Watchlist Adaptation: Protecting the Innocent

Manuel Günther[1], Akshay Raj Dhamija[2], Terrance E. Boult[2]

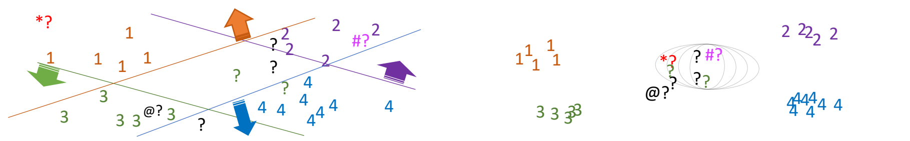
Abstract: One of the most important government applications of face recognition is the watchlist problem, where the goal is to identify a few people enlisted on a watchlist while ignoring the majority of innocent passersby. Since watchlists dynamically change and training times can be expensive, the deployed approaches use pre-trained deep networks only to provide deep features for face comparison. Since these networks never specifically trained on the operational setting or faces from the watchlist, the system will often confuse them with the faces of innocent non-watchlist subjects leading to difficult situations, e.g., being detained at the airport to resolve their identity. We develop a novel approach to take an existing pre-trained face network and use adaptation layers trained with our recently developed Objectosphere loss to provide an open-set recognition system that is rapidly adapted to the gallery while also ignoring non-watchlist faces as well as any background detections from the face detector. While our adapter network can be quickly trained without the need of re-training the entire representation network, it can also significantly improve the performance of any state-of-the-art face recognition network like VGG2. We experiment with the largest open-set face recognition dataset, the UnConstrained College Students (UCCS). It contains real surveillance camera stills including both known and unknown subjects, as well as many non-face regions from the face detector. We show that the Objectosphere approach is able to reduce the feature magnitude of unknown subjects as well as background detections, so that we can apply a specifically designed similarity function on the deep features of the Objectosphere network, which works much better than the direct prediction of the very same network. Additionally, our approach outperforms the VGG2 baseline by a large margin by rejecting the non-face data, and also outperforms prior state-of-the-art open-set recognition algorithms on the VGG2 baseline data.

Keywords: Open-Set Face Recognition, Watchlist, Gallery Adaptation.

1 Introduction

In recent years, face biometric systems using deep networks have matured into an age of high performance. These advances have lead face biometrics into daily-use applications such as access control for mobile devices or tagging friends on social media, but also an increasing usage by governments and law-enforcement for security can be observed. However, there remains at least one application for which their performance is insufficient and where errors impact innocent citizens: watchlists. A watchlist is an open-set problem and, because most people are not in the gallery of subjects of interest, the system must operate at a very low false alarm rate to reject the predominately unknown people. Recently, one of the vendors faced considerable criticism for matching US congress members to mugshots of criminals [Ro17]. That research was an eye-opener on the state of such commercial recognition systems since false alarms can substantially bias the interaction of security personnel with the person in question while increasing the responsibility for officers to verify the outputs of the system. Consequently, the latest NIST evaluations [GNH19] also include watchlist protocols, though only on images in controlled conditions.

[1] University of Zurich, Department of Informatics, Binzmühlestrasse 14, CH-8050 Zurich, guenther@ifi.uzh.ch
[2] University of Coloado Colorado Springs, Vision and Security Technology Lab, 1420 Austin Bluffs Parkway, CO-80933 Colorado Springs, {adhamija,tboult}@vast.uccs.edu



(a) Linear classifiers have unbounded open-space risk and unknowns are often misclassified

(b) Watchlist adaption refines features into tighter clusters and brings unknowns to the center

Fig. 1: OBJECTOSPHERE. *(a) Features from pre-trained face recognition networks often leave innocent subjects (?) near watchlist subjects (numbers), so any distance-based rejection fails. Gallery adaptation, e.g. using linear SVMs, can leave innocent subjects (*?, ?? #?) associated with a watchlist subject, often with high confidence. In our watchlist adaptation approach (b), we use an Objectosphere-based deep feature adapter to learn to specialize the features to the watchlist samples while mapping innocent unknown (*?, @? #?) samples to be near the origin. Watchlist subjects (numbers) have more compact representations, and when an innocent unknown (?) is presented, it has reduced magnitude and its weighted cosine distance to any class allows for rejection. Known samples will still map to be near their training samples, and the weighted cosine will correctly match them.*

In our recent work [DGB18], we showed that one of the difficulties with open-set recognition in deep networks is that deep features for unknown inputs will often map near or directly on top of the features of known inputs, as indicated in Fig. 1. If the features overlap, no distance-based algorithm can separate them. Thus we argue that for dealing with unknown subjects in a watchlist, we ideally want to learn deep features that separate the known subjects from unknown inputs. Note that unlike prior work, our ”unknowns” include very different objects, e.g., backgrounds that make it through the face detector as well as faces of unknown subjects. The latter are very similar to the known subjects, which makes this a more difficult separation.

Currently, face research/practice eschews ”training” on the gallery [Lu12]. That deeply ingrained decision is more folklore than science and we argue that difficult open-set problems such as watchlists require adapting the deep features to the specific gallery, what we call *watchlist adaptation*. There are two primary arguments against gallery training: cost and generalizability. For large face recognition, such as passport or visa management, there are many people in the gallery, and the gallery is constantly changing so training on it is impractical. However, watchlists tend to be small and rarely modified, and as we shall show in this paper, with the right design, retraining/fine-tuning for a modified watchlist is quick.

The question of generalizability is more subtle. Systems are trained across demographics to ensure that they generalize well. To obtain such a large training dataset, the data is not likely consistent with the domain of the watchlist problem. We argue that face watchlist adaptation is not generalization but a rather proper specialization to the operational domain.

Our contributions in this paper are: *(a)* We develop the first system with watchlist adaptation, providing features tuned to separate the watchlist identities from unknown identities and objects. *(b)* We develop the first novel approach to adapt a pre-trained network using Objectosphere loss. *(c)* We demonstrate that the Objectosphere adapter learns feature representations that are more robust and help to protect innocents in watchlist scenarios. *(d)* We provide state-of-the-art results on the UCCS face watchlist dataset.

2 Related Work

Since deep learning was introduced to face recognition, all modern face recognition algorithms [PVZ15, SKP15, CPC16, Sa16, Ca18, De19] rely on deep neural networks (DNNs). Many algorithms implemented special ways of training the networks in order to provide better performance on difficult images, such as triplet loss embedding [PVZ15, Sa16] or different network topologies [SKP15, Ca18]. Since these networks require large amounts of training data, usually these algorithms are trained on secondary datasets [Ba17, Ca18] that cannot have overlapping identities with tested datasets. Deep feature representations extracted from the penultimate layer are compared using simple distances such as Euclidean [PVZ15] or cosine [Sa16, CPC16, De19]. While these networks provide brilliant performance on imagery with relatively high quality, they cannot handle facial images with low (optical) resolution, or even background detections of the face detector. Generally, images with difficult content have very similar deep features and cluster in the center of the deep feature space [O'18]. Thus, none of these networks is able to reject background detections in a principled way and, therefore, they cannot be applied in real-world applications where (false) alarms need to be handled by a human operator.

A few prior works also saw the need for gallery-related training. Klare *et al.* [Kl15] argue that "training could occur on an active gallery to learn the nuances of subjects that are labeled in a gallery" but did not experiment with such gallery/subject-specific modeling. Chowdhury *et al.* [Ch16] trained one-vs-rest linear SVMs on the gallery. In neither case did they explore gallery specific features or design to separate the features of known and unknown inputs.

The most well-known open-set face recognition datasets come from the IARPA Janus benchmark (IJB) series [Kl15, Wh17, Ma18]. The biggest issue with the IJB datasets is that their protocols only include detected and manually marked faces but no background detections. In contrast, the UnControlled College Students (UCCS) dataset [SB13] and its corresponding protocol [Gü17b] mandate for faces to be detected as part of the recognition pipeline. Hence, the recognition system needs to classify both background detections and unknown faces as unknown. Due to this unique property and its true open-set nature, we use this dataset for our experiments.

3 Approach

Watchlist is a typical open-set recognition scenario where a probe may include an unknown identity. The system should only provide an alert if the probe belongs to one of the known subjects from the gallery G, but not when the probe sample is of an unknown subject $u \in U$ with $G \cap U = \emptyset$. Given a probe sample x_p of subject $g \in G$, the system D needs to produce $D(x_p) \to g$. If x_p does not belong to any subject in G, the system needs to produce $D(x_p) \to U$, even when the system never saw this specific subject.

A real-world face watchlist system consists of two sub-systems, i.e., $D = D_d \to D_r$ where D_d is a detection system and D_r is a representation system that is used to represent the output of D_d for recognition. Thus, the performance of the complete system D is tied to the performance of the detector D_d, and the representation system D_r should act as the last line of defense for overcoming the drawbacks of D_d. Therefore, D_r is susceptible to two types of samples that it should identify as unknown, i.e., when x_p is a face that does not belong to G, or when x_p is not even a face. In either case, D_r should be able to mark this probe image as not belonging to one of the known faces.

Traditionally, the use of deep networks in face watchlists is limited to representational networks, which enable researchers to decouple the training and the testing pipelines and recognize subjects that the network was not initially trained to identify. During enrollment, representations $R_g = D_r(x_g)$ are obtained for faces belonging to subject $g \in G$. These representations are used to create a gallery template G_g for the subject g. To avoid enrolling bad samples, the detection step D_d is either avoided by hand-labeling the face or at least monitored. During inference, a representation $R_p = D_d \rightarrow D_r(x_p)$ is obtained for a probe image x_p and a similarity score $s(G_g, R_p)$ is calculated between the representations of the probe and the gallery. This score is then thresholded in order to reject probe samples with low similarity to all gallery templates as unknown.

In this paper, we present a new approach to the watchlist problem, where we use the gallery for training new features so that we separate feature representations for persons of interest from representations of unknown samples. Since this approach creates a drastic difference between gallery subjects and unknown faces, it is not possible to perform enrollment for a new subject without retraining. Fortunately, since we rely on features extracted from another representational network, retraining the network is fast and could be performed whenever a new subject needs to be enrolled.

3.1 Training

We use a secondary network (D_c) containing multiple fully connected layers to classify a given feature representation $R = D_r(x)$. We use two different loss functions to train D_c, namely the Objectosphere loss as introduced in [DGB18] and the standard softmax loss with an additional background class, which is often seen for training object detectors.

Objectosphere The Objectosphere loss introduced in [DGB18] is based on the entropic open-set loss J_E. The entropic open-set loss works similarly to the traditional softmax loss for the samples x_g belonging to the known subjects, where each node S_g of the softmax output represents one of the G known subjects. Unknown samples x_u are considered as equal members of each of the possible classes:

$$J_E(x) = \begin{cases} -\log S_g(D_c(R)) & \text{if } x \text{ belongs to } g \\ -\frac{1}{G}\sum\limits_{g'=1}^{G} \log S_{g'}(D_c(R)) & \text{if } x \text{ is unknown} \end{cases} \tag{1}$$

It addition to J_E, Objectosphere applies a constraint on the magnitudes of the features representations. For unknown samples, the objective function forces the magnitudes of the penultimate layer $D_c(R)$ to be close to zero, while pushing the feature magnitudes of known samples to at least ξ, a predefined hyperparameter, which we have set to $\xi = 5$ in our experiments:

$$J_R = J_E + \lambda \begin{cases} \max(\xi - ||D_c(R)||, 0)^2 & \text{if } x \text{ is known} \\ ||D_c(R)||^2 & \text{if } x \text{ is unknown} \end{cases} \tag{2}$$

Softmax For comparison, we apply a technique that is commonly used in object detectors. Similar to the above, we utilize softmax to classify a given input as one of the subjects present in the gallery, and we add an additional output node for the unknown samples. Hence, the last layer of the softmax network has $|G|+1$ nodes.

3.2 Inference

During inference, a representation for the probe image x_p is obtained using the representational network, i.e., $R_p = D_r(x_p)$. This representation is then fed into the secondary network D_c to identify the sample as belonging to one of the subjects in the gallery or not belonging to any of them. We achieve this using the following two approaches:

Classification In this approach, we use the scores of the softmax layer S_g of the classification network D_c to link the probe to a known subject: $s(g,x_p) = S_g(D_c(R_p))$. To obtain an open-set measure, we threshold the softmax score at θ and reject the probe sample as unknown if the softmax score is below θ.

Similarity This approach is the traditional use of deep networks as feature extractors for face recognition. Since our classification network D_c contains multiple fully connected layers, it is able to learn its own representation of the incoming samples. As common, we remove the last fully-connected and the softmax layers of the network and extract $P_p = D_c(R_p)$ from the deep feature layer of the network. We enroll gallery templates G_g by averaging the normalized gallery features $P_g = D_c(R_g)$ of each known subject. For inference, we compute similarity scores between the probe and the gallery using two different similarity functions. First, we compute the cosine similarity $\cos(G_g,P_p)$ between gallery template G_g and probe feature P_p. Since Objectosphere specifically aims at manipulating the magnitude of the deep features P_p, we also multiply the cosine similarity with the magnitude of the probe feature:

$$\text{mcos}(G_g,P_p) = \cos(G_g,P_p)\cdot ||P_p|| = \frac{{G_g}^T P_p}{||G_g||} \tag{3}$$

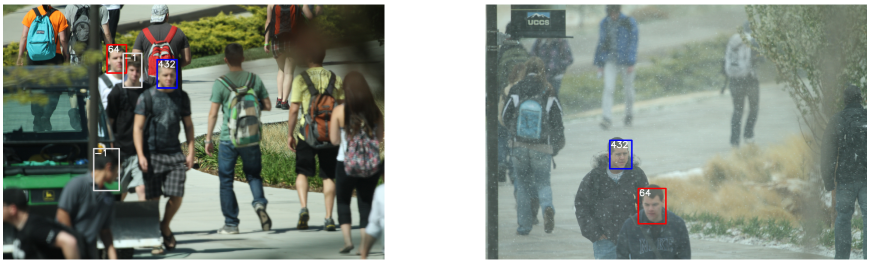
As before, the maximum similarity to any gallery template is thresholded to reject probe samples as unknown.

4 Experiments

4.1 Evaluation

To evaluate the open-set face recognition performance, we employ an adaptation of the detection and identification rate (DIR) curve [Gü17a], which usually is plotted against the probability of false alarms [PGM11]. The DIR (here we only evaluate the DIR at rank 1) is computed solely on the probe samples of known subjects $\mathbf{K}$. In the DIR, we consider probes to be correctly identified if the similarity to the correct subject g^* is the highest and above similarity threshold θ:

$$\begin{aligned}\text{DIR}(\theta) = \frac{1}{|\mathbf{K}|}\Big|\big\{P_p \mid \operatorname*{arg\,max}_g\ s(G_g,P_p) = g^* \wedge \\ s(G_{g^*},P_p) \geq \theta;\ P_p \in \mathbf{K}\big\}\Big|.\end{aligned} \tag{4}$$



(a) Sunny Day

(b) Snowy Day

Fig. 2: UCCS DATASET EXAMPLES. *Two images show examples from the UCCS dataset [Gü17b] including their ground-truth bounding boxes and labels. Identical subjects are marked with identical color, while unknown subjects are marked in white.*

The original definition of the DIR curve plots the detection and identification rate (4) over the probability of false alarms $\mathscr{P}_{\text{FA}}$ [PGM11]. The $\mathscr{P}_{\text{FA}}$ is computed on the unknown samples $\mathbf{U}$ only. A false alarm is issued when the similarity of an unknown probe sample P_p to any of the known subjects G_g is larger than θ:

$$\mathscr{P}_{\text{FA}}(\theta) = \frac{1}{|\mathbf{U}|} \left| \left\{ P_p \mid \max_g s(G_g, P_p) \geq \theta ;\; P_p \in \mathbf{U} \right\} \right| . \tag{5}$$

Using the $\mathscr{P}_{\text{FA}}$ in our evaluation has the issue that the number of unknown samples $\mathbf{U}$ might vary based on the quality of the employed face detector. Hence, a poor face detector that provides many background detections that are easy to reject by the face recognition system would be favored since it lowers the $\mathscr{P}_{\text{FA}}$. Therefore, in [Gü17b] we only computed the total number of false alarms (which we called false identifications), i.e., without normalizing by the number of unknown samples $|\mathbf{U}|$. Unfortunately, the total number of false alarms is not very intuitive and might vary based on the number of probe images. Hence, we divide the number of false alarms by the total number of probe *images* $\mathbf{I}$, where each image contains several probe faces, to obtain the average number of false alarms per image (FAI):

$$\text{FAI}(\theta) = \frac{1}{|\mathbf{I}|} \left| \left\{ P_p \mid \max_g s(G_g, P_p) \geq \theta ;\; P_p \in \mathbf{U} \right\} \right| . \tag{6}$$

We believe that this metric is best suited for selecting a threshold θ according to specific requirements. For example, if a CCTV camera captures an image every 6 seconds, and we want to limit the impact on innocent subjects by needing human operator intervention once every ten minutes, then the threshold θ should be based on an FAI of 0.01.

4.2 Dataset and Experimental Setup

We evaluate our experiments on the validation set of the UCCS dataset [Gü17b], examples of which are shown in Fig. 2. We use the source code package[3] of the challenge, which includes the evaluation scripts that we used in our evaluation. We actually found a small bug in the evaluation code, which we corrected. Additionally, we modified the false alarms axis to divide by the number of probe images to arrive at the FAI.

[3] http://pypi.org/project/challenge.uccs

In our experiments, we use the publicly available MTCNN2 face detector [Zh16] and the VGG2 face recognition [Ca18] network as our detection D_d and representation systems D_r, respectively. Since the images in the UCCS dataset are very difficult and most of the faces were not detected with the default face detector parameters, we had to lower the detection thresholds to $(0.1, 0.2, 0.2)$. With this setting, most of the faces were detected, but also a large number of background regions were marked as faces.

We detected all faces in all images and extracted the 2048-dimensional deep features of the VGG2 face recognition network for all detected bounding boxes. In total, we obtained 11299 of the 11315 known and 15792 of the 15551 unknown faces,[4] as well as 74962 background detections in the training set. Additionally to the deep features from the training set images of the known and the unknown subjects, we added all background detections of the face detector, which we used as additional unknown samples.[5]

The topology of our Objectosphere adapter network is a simple three-layer fully-connected network with 128 and 64 neurons in the first two layers, and the number of known faces in the UCCS dataset in the last layer. We trained the network with our Objectosphere loss on 90 % of the training data that contained known and unknown subjects, leaving 10 % for validation and ran 1000 training epochs using tensorflow [Ab16]. The training procedure took around 20 minutes on a regular desktop computer with a single NVidia Titan X GPU until convergence on the validation set, which was achieved after around 100 epochs. If more speed is required, the network topology can surely be adapted without a significant loss in accuracy, or more GPU resources could be used.

After training, we extracted the features from our Objectosphere network for all of the known subjects. We enrolled the gallery templates by a simple average of the normalized features so that we had a 64-dimensional template representation G_g of each subject g. Particularly, we also enrolled one gallery template for the unknown faces by computing the average 64-dimensional feature vector over all unknown faces.[6] During testing, for each detected bounding box in the validation set, we used the VGG2 features and the 64-dimensional Objectosphere features as probes.

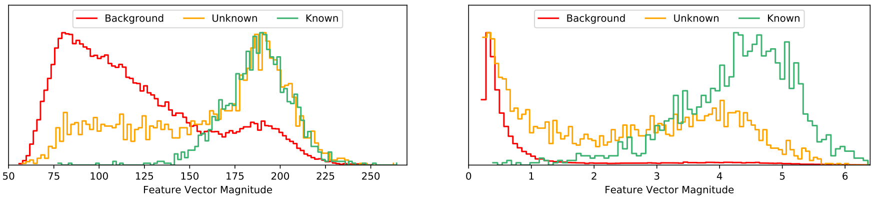
4.3 Deep Feature Magnitudes

One of the goals of Objectosphere is that the deep features $P = D_c(R)$ of unknown samples have a much smaller magnitude than those of known samples. Fig. 3 shows histograms of the feature magnitudes of all probe features of the validation set. When looking at the distribution of the Objectosphere feature magnitudes $||P||$ in Fig. 3(b), we can observe that the background samples are well-separated from the known samples and have very low magnitudes. The known samples are distributed around the desired target magnitude of $\xi = 5$, while the unknown samples have a peak close to 0, but are distributed throughout the range of magnitudes, which might be an effect of badly labeled images in the UCCS

[4] Several faces were detected multiple times and we used all of the detections.

[5] The additional unknown samples provide only a minor improvement. The background detections are to dissimilar to real faces, whereas Objectosphere requires hard negative samples to obtain good results.

[6] This additional template does not change the shape of the DIR in Fig. 4, neither for Objectosphere nor for Softmax. It only reduces the number of false alarms with very low scores and, thus, the plots in the DIR do not extend further to the right. It can be removed in an operational setting that relies on a low FAI.



(a) Original VGG2 features (b) Objectosphere Features

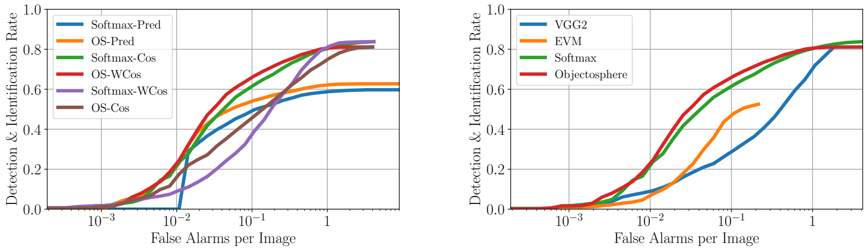
Fig. 3: FEATURE MAGNITUDE HISTOGRAMS. *Histograms of magnitudes of deep features are shown for known subjects, unknown subjects, and background detections, with deep features being extracted (a) from the original VGG2 network and (b) from the Objectosphere network. For visualization purposes, histograms are normalized individually to have the same maximum value.*

dataset, cf. Sec. 4.6. Hence, it is very easy to separate background detections from faces, but a little bit more difficult to separate known from unknown faces.

For comparison, the magnitudes $||R||$ of the original VGG2 features are shown in Fig. 3(a). As expected, the known and unknown samples have similar magnitudes, though many of the unknown samples still have lower values, which we attribute to the fact that many of the unknown faces are of very bad quality. Also, the background detections have relatively low magnitudes, but the overlap with the known samples is large. Hence, the feature magnitude $||R||$ of the original VGG2 network cannot be taken directly as an indicator if features belong to faces or to the background.

4.4 Softmax vs. Objectosphere

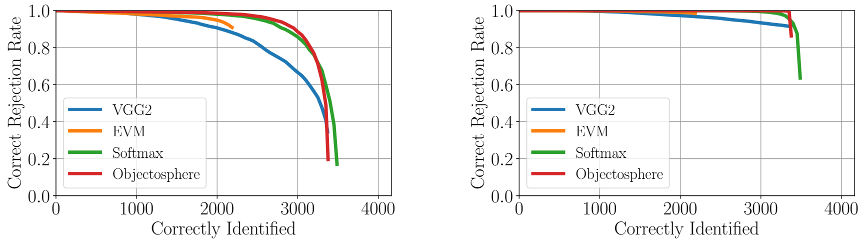
To show the advantage of our Objectosphere training procedure over softmax, we trained a network with identical topology and training strategy with softmax loss, where the negative class contained the same combination of unknown subjects and background detections from the training set. For both networks, we apply three different strategies to identify probe samples. First, we take the network predictions $S_g(D_c(R))$ as similarity values between the probe and all gallery samples. For the softmax-trained network, we ignore the unknown class prediction, but threshold on the predictions of the known classes. For the Objectosphere-trained network, we also obtain the predictions, but we additionally multiply them with the feature magnitude $||P_p||$. As the second alternative, we extract the deep features from both networks, enroll a template G_g for each training subject, and compare template and probe features P_p using the simple cosine similarity. Third, we use the same templates and probes as before, but this time we multiply the cosine distance by the probe feature magnitude (mcos) as in (3). From the DIR plots in Fig. 4(a), we can observe that the extraction of deep features from our networks performs considerably better than the predictions. As anticipated, comparing deep features from the softmax-trained network works far better with the simple cosine similarity rather than the weighted cosine. On the other hand, for Objectosphere the exact opposite is the case, and the Objectosphere network performs considerably better than the softmax-trained network, particularly at lower FAIs.



(a) Softmax vs. Objectosphere

(b) State of the Art

Fig. 4: DIR PLOTS. *In (a), DIR curves are generated for two approaches: using the final network prediction as well as enrolling models from deep features and comparing them with two similarity functions; on two different networks: softmax trained with a background class and Objectosphere. In (b), we show the comparison of our two watchlist-adapted networks (Softmax and Objectosphere) with respect to the results of the best participant (EVM) of the face recognition challenge on the UCCS dataset [Gü17b] and the original VGG2 features.*



(a) Unknown Subjects

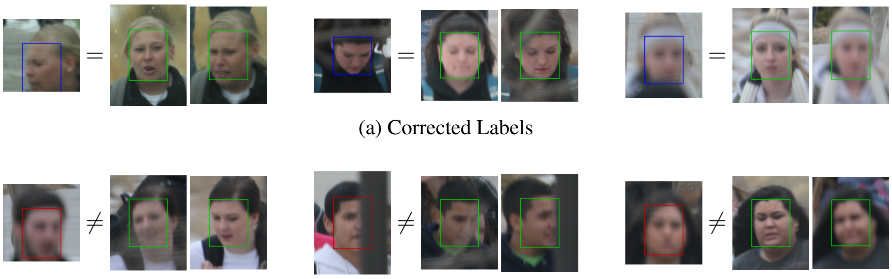
(b) Background Detections

Fig. 5: REJECTION RATES BY TYPE. *Unknown samples are split into (a) unknown subjects and (b) background detections. We plot, how many samples are correctly rejected (i.e., not identified as any known subject) for given thresholds that are based on the number of correctly identified known subjects. Both adapters (Softmax and Objectosphere) significantly improve rejection of unknowns.*

4.5 Comparison to the State of the Art

In order to compare our results to other work that reported on the UCCS dataset, we plot the results of the best participant, who used the extreme value machine (EVM). We generated the DIR curves on the validation set, where we include the VGG2 baseline that was the basis of our networks trained with softmax and Objectosphere, the results of those two networks, and the current state of the art on the UCCS dataset. The resulting DIR curve can be found in Fig. 4(b). Compared to the VGG2 baseline, which is the state of the art [Ca18] on the IARPA Janus Benchmark-A dataset, our Objectosphere improved results drastically, especially at relevant FAI thresholds. Our improved performance might be due to the very different imagery in the UCCS and IJB-A dataset and, in opposition to VGG2, we trained our networks on this type of data. More importantly, we outperform the EVM algorithm, which also trained on the UCCS dataset.

To further investigate the performance of the different systems, we evaluate the number of correctly rejected unknown samples as these samples are disregarded in DIR plots. Basing our score threshold θ on the number of correctly identified known subjects, in Fig. 5 we



(a) Corrected Labels

(b) Incorrect Predictions

Fig. 6: Error Analysis. *This figure includes six probe images (left in each triplet) and a selection of two automatically assigned gallery images (right in the triplets, green boxes). In (a) indicated by blue boxes, we present faces for which Objectosphere provided correct labels. As evident, some of these are difficult cases. (b) shows failure cases, indicated by red boxes, where Objectosphere provided an incorrect label.*

plot the percentage of unknown samples that are correctly rejected under this threshold. We separate the unknown samples into unknown subjects and background detections. From Fig. 5(a) we can observe that Objectosphere is able to identify 2000 out of the 4171 known probe faces while rejecting almost all unknown subjects. This is better than all of the related algorithms and particularly much better than the VGG2 baseline. Only the softmax-trained network comes close to our proposed Objectosphere network and exceeds it slightly on the right-hand side of the plot. Checking the rejection performance for background detections in Fig. 5(b) it is clear that Objectosphere is able to reject almost all of the background samples. Even with a threshold that allowed 3000 out of the 4171 known probe faces to be identified correctly, the rejection rate of background detections is very close to 100 %, which is considerably higher than all other algorithms. Even the softmax trained network starts accepting background detections as known subjects. Thus, we can conclude that background detections cannot harm the Objectosphere network anymore, they are successfully rejected by providing very low mcos similarity scores for all subjects in the gallery.

4.6 Failure Analysis

To analyze the errors made by our Objectosphere network, we checked the first 1000 false alarms, i.e., the detected face bounding boxes that were labeled to be unknown in the dataset but that Objectosphere identified as a certain known subject. Out of these error cases, we identified 621 faces, where the automatic label was assigned to a subject that had images from the same time stamp, but for which the assigned subject label and the ground-truth label disagreed. We manually checked those faces by showing a pair of probe face and faces from the anticipated gallery to a human who decided if the pair contains the same face. We selected to use only images with the same time stamp since in this case, additional cues like clothes and neighboring subjects could be included into the manual decision process.

With this process, we found that 573 of the 621 faces actually have a wrong label, most of the faces are labeled as unknown, and our Objectosphere approach was correct. Only in 48 of the 621 cases, our assigned label was actually wrong. Therefore, we assume that the plots shown in Fig. 4 do not reflect the reality, but false alarms actually happen far less frequently. A few examples are presented in Fig. 6, wherein Fig. 6(a) shows difficult cases where Objectosphere was able to identify the correct subject, which would have been a very hard task for face recognition systems a decade ago. On the other hand, the failure cases displayed in Fig. 6(b) indicate that the network still uses different features than humans would since many of the failure cases are really obvious to humans, at least when the local context around the face is included.

5 Conclusion

Pre-trained networks are widely used for recognition tasks. This is the first paper to demonstrate how they can be adapted to improve open-set recognition in a detection-recognition pipeline. Our approach should adapt any existing state-of-the-art detection and recognition approaches to improve support for rejecting both unknown inputs and background detections.

In this paper, we approached face recognition watchlist as an open-set problem by focusing on decreasing the false alarms of the non-gallery subjects while maintaining/improving the performance of identifying the watchlist subjects. For this purpose we used a novel open-set classification technique called Objectosphere and evaluated its effectiveness on popular face recognition metrics. With the UnConstrained College Students (UCCS) dataset, we employed the largest open-set face dataset to demonstrate this effectiveness. Using the deep features from the VGG2 face recognition network for all the detections, we trained an Objectosphere adapter with background detections and unknown faces and demonstrated the generalization to the test set. We also demonstrate that our adapter network can be used for its representation ability rather than just classification ability it was originally trained for. For probe detections, we showed that enrolling gallery templates and computing similarity scores perform better than using the raw features from a pretrained network, especially when we applied our specifically designed mcos similarity. We found that Objectosphere performs better than the state-of-the-art algorithms that were reported on the UCCS dataset.

The protocol of the UCCS dataset permits training on the gallery, which is known to favor certain types of algorithms [Lu12]. In this paper, we followed that protocol and trained on the gallery, but we are confident that the proposed algorithm would also work in rejecting background detections in protocols that do not allow gallery adaptation. However, since for the representation network all faces from the evaluation dataset are unknown, there is naturally no way of telling apart known from unknown faces without training on the gallery.

Our focus in this paper is on face watchlists which we see as a critically under-studied and socially important problem. While government/system operators may prefer the simplicity of a pre-trained system, this paper shows a significant improvement from using watchlist adaptation. We argue that modern watchlists operators should accept the mild cost of doing watchlist/gallery adaptation to protect the innocent.

References

[Ab16] Abadi, Martín et al.: TensorFlow: A System for Large-scale Machine Learning. In: USENIX Conference on Operating Systems Design and Implementation. 2016. 7

[Ba17] Bansal, Ankan; Nanduri, Anirudh; Castillo, Carlos D.; Ranjan, Rajeev; Chellappa, Rama:

UMDFaces: An Annotated Face Dataset for Training Deep Networks. In: IJCB. 2017. 3

[Ca18] Cao, Qiong; Shen, Li; Xie, Weidi; Parkhi, Omkar M.; Zisserman, Andrew: VGGFace2: A Dataset for Recognising Faces across Pose and Age. In: FG. 2018. 3, 7, 9

[Ch16] Chowdhury, Aruni Roy; Lin, Tsung-Yu; Maji, Subhransu; Learned-Miller, Erik: One-to-many Face Recognition with Bilinear CNNs. In: WACV. 2016. 3

[CPC16] Chen, Jun-Cheng; Patel, Vishal M.; Chellappa, Rama: Unconstrained Face Verification using Deep CNN Features. In: WACV. 2016. 3

[De19] Deng, Jiankang; Guo, Jia; Xue, Niannan; Zafeiriou, Stefanos: ArcFace: Additive Angular Margin Loss for Deep Face Recognition. In: CVPR. 2019. 3

[DGB18] Dhamija, Akshay Raj; Günther, Manuel; Boult, Terrance E.: Reducing Network Agnostophobia. In: NeurIPS. 2018. 2, 4

[GNH19] Grother, Patrick; Ngan, Mei; Hanaoka, Kayee: Face Recognition Vendor Test (FRVT) Part 2: Identification. Technical report, NIST, 2019. 1

[Gü17a] Günther, Manuel; Cruz, Steve; Rudd, Ethan M.; Boult, Terrance E.: Toward Open-Set Face Recognition. In: CVPR Workshops. 2017. 5

[Gü17b] Günther, Manuel et al.: Unconstrained Face Detection and Open-Set Face Recognition Challenge. In: IJCB. 2017. 3, 6, 9

[Kl15] Klare, Brendan et al.: Pushing the Frontiers of Unconstrained Face Detection and Recognition: IARPA Janus Benchmark A. In: CVPR. 2015. 3

[Lu12] Lui, Yui Man et al.: Preliminary Studies on the Good, the Bad, and the Ugly Face Recognition Challenge Problem. In: CVPR Workshops. 2012. 2, 11

[Ma18] Maze, Brianna et al.: IARPA Janus Benchmark - C: Face Dataset and Protocol. In: ICB. 2018. 3

[O'18] O'Toole, Alice J. et al.: Face Space Representations in Deep Convolutional Neural Networks. Trends in Cognitive Sciences, 2018. 3

[PGM11] Phillips, P. Jonathon; Grother, Patrick; Micheals, Ross: Evaluation Methods in Face Recognition. In: Handbook of Face Recognition. Springer, 2nd edition, 2011. 5, 6

[PVZ15] Parkhi, Omkar M.; Vedaldi, Andrea; Zisserman, Andrew: Deep Face Recognition. In: BMVC. 2015. 3

[Ro17] Romm, Tony: Amazon's facial-recognition tool misidentified 28 lawmakers as people arrested for a crime, study finds. Washington Post, July 2017. Retrieved from http://www.washingtonpost.com. 1

[Sa16] Sankaranarayanan, Swami; Alavi, Azadeh; Castillo, Carlos D.; Chellappa, Rama: Triplet Probabilistic Embedding for Face Verification and Clustering. In: BTAS. 2016. 3

[SB13] Sapkota, Archana; Boult, Terrance E.: Large Scale Unconstrained Open Set Face Database. In: BTAS. 2013. 3

[SKP15] Schroff, Florian; Kalenichenko, Dmitry; Philbin, James: FaceNet: A Unified Embedding for Face Recognition and Clustering. In: CVPR. 2015. 3

[Wh17] Whitelam, Cameron et al.: IARPA Janus Benchmark-B Face Dataset. In: CVPR Workshops. 2017. 3

[Zh16] Zhang, K.; Zhang, Z.; Li, Z.; Qiao, Y.: Joint Face Detection and Alignment using Multitask Cascaded Convolutional Networks. Signal Processing Letters, 2016. 7