Incremental Learning of People Identities

Federico Bartoli, Federico Pernici, Matteo Bruni, and Alberto Del Bimbo(✉)

MICC, Media Integration and Communication Center,
Department of Information Engineering, University of Firenze, Florence, Italy
alberto.delbimbo@unifi.it

Abstract. Face recognition in unconstrained open-world settings is a challenging problem. Differently from the closed-set and open-set face recognition scenarios that assume that the face representations of known subjects have been manually enrolled in a gallery, the open-world scenario requires that the system learns identities incrementally from frame to frame, discriminate between known and unknown identities and automatically enrolls every new identity in the gallery, so to be able to recognize it every time it is observed again in the future. Performance scaling with large number of identities is likely to be needed in real situations. In this paper we discuss the problem and present a system that has been designed to perform effective open-world face recognition in real time at both small-moderate and large scale.

Keywords: Open-world recognition · Incremental learning · Large scale

1 Introduction

Deep face recognition is now believed to surpass human performance in many scenarios of face identity face verification and authentication [1,2] and is widely used in many fields such as military, public security and many other contexts of ordinary daily life. Despite this progress, many fundamental questions are still open and should be answered in order to build robust applications for real world. Different tasks have been identified for the evaluation of face recognition systems: face verification is traditionally relevant in access control systems, re-identification in multi camera systems; closed-set face identification [32] is relevant when searching individuals into a gallery of known subjects, such as for example in forensics applications; open-set identification [4] is relevant to search systems where the system should also be able to reject probes that are not present in the gallery; open-world recognition [6] is finally relevant to those cases where the system should also be able to reject probes that are not present in the gallery and at the same time enroll the rejected subjects as new identities in the gallery. While most of the research has addressed the verification, closed set and open set tasks, very little research has been done on the latter task, the open world task, despite of the highest relevance it has in real contexts, mainly due to the difficulty to solve the many problems that this task implies.

I. Nyström et al. (Eds.): CIARP 2019, LNCS 11896, pp. 3–15, 2019.
https://doi.org/10.1007/978-3-030-33904-3_1

In this paper, we discuss face recognition in unconstrained open-world settings and present a system that has been designed to operate in real time at both small-moderate and large scale. The open-world scenario inherits the basic working principles of face (re-)identification of the closed-set and open-set scenarios in which recognition of identities are performed by considering the distance (or similarity) between meaningful features of the subject. Features of the same identity are expected to be close in the representation space, while for different identities they are expected to be far apart [3,9,10]. However, while both closed-set and open-set scenarios require that the face representations of known subjects have been manually enrolled in a gallery, in the open-world setting the system must learn identities incrementally from frame to frame, discriminate between known and unknown identities and automatically enroll every new identity in the gallery, so to be able to recognize it every time it is observed again in the future. Large scale scenarios are likely to be implied in real situations. Key challenges of open-world recognition are therefore to avoid the possible indefinite fragmentation of identities and performance scaling.

1.1 Main Issues

Deep face recognition is a mature field of research. Network architectures, such as Deepface [11], DeepID [12], VGGFace [3], FaceNet [13], and VGGFace2 [14], have been demonstrated to be able to provide very discriminative face descriptors and effective recognition. Pose invariance is traditionally a critical issue of face recognition. In the real case, we expect that observations of the same subject under changes of pose or illumination or partial occlusions originate different (although correlated) representations. While pose invariance is mandatory in the open-world recognition context, changes in facial appearance by the aging process is not an issue instead (in most real cases there is only a short time lag between the first and last appearance of a subject). A few attempts have been published to obtain pose invariance in the deep face representation [15,16].

However, on-line incremental learning from video streams almost naturally suggests that a complete model of the identity is built as a set a collection of distinct representations, each of which refers to a specific observed pose of the face. Collecting such distinct representations as faces are detected in the video sequence unsupervisedly, so that the models are continuously self-updated by the novel information observed, is not anyway free of complexity.

This feature is not supported by the current Deep face recognition systems that rely on a separate (off-line) training phase. In fact, these systems assume that a training phase is performed off-line exploiting large face datasets. Such architectures are therefore unsuited to unsupervisedly and incrementally learn person identities, since would require continuous retraining of the network as new identities are discovered.

On-line incremental learning from video streams requires therefore **the inclusion of a memory mechanism in learning** [33]. The presence of memory allows to break the temporal correlations of the observations and combine more recent and less recent representations to complete the model of appearance of the observed subject. Different memory mechanisms for incremental learning have been proposed in the literature. In Reinforcement Learning [17,18], memory has been used to store the past experience with some priority, assuming temporal coherence. Mini batches are sampled to perform incremental learning. More recently, deep network architectures named Neural Turing Machine have been proposed in [19,20] and [21] that train an external memory module to quickly encode and retrieve new information. These architectures have the ability to rapidly bind never-before-seen information after a single presentation. Both these solutions are anyway unfit to our problem. In both cases training is provided supervisedly. In Reinforcement Learning, feedbacks are explicitly provided by humans, by assigning weights of relevance to the observations and the network is updated periodically. In the Neural Turing machine, the memory is trained offline in a supervised way. Moreover, both these solutions don't scale with massive video streams.

As in the open-set setting, open-world recognition requires the capability to discriminate between already known and unknown classes [4]. The open-set classification has been modeled as a problem of balancing known space (specialization) and unknown open space (generalization) according to the class rejection option. Solutions have formalized the open space risk as the relative measure of open space compared to the overall space [4,5,7,8]. The underlying assumption is that data is independent and identically distributed, in order to allow sampling the overall space uniformly. However, in our open-world context, since observations come from a continuous video stream, **discrimination between already known and unknown classes observations cannot assume independent and identically distributed data**.

Finally, in a typical open-world recognition context, the system should manage a very large number of different identities possibly in real time. So it should scale with massive data. If identities are represented as sets of distinct representations, **some forgetting mechanism is required in the memory** that discards unuseful representations and only retains the representations useful to build unique identities. Moreover, it is unlikely that such massive data can be retained in the memory. So, **some smart mechanism that allows to switch between main and secondary memories with effective indexing is required**.

2 Principles of Operation

The block diagram of the solution proposed is shown in Fig. 1. We used the state of the art Tiny Face Detector [22] for detection and the VGG-face descriptor [3] to represent faces. A memory module is used to collect the face descriptors [36]. The matching module is a discriminative classifier that associates to each new observation the same identity id of the most similar past observations already in the memory. So, clusters of descriptors are dynamically formed each of which is ideally representative of a single identity. The memory controller has the task of discarding redundant descriptors and implements a forgetting mechanism that attempts to keep descriptors of both the most recent and frequent and

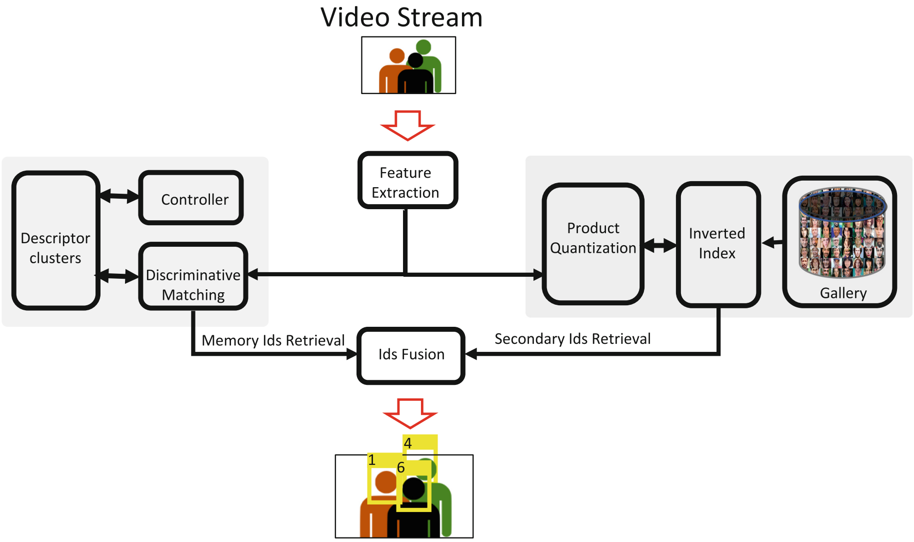


Fig. 1. Block diagram of the incremental identity learning with the main and secondary memories.

the rare observations. Consolidated clusters are periodically transferred into the secondary memory and indexed to guarantee fast access at large scales. Indexing is implemented using the FAISS framework [27] that guarantees high retrieval performance for very large number of instances. At regular time intervals the index is updated with the memory clusters, while the system continues to collect face descriptors in the main memory. As soon as a face is detected, its descriptor is matched first with the gallery in the secondary memory. Then it is either associated to an existing cluster in the memory (with its identity label) or a new cluster is formed based on Euclidean distance and Reverse Nearest Neighbor. Ideally, a new identity should be created whenever a new individual is observed that has not been observed before.

3 Matching Face Representations in Memory

In our open-world scenario, we cannot exploit Nearest Neighbor with distance ratio criterion to assess matching between face descriptors in the frame and descriptors in the memory. In fact, it is likely that faces of the same subject in consecutive frames have little differences. So, similar feature descriptors will rapidly be accumulated in the memory. Due to this, in most cases the distance ratio between the descriptor of a face observation to its nearest and the second nearest descriptor in memory will be close to 1 and the matching will be undecidable. Reverse Nearest Neighbor (ReNN) with the distance ratio criterion [31] is therefore used to assess matching. With ReNN, each descriptor in memory is NN-matched with the descriptors of the faces detected in the frame and distance ratio is used to assess matching. Since the faces detected in the frame are of different persons the distance ratio criterion can be used effectively in this case. ReNN matching could determine ambiguous assignments when distinct face observations match with descriptors of the same identity in memory, or an observation matches with descriptors of different identities. To resolve such ambiguities, we assign no identity id to the observations in the first case, and the id with the largest number of descriptors in memory, in the second case. Duplicated ids assignments to distinct face observations in the same frame are not allowed.

Accumulating matched descriptors in memory allows to dynamically create models of identities with no need of prior information about the identities and their number. At the same time, it allows to disregard the non-iid nature of data. Time of observation is not considered anymore and the descriptors in memory don't maintain the order of occurrence of the observations in the video sequence. However, the temporal coherence of the observation is useful as a form of supervision to decide whether non matched observations should be considered as new identities. Assuming that faces of the same individual have similar descriptors in consecutive frames, non-matched descriptors are assigned a new identity id only if the same identity is assessed also in the following frames (two consecutive assignments of the same id and at least one in the following three frames was verified to provide good results).

This incremental learning mechanism of memory module has two drawbacks. On the one hand a large amount of redundant information is likely will be included for each identity model (consecutive frames have similar face descriptors); on the other hand, the matching mechanism would not scale its performance at very large scales. To solve these drawbacks, we implemented the forgetting mechanism for main memory and the secondary memory indexing, respectively.

4 Forgetting Mechanism

The forgetting mechanism has the goal to avoid redundancy in the identity clusters, so that ideally, they only retain the most useful descriptors to discriminate between distinct identities. To avoid redundancy, we associate to each i-th descriptor-identity pair a dimensionless quantity e_i referred to as *eligibility-to-be-learned* (shortly *eligibility*) that indicates the relevance of the descriptor to be used as representative of the identity. Eligibility is set to 1 when the descriptor is loaded into the memory. At each match eligibility is down-weighted by a proportional amount to the matching distance ratio:

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

So descriptors in memory that have smaller distance ratio (i.e. are more similar to the observation) will have their eligibility decreased more than the others and therefore will have higher chance to be replaced in the future. In this equation, distance ratio threshold $\bar{\rho}$ is used for normalization and parameter α helps to emphasize the effect of the distance-ratio. As the eligibility of a face descriptor in memory drops below a given threshold (that happens after a number of matches), the descriptor is removed from the memory and will not be used as a representative of the identity. Effects of Eq. 1 can be appreciated in Fig. 2 that simulates a matching condition. Descriptors in memory that are very similar to observation $\mathbf{o}_1$ and dissimilar to $\mathbf{o}_2$ (dark red region) have have low matching distance ratio η; their eligibility is more down-weighted and will have higher chance to be replaced in the future. Descriptors less similar to $\mathbf{o}_1$ and dissimilar to $\mathbf{o}_2$ (light red) have higher η and their eligibility is less down-weighted; so remaining in memory is higher.

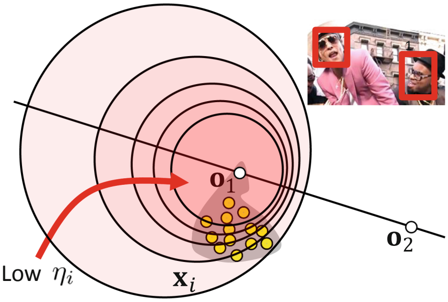


Fig. 2. Matching of descriptors in memory ($\mathbf{x}_i$) with observations in the frame ($\mathbf{o}_1$,$\mathbf{o}_2$) and effects on the eligibility (Color figure online)

According to this, the set of descriptors of a face identity in memory includes both the rare views and the most recent occurrences of frequent. The eligibility-based forgetting mechanism is accompanied by removal by aging, that helps to remove descriptors that did not receive matches for a long time (typically false positives of the detector). Similarly to [26], removal is made according to Least Recently Used Access strategy. This learning schema is well suited for the *open-world* face recognition scenario.

5 Matching Face Representations to the Secondary Memory

If the number of identities increases indefinitely, ReNN matching with forgetting may not be sufficient to maintain high recognition performance and cannot prevent memory overflow in the long term.

So, storage of descriptors in secondary memory must be used also. In this case, the descriptors extracted from the frame are matched against descriptors in both the main and secondary memory. The two outputs are taken into account to assign identity id to face descriptor.

In order to perform efficient similarity search of face descriptors in the secondary memory, we use the FAISS indexing [27] based on Inverted Index [34] and Product Quantization [30]. FAISS supports efficient search and clustering of compressed representations of the vectors at the cost of a less precise search, but allows to scale matching to billions of vectors on a single server. The inverted index groups similar descriptors in the same bucket, and represents each bucket by its centroid. Product Quantization compresses face descriptors of size 4096 into a representation of size 64, before their storage in the secondary memory. Only a limited set of candidates is considered for the nearest neighbour matching, so drastically reducing the computational cost.

6 Experiments

In the following, we will present and discuss performances of our open-world recognition system without and with the secondary memory module, in order to assess the performance in the two scenarios of small-moderate scale and large scale open-world recognition.

6.1 Small-Moderate Scale

As discussed in the introduction, besides its highest relevance in real world contexts open-world recognition has received little attention. For the small-moderate scale, considering the affinity of the problem with the Multiple Object Tracking (MOT) problem, we have compared our system (referred to as IdOL, Identity Online Learning) against a few of the most effective MOT methods published in the literature. For a correct comparative evaluation it is anyway appropriate the main differences between the two tasks. MOT methods perform data associations off-line and build identity shot-level tracklets on the basis of the whole video information (at each time instant they exploit past, present and future information).

We used the publicly available Music [23] and Big Bang Theory [24] datasets. The Music dataset includes short YouTube videos of live vocal concerts with limited number of annotated characters in continuous fast movement. In total, there are 117,598 face detections and 3,845 face tracks. The difficulty of the dataset is mainly due to the presence of frequent shot changes, rapid changes in pose, scale, viewpoint, illumination, camera motion, makeup, occlusions and special effects. The Big Bang Theory dataset collects six episodes of Big Bang Theory TV Sitcom, Season 1, approx 20' each. They include indoor ordinary scenes under a variety of settings and illumination conditions and crowding conditions. In total the dataset contains a much larger number of identities (approximately 100). In total, there are 373,392 face detections and 4,986 face tracks. Faces have large variations of appearance due to rapid changes in pose, scale, makeup, illumination, camera motion and occlusions.

Tables 1 and 2 provide the MOTA and IDS scores [25] for the experiments with the Music [23] and Big Bang Theory [24] datasets, respectively. In the Music dataset our system has lower MOTA in most videos (although almost the same of ADMM and IHTLS) but comparable IDS for HELLOBUBBLE, APINK PUSSYCATSDOLLS and WESTLIFE videos and lower IDS for T-ARA. We obtained similar results for the Big Bang Theory dataset. However, the presence of less frequent cuts and less extreme conditions due to editing effects and camera takes determines sensibly lower Identity Switch and similar MOTA in almost all the videos.

Table 1. MOTA and ID SWITCH scores comparative. *Music* dataset

	Apink		BrunoMars		Darling		GirlsAloud	
METHOD	IDS ↓	MOTA ↑	IDS ↓	MOTA ↑	IDS ↓	MOTA ↑	IDS ↓	MOTA ↑
mTLD2*	173	77.4	278	52.6	278	59.8	322	46.7
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	191	55.1	420	48.8	449	62.1	339	49.3

	HelloBubble		PussycatDolls		Tara		Westlife	
METHOD	IDS ↓	MOTA ↑	IDS ↓	MOTA ↑	IDS ↓	MOTA ↑	IDS ↓	MOTA ↑
mTLD2*	139	52.6	296	68.3	251	56.0	177	58.1
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	88	51.4	83	30.7	270	39.5	76	58.9

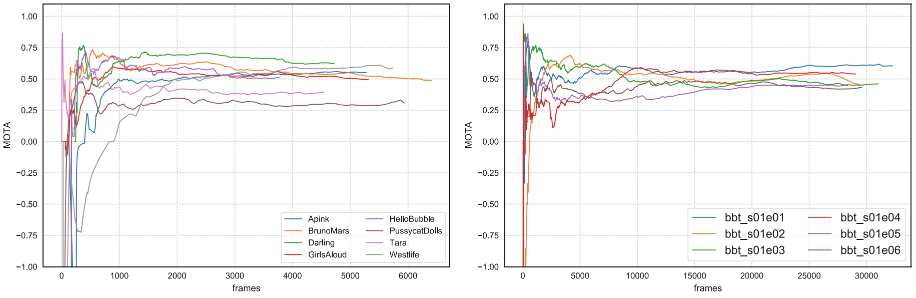
* Values reported from [24]

Table 2. MOTA and ID SWITCH scores comparative. *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 ↑
mTLD2*	223	58.4	174	43.6	142	38.0	103	11.6	169	46.4	192	37.7
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	60.37	55	45.2	14	46.1	75	53.9	35	44.7	204	43.0

* Values reported from [24]

Figure 3 shows plots of MOTA computed at each frame for the two cases. Note that MOTA has low score initially, when identity models are largely incomplete and then stabilizes at good asymptotic values as more observations are received.



(a) (b)

Fig. 3. MOTA computed at each frame for the videos in the *Music* (left) and *Big Bang Theory* (right) dataset.

6.2 Large Scale

A realistic scenario for online incremental learning of face identities at large scale is surveillance of public areas, such as railway stations, subway access, airports, malls, or open air crowded places. According to this, we evaluated the system in the large scale scenario with two different datasets, namely the SubwayFaces [35] and the ChokePoint [29], that have different image resolutions and represent different setting conditions.

The SubwayFaces is a dataset that has been used for face tracking. It includes four video sequences of real crowded subway scenes, 25 frames per second with 1920×1080 resolution (full HD). Faces are annotated with their bounding boxes and identity id (see Fig. 4). Most of the faces are of caucasian people. The dataset is available upon request for research purposes only.

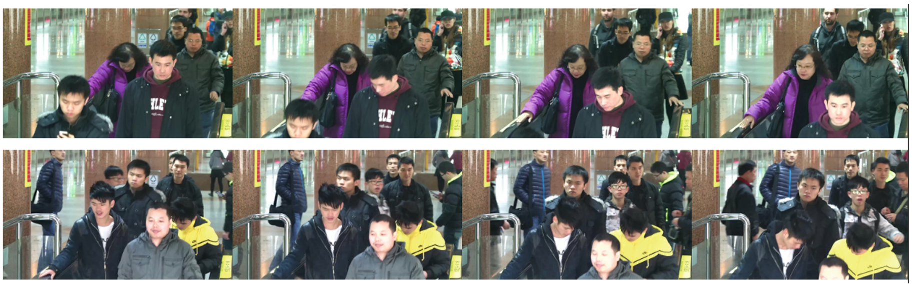


Fig. 4. Consecutive frames from the SubwayFaces dataset.

The ChokePoint dataset includes 48 sequences of gates and portals from three different cameras. The sequences are collected at 30 frames per second with 800×600 resolution. The faces have different lighting conditions, image sharpness, face poses and alignments due to camera settings. We added two more sequences of higly crowded scenes where faces also have strong continuous occlusions (see Fig. 5).

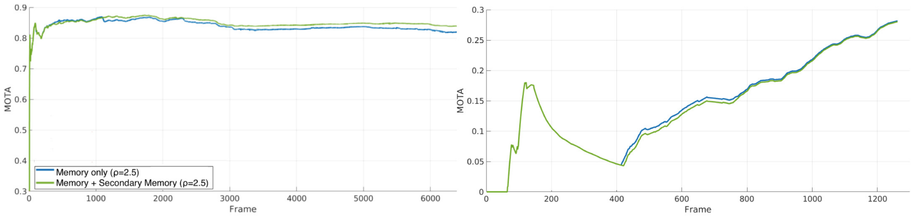
Fig. 5. Images from different portals in the ChokePoint dataset.

To simulate the large scale scenario, we populated the secondary memory with the faces of the UMDFaces dataset [28], about 400.000 face images of about 8.000 different subjects. The FAISS index was initialized with this dataset and performance of incremental learning was evaluated separately for the Subway-Faces [35] and ChokePoint [29] datasets. Descriptors removed from the main memory according to the forgetting mechanism are stored in the secondary memory at regular time intervals, and the index is updated. The system performance was evaluated with different distance ratio thresholds $\bar{\rho}$ and different parameters of descriptor quantization and FAISS index. We present here only the best results obtained with $\bar{\rho} = 2.4$.

Figure 6 shows the MOTA plots of the system for the two datasets. For each case we compared the system performance in the large scale scenario with the performance measured in the small-moderate scenario (where incremental learning does not consider the faces in the secondary memory, but applies the descriptor forgetting mechanism to expel redundant descriptors from memory).

In general, it can be noticed that for both datasets, MOTA plots of the two scenarios are very close to each other. This shows that the system is able to scale to large number of identities almost with no loss of performance. Some little improvements can be observed in the large scale scenario that can be explained with the fact that in this case, descriptors that are removed from memory are not discarded but stored in the secondary memory.

MOTA plots show different behavior and performance differences between the two datasets. For the Subwayfaces dataset, plots have similar behavior as those of the Music and Big Bang theory: MOTA stabilizes almost at the same



(a) (b)

Fig. 6. MOTA computed at each frame for the sequences in the *SubwayFaces* (left) and *ChokePoint* (right) dataset.

values after an initial interval. For the ChokePoint dataset, there are several factors that determine the large performance drop. First the camera setting: the sequences are taken from cameras placed on top of the gate, that determines face crops with smaller size and face images have lower resolution than in the Subwayfaces dataset. Second, in the sequences, people remain at the gate area only for a limited number of frames that makes identity learning largely flawed.

7 Conclusions

We have discussed open-world face recognition for both small-moderate and large scale scenarios. in both cases incremental learning of identities is performed online, at real time pace. We demonstrated how it can be performed with almost no performance drop between the two cases. Most of the good performance of the system is to be ascribed to the smart forgetting mechanism in the main memory that allows to keep both the most recent frequent descriptors and the rare descriptors in memory, and the effective indexing that supports operations in the large scale scenario.

Acknowledgments. This research has been partially supported by Leonardo SpA and TICOM, Consorzio Per Le Tecnologie Dell'informazione E Comunicazione, Italy.

References

1. Deng, W., Hu, J., Zhang, N., Chen, B., Guo, J.: Fine-grained face verification: FGLFW database, baselines, and human-DCMN partnership. Pattern Recogn. **66**, 63–73 (2017)
2. Phillips, P.J., et al.: Face recognition accuracy of forensic examiners, super recognizers, and face recognition algorithms. In: Proceedings of the National Academy of Sciences, p. 201721355 (2018)
3. Parkhi, O.M., Vedaldi, A., Zisserman, A., et al.: Deep face recognition. In: BMVC, vol. 1, p. 6 (2015)
4. Scheirer, W.J., de Rezende Rocha, A., Sapkota, A., Boult, T.E.: Toward open set recognition. IEEE Trans. Pattern Anal. Mach. Intell. **35**(7), 1757–1772 (2013)

5. Scheirer, W.J., Jain, L.P., Boult, T.E.: Probability models for open set recognition. IEEE Trans. Pattern Anal. Mach. Intell. (T-PAMI) **36** (2014)
6. Bendale, A., Boult, T.: Towards open world recognition. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 1893–1902 (2015)
7. Bendale, A., Boult, T.E.: Towards open set deep networks. In: The IEEE Conference on Computer Vision and Pattern Recognition (CVPR), June 2016
8. Rudd, E.M., Jain, L.P., Scheirer, W.J., Boult, T.E.: The extreme value machine. IEEE Trans. Pattern Anal. Mach. Intell. **40**, 762–768 (2017)
9. Wen, Y., Zhang, K., Li, Z., Qiao, Y.: A discriminative feature learning approach for deep face recognition. In: Leibe, B., Matas, J., Sebe, N., Welling, M. (eds.) ECCV 2016. LNCS, vol. 9911, pp. 499–515. Springer, Cham (2016). https://doi.org/10.1007/978-3-319-46478-7_31
10. Wang, H., et al.: CosFace: large margin cosine loss for deep face recognition. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 5265–5274 (2018)
11. Taigman, Y., Yang, M., Ranzato, M., Wolf, L.: DeepFace: closing the gap to human-level performance in face verification. In: CVPR, pp. 1701–1708 (2014)
12. Sun, Y., Liang, D., Wang, X., Tang, X.: DeepID3: face recognition with very deep neural networks. arXiv preprint arXiv:1502.00873 (2015)
13. Schroff, F., Kalenichenko, D., Philbin, J.: FaceNet: a unified embedding for face recognition and clustering. In: CVPR, pp. 815–823 (2015)
14. Cao, Q., Shen, L., Xie, W., Parkhi, O.M., Zisserman, A.: VGGFace2: a dataset for recognising faces across pose and age. arXiv preprint arXiv:1710.08092 (2017)
15. Chen, G., Shao, Y., Tang, C., Jin, Z., Zhang, J.: Deep transformation learning for face recognition in the unconstrained scene. Mach. Vis. Appl. **29**, 1–11 (2018)
16. Zhao, J., Cheng, Y., et al.: Towards pose invariant face recognition in the wild. In: CVPR, pp. 2207–2216 (2018)
17. Mnih, V., et al.: Human-level control through deep reinforcement learning. Nature **518**(7540), 529–533 (2015)
18. Schaul, T., Quan, J., Antonoglou, I., Silver, D.: Prioritized experience replay. In: International Conference on Learning Representations, Puerto Rico (2016)
19. Graves, A., Wayne, G., Danihelka, I.: Neural turing machines. arXiv preprint arXiv:1410.5401 (2014)
20. Graves, A., et al.: Hybrid computing using a neural network with dynamic external memory. Nature **538**(7626), 471–476 (2016)
21. Santoro, A., Bartunov, S., Botvinick, M., Wierstra, D., Lillicrap, T.: Meta-learning with memory-augmented neural networks. In: International Conference on Machine Learning, pp. 1842–1850 (2016)
22. Hu, P., Ramanan, D.: Finding tiny faces. In: The IEEE Conference on Computer Vision and Pattern Recognition (CVPR), July 2017
23. Zhang, S., et al.: Tracking persons-of-interest via adaptive discriminative features. In: Leibe, B., Matas, J., Sebe, N., Welling, M. (eds.) ECCV 2016. LNCS, vol. 9909, pp. 415–433. Springer, Cham (2016). https://doi.org/10.1007/978-3-319-46454-1_26
24. Bäuml, M., Tapaswi, M., Stiefelhagen, R.: Semi-supervised learning with constraints for person identification in multimedia data. In: IEEE Conference on Computer Vision and Pattern Recognition (CVPR), June 2013
25. Leal-Taixé, L., Milan, A., Reid, I., Roth, S., Schindler, K.: Motchallenge 2015: towards a benchmark for multi-target tracking. arXiv preprint arXiv:1504.01942 (2015)

26. Santoro, A., Bartunov, S., Botvinick, M., Wierstra, D., Lillicrap, T.: One-shot learning with memory-augmented neural networks. arXiv preprint arXiv:1605.06065 (2016)
27. Johnson, J., Douze, M., Jégou, H.: Billion-scale similarity search with GPUs. arXiv preprint arXiv:1702.08734 (2017)
28. Bansal, A., Nanduri, A., Castillo, C.D., Ranjan, R., Chellappa, R.: UMDFaces: an annotated face dataset for training deep networks. arXiv (2016)
29. Wong, Y., Chen, S., Mau, S., Sanderson, C., Lovell, B.C.: Patch-based probabilistic image quality assessment for face selection and improved video-based face recognition. In: Computer Vision and Pattern Recognition (CVPR) Workshops, pp. 81–88 (2011)
30. Jégou, H., Douze, M., Schmid, C.: Product quantization for nearest neighbor search. IEEE Trans. Pattern Anal. Mach. Intell. **33**(1), 117–128 (2011). https://doi.org/10.1109/TPAMI.2010.57. inria-00514462v2
31. Korn, F., Muthukrishnan, S.: Influence sets based on reverse nearest neighbor queries. In: Proceedings of the 2000 ACM SIGMOD International Conference on Management of Data, pp. 201–212. ACM, New York (2000)
32. Liu, W., Wen, Y., Yu, Z., Li, M., Raj, B., Song, L.: SphereFace: deep Hypersphere embedding for face recognition. In: The IEEE Conference on Computer Vision and Pattern Recognition (CVPR), June 2017
33. Kumaran, D., Hassabis, D., McClelland, J.L.: What learning systems do intelligent agents need? Complementary learning systems theory updated. Trends Cogn. Sci. **20**, 512–534 (2016)
34. Sivic, J., Zisserman, A.: The inverted file from "Video Google: a text retrieval approach to object matching in videos." In: ICCV (2003)
35. Wen, L., Lei, Z., Lyu, S., Li, S.Z., Yang, M.H.: Exploiting hierarchical dense structures on hypergraphs for multi-object tracking. IEEE Trans. Pattern Anal. Mach. Intell. (TPAMI) **38**, 1983–1996 (2016)
36. Pernici, F., Bartoli, F., Bruni, M., Del Bimbo, A.: Memory based online learning of deep representations from video streams. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 2324–2334 (2018)