Markus Roth, Martin Bäuml, Ram Nevatia, and Rainer Stiefelhagen. Robust multi-pose face tracking by multi-stage tracklet association. In *Pattern Recognition (ICPR), 2012 21st International Conference on*, pages 1012–1016. IEEE, 2012. 2

[38] Bing Wang, Gang Wang, Kap Luk Chan, and Li Wang. Tracklet association by online target-specific metric learning and coherent dynamics estimation. *IEEE transactions on pattern analysis and machine intelligence*, 39(3):589–602, 2017. 2

[39] Cheng-Hao Kuo and Ram Nevatia. How does person identity recognition help multi-person tracking? In *Computer Vision and Pattern Recognition (CVPR), 2011 IEEE Conference on*, pages 1217–1224. IEEE, 2011. 2

[40] Ramazan Gokberk Cinbis, Jakob Verbeek, and Cordelia Schmid. Unsupervised metric learning for face identification in tv video. In *Computer Vision (ICCV), 2011 IEEE International Conference on*, pages 1559–1566. IEEE, 2011. 2

[41] Fengwei Yu, Wenbo Li, Quanquan Li, Yu Liu, Xiaohua Shi, and Junjie Yan. Poi: Multiple object tracking with high performance detection and appearance feature. In *European Conference on Computer Vision*, pages 36–42. Springer, 2016. 2

[42] Alex Bewley, Zongyuan Ge, Lionel Ott, Fabio Ramos, and Ben Upcroft. Simple online and realtime tracking. In *2016 IEEE International Conference on Image Processing (ICIP)*, pages 3464–3468, 2016. 2

[43] Amir Sadeghian, Alexandre Alahi, and Silvio Savarese. Tracking the untrackable: Learning to track multiple cues with long-term dependencies. In *The IEEE International Conference on Computer Vision (ICCV)*, Oct 2017. 2

[44] Nicolai Wojke, Alex Bewley, and Dietrich Paulus. Simple online and realtime tracking with a deep association metric. *CoRR*, abs/1703.07402, 2017. 2

[45] Baoyuan Wu, Yifan Zhang, Bao-Gang Hu, and Qiang Ji. Constrained clustering and its application to face clustering in videos. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 3507–3514, 2013. 2

[46] Makarand Tapaswi, Omkar M Parkhi, Esa Rahtu, Eric Sommerlade, Rainer Stiefelhagen, and Andrew Zisserman. Total cluster: A person agnostic clustering method for broadcast videos. In *Proceedings of the 2014 Indian Conference on Computer Vision Graphics and Image Processing*, page 7. ACM, 2014. 2

[47] Shijie Xiao, Mingkui Tan, and Dong Xu. Weighted block-sparse low rank representation for face clustering in videos. In *European Conference on Computer Vision*, pages 123–138. Springer, 2014. 2

[48] Shun Zhang, Yihong Gong, Jia-Bin Huang, Jongwoo Lim, Jinjun Wang, Narendra Ahuja, and Ming-Hsuan Yang. Tracking persons-of-interest via adaptive discriminative features. In *European Conference on Computer Vision*, pages 415–433. Springer, 2016. 2, 6, 7, 8

[49] Long term detection and tracking workshop. In *Conjunction With The IEEE Conference on Computer Vision and Pattern Recognition (CVPR) Workshops (LTDT2014)*, June 2014. 2

[50] Z. Kalal, J. Matas, and K. Mikolajczyk. P-n learning: Bootstrapping binary classifiers by structural constraints. In *CVPR*, june 2010. 2

[51] Thang Ba Dinh, Nam Vo, and G. Medioni. Context tracker: Exploring supporters and distracters in unconstrained environments. In *CVPR*, june 2011. 2

[52] Federico Pernici and Alberto Del Bimbo. Object tracking by oversampling local features. *IEEE transactions on pattern analysis and machine intelligence*, 36(12):2538–2551, 2014. 2

[53] Yang Hua, Karteek Alahari, and Cordelia Schmid. Occlusion and motion reasoning for long-term tracking. In *Computer Vision–ECCV 2014*, pages 172–187. Springer, 2014. 2

[54] Zhibin Hong, Zhe Chen, Chaohui Wang, Xue Mei, Danil Prokhorov, and Dacheng Tao. Multi-store tracker (muster): A cognitive psychology inspired approach to object tracking. June 2015. 2

[55] David G. Lowe. Distinctive image features from scale-invariant keypoints. *IJCV*, 60:91–110, 2004. 3

[56] Flip Korn and S. Muthukrishnan. Influence sets based on reverse nearest neighbor queries. In *Proceedings of the 2000 ACM SIGMOD International Conference on Management of Data*, SIGMOD '00, pages 201–212, New York, NY, USA, 2000. ACM. 3

[57] Shengren Li and Nina Amenta. Brute-force k-nearest neighbors search on the gpu. In *International Conference on Similarity Search and Applications*, pages 259–270. Springer, 2015. 4

[58] Lukasz Kaiser, Ofir Nachum, Aurko Roy, and Samy Bengio. Learning to remember rare events. *International Conference on Learning Representations (ICLR)*, 2017. 5

[59] Z Kalal, J Matas, and K Mikolajczyk. Tracking-learning-detection. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2011. 7, 8

[60] Caglayan Dicle, Octavia I Camps, and Mario Sznaier. The way they move: Tracking multiple targets with similar appearance. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 2304–2311, 2013. 6, 7, 8

[61] Mustafa Ayazoglu, Mario Sznaier, and Octavia I Camps. Fast algorithms for structured robust principal component analysis. In *Computer Vision and Pattern Recognition (CVPR), 2012 IEEE Conference on*, pages 1704–1711. IEEE, 2012. 6, 7, 8

[62] O. M. Parkhi, A. Vedaldi, and A. Zisserman. Deep face recognition. In *British Machine Vision Conference*, 2015. 6

[63] Laura Leal-Taixé, Anton Milan, Ian Reid, Stefan Roth, and Konrad Schindler. Motchallenge 2015: Towards a benchmark for multi-target tracking. *arXiv preprint arXiv:1504.01942*, 2015. 6

[64] João Gama, Indrė Žliobaitė, Albert Bifet, Mykola Pechenizkiy, and Abdelhamid Bouchachia. A survey on concept drift adaptation. *ACM Comput. Surv.*, 46(4):44:1–44:37, March 2014. 6

[65] Zdenek Kalal, Krystian Mikolajczyk, and Jiri Matas. Tracking-learning-detection. *IEEE transactions on pattern analysis and machine intelligence*, 34(7):1409–1422, 2012. 6

[66] Junliang Xing, Haizhou Ai, and Shihong Lao. Multi-object tracking through occlusions by local tracklets filtering and global tracklets association with detection responses. In *Computer Vision and Pattern Recognition, 2009. CVPR 2009. IEEE Conference on*, pages 1200–1207. IEEE, 2009. 7

[67] Markus Mathias, Rodrigo Benenson, Marco Pedersoli, and Luc Van Gool. Face detection without bells and whistles. In *European Conference on Computer Vision*, pages 720–735. Springer, 2014. 7

[68] Peiyun Hu and Deva Ramanan. Finding tiny faces. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, July 2017. 7