Learning from a Single Labeled Face and a Stream of Unlabeled Data

Branislav Kveton, Michal Valko

Learning from a Single Labeled Face and a Stream of Unlabeled Data

Branislav Kveton
Technicolor Labs
Palo Alto, CA
branislav.kveton@technicolor.com

Michal Valko
INRIA Lille - Nord Europe, team SequeL
Lille, France
michal.valko@inria.fr

***Abstract*— Face recognition from a single image per person is a challenging problem because the training sample is extremely small. We study a variation of this problem. In our setting, only a single image of a single person is labeled, and all other people are unlabeled. This setting is very common in authentication on personal computers and mobile devices, and poses an additional challenge because it lacks negative examples. We formalize our problem as one-class classification, and propose and analyze an algorithm that learns a non-parametric model of the face from a single labeled image and a stream of unlabeled data. In many domains, for instance when a person interacts with a computer with a camera, unlabeled data are abundant and easy to utilize. We show how unlabeled data can help in learning better models and evaluate our method on 43 people. The people are identified 90% of the time at nearly zero false positives. This is 15% more often than by Fisherfaces at the same false positive rate. Finally, we conduct a comprehensive sensitivity analysis of our method and provide a guideline for setting its parameters.**

I. Introduction

Face recognition from a single image per person is a hard problem because the training sample is extremely small [1]. Yet this setting is very common in practice and therefore has been of great interest. For instance, extensive databases with one labeled image per person already exist, such as those for ID cards, and face recognition from these data could enable population-wide security screening at airports. The challenge in learning from a single labeled image is that the appearance of the face changes due to many factors, such as aging, facial expressions, or growing a mustache. In general, such changes are hard to model, especially from a single image per person.

Face recognition research has made many advances due to learning discriminative projections [2] and all state-of-the-art methods employ them in one way or another. Unfortunately, learning of high-quality discriminative projections requires a lot of labeled data. Therefore, it is not surprising that state-of-the-art face recognition methods perform poorly when only a single image per person is labeled (Section II). This problem becomes even more challenging when only a single person is labeled, and all other people are unlabeled. This is the setting considered in our paper.

We study face recognition from a single labeled image per person in the online setting. In addition to the labeled image, we observe a stream of unlabeled data, for instance recorded by a video camera. In this setting, the lack of labeled images can be compensated for by a large amount of unlabeled data. Computer vision problems usually exhibit a low-dimensional manifold structure [3] and these data can be used to learn it. We propose a new face recognition method, *online manifold tracking (OMT)*, that learns the structure of the manifold on-the-fly and can adapt to changes in data. The time and space complexity of our approach are bounded and do not increase with time. We compare our approach to several baselines and demonstrate its superiority. Finally, we evaluate its sensitivity to the setting of the parameters and discuss how to set them.

Online manifold tracking has several advantages. First, the algorithm is relatively easy to implement. Second, it does not require extensive offline training and is sufficiently fast to run in real time. In Sections IV-D and IV-E, we show that OMT recognizes faces in as little as 0.05 second on average. Third, our approach is non-parametric. We make no assumptions on recognized faces, and can adapt to various facial expressions and poses. Finally, our method is by design robust to outliers and thus suitable for open-world domains.

Non-parametric learning tends to be viewed as an alternative to learning with sophisticated features. We want to stress that our approach is complementary and benefits from better features (Section IV-C). Similarly, we believe that most face recognition algorithms, which rely on discriminative features, could benefit from adaptation and handling the concept drift. This work shows how to incorporate such features into these algorithms.

II. Face recognition from a single labeled face

Face recognition from a *single image per person* is a difficult problem [1] because all state-of-the-art face recognizers rely on a large number of training data, which are unavailable in this setting. In Fisherfaces [4], two or more images of the person are needed to estimate the within-class variance. This problem is ill posed when only one training face is available. In Laplacianfaces [3], several images of the same person are necessary to estimate the low-dimensional manifold of faces. When only one image per person is available, Laplacianfaces reduce to maximizing the between-class variance [1] and are similar to eigenfaces. Eigenfaces [5] are maximum variance projections of data obtained by principal component analysis (PCA).

In this work, we study a variation of face recognition from a single image. In our setting, only one image of one person is labeled, and many other people are unlabeled. This setting is common in open-world domains, where the class of other people is hard to model explicitly. For instance, in face-based authentication on a computer, the owner of the computer has to be modeled but it is hard, even impossible, to individually model all other people. A major challenge in this problem is the lack of negative examples. Therefore, the problem cannot

be directly formulated as learning a discriminator of positive and negative examples, as is common in face recognition [2].

One-class classification [6] is a natural way of formulating our problem. In one-class classification, the goal is to learn a hypersphere that covers positive examples. Nearest-neighbor (NN) classification with one positive example is the simplest instance of such techniques. This classifier can be written as:

$$f_R^{\mathrm{nn}}(\mathbf{x}) = \begin{cases} 1 & d(\mathbf{x}, \mathbf{x}_l) \leq R \\ 0 & \text{otherwise,} \end{cases} \quad (1)$$

where $\mathbf{x}_l$ is the labeled example, $d(\cdot, \cdot)$ is a distance function, and R is the radius of the hypersphere. In this work, we refer to R as a *generalization radius* and assume that the distance $d(\cdot, \cdot)$ is Euclidean.

The accuracy of one-class classifiers is typically measured by the *true positive (TPR)* and *false positive (FPR) rates*. The TPR is the fraction of positives classified as positives and the FPR is the fraction of negatives classified as positives. In the NN classifier $f_R^{\mathrm{nn}}(\mathbf{x})$, both rates monotonically increase with the generalization radius R. The radius R should be set such that the classifier has high TPR and acceptably low FPR.

III. Face recognition from a stream of unlabeled faces

Many face recognition algorithms can be viewed as batch-mode NN classifiers in some metric space $d(\cdot, \cdot)$ (Section V). This space is defined by discriminative features. In principle, it is hard to learn good discriminative features when only one example is labeled (Section II). So instead, we take advantage of the structure of unlabeled data and learn which part of the feature space belongs to the same person as the labeled face $\mathbf{x}_l$.

In particular, we learn a non-parametric predictor of a face from a single labeled face and a stream of unlabeled images. This problem is challenging for a few reasons. First, the data are unlabeled and may contain images of other people. As a result, it is necessary to be cautious when generalizing. This is why state-of-the-art face recognizers often perform poorly in practice. Second, the sequence of unlabeled faces may be long, and even infinite. Therefore, our non-parametric model should be compact and sublinear, or constant, in the number of observed faces.

Formally, our learning problem is modeled as a repeating game against a potentially adversarial nature. At each step t of this game, we observe an example $\mathbf{x}_t$ and then predict its label based on all observations $\mathbf{x}_1, \ldots, \mathbf{x}_t$ up to time t. This problem is challenging because only one example is labeled. Therefore, if we want to learn in this setting, we have to rely on *indirect* forms of *feedback*, such as the similarity between the observations $\mathbf{x}_1, \ldots, \mathbf{x}_t$.

This section is organized as follows. In Section III-A, we show how to compactly represent a potentially infinite stream of data. In Sections III-B and III-C, we discuss how to infer the identity of a person based on our compact representation. In Section III-D, we discuss how to set the parameters of our algorithm.

Algorithm 1 Online manifold tracking.

Input:
Representative faces u_t
Observed face $\mathbf{x}_t$
Generalization radius R
Cover radius r

if $(d(\mathbf{x}_t, \mathbf{x}_l) \leq R)$ **then**
  if $(\forall i \in u_t \ (d(\mathbf{x}_t, \mathbf{x}_i) > r))$ **then**
    $u_{t+1} \leftarrow u_t \cup \{t\}$
  else
    $u_{t+1} \leftarrow u_t$
  end if
  while $(|u_{t+1}| = k + 1)$ **do**
    $r \leftarrow 2r$
    $u_{\mathrm{old}} \leftarrow u_{t+1}$
    Greedily select face indices $u_{t+1} \subseteq u_{\mathrm{old}}$ such that:
      $\forall i \in u_{\mathrm{old}} \ \exists j \in u_{t+1} \ (d(\mathbf{x}_i, \mathbf{x}_j) \leq r)$
      $\forall i \in u_{t+1} \ \forall j \in (u_{t+1} \setminus i) \ (d(\mathbf{x}_i, \mathbf{x}_j) > r)$
  end while
end if

Output:
Representative faces u_{t+1}
Cover radius r

A. Online manifold tracking

One way of summarizing data is by mapping each example to the closest representative example. This approach is known as *data quantization* [7] and the representative examples can be found by various techniques, such k-means clustering and random sampling. In our setting, we want to summarize data on-the-fly. Two popular methods for online data quantization are online k-center clustering [8] and cover trees [9].

In this paper, we quantize faces by online k-center clustering [8]. At time t, all previously seen faces are summarized by indices u_t of up to k *representative faces*. The indices are updated as follows. If the face $\mathbf{x}_t$ at time t is at least r away from all representative faces u_t, $u_{t+1} = u_t \cup \{t\}$. Otherwise, $u_{t+1} = u_t$. Finally, when $|u_{t+1}| = k + 1$, the *cover radius* r is doubled and the representative faces are repartitioned such that no two faces are closer than r.

Our implementation of online k-center clustering is shown in Algorithm 1. Note that the example $\mathbf{x}_t$ is quantized only if it is sufficiently close to the labeled example $\mathbf{x}_l$, $d(\mathbf{x}_t, \mathbf{x}_l) \leq R$. Therefore, the *generalization radius* R essentially controls how much space is covered. In practice, it should be set such that we do not cover parts of the space that are too far away from the labeled example $\mathbf{x}_l$ and may be irrelevant when we extrapolate from it. More discussion on how to set the value of R can be found in Section III-D.

Because online k-center clustering provides guarantees on the error of its approximation [8], we can bound the error of Algorithm 1. In particular, at any time t, the distance between

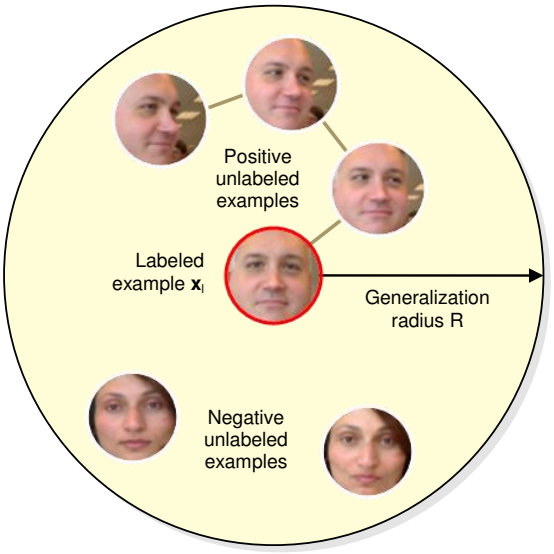

Fig. 1: An illustration of the face manifold tracked by OMT. The labeled example $\mathbf{x}_l$ is shown in the middle.

any previously seen face and the closest representative face:

$$d^t_{\max}(u_t) = \max_{\substack{i<t \\ d(\mathbf{x}_i,\mathbf{x}_l)\leq R}} \min_{j\in u_t} d(\mathbf{x}_i, \mathbf{x}_j) \tag{2}$$

is bounded by $2r$. The *error of the cover* $d^t_{\max}(u_t)$ is always smaller than 8 times of that of the optimal cover. The optimal cover of cardinality k minimizes $d^t_{\max}(\cdot)$ and its computation is NP hard.

Because the error $d^t_{\max}(u_t)$ is bounded, we can also bound the error of our identify inference algorithm in Section III-B. This proof would proceed along the lines of Valko *et al.* [10].

B. Inference

Identity inference on a manifold of faces can be formulated as a random walk on a graph, where the vertices are the faces and the edges are weighted by the similarity w_{ij} of the faces [11]. This random walk starts at an unlabeled face $\mathbf{x}_i$, jumps to neighboring faces $\mathbf{x}_j$ proportionally to their similarity w_{ij}, and is absorbed at labeled faces. The absorption probabilities $F \in [0,1]^{|u|\times|l|}$ can be computed as:

$$F = (L_{uu})^{-1} W_{ul}, \tag{3}$$

where $W \in \mathbb{R}^{n\times n}$ is the matrix of pairwise face similarities, L is its combinatorial Laplacian, l is the set of labeled faces, u is the set of unlabeled faces, and n is the number of faces. Equation 3 is well known as the *harmonic solution (HS)* and is a basis for many semi-supervised learning algorithms [12].

The main challenge in computing the harmonic solution in our setting is that we have only one labeled positive example $\mathbf{x}_l$ (Section II). Therefore, all random walks on the graph W ultimately terminate in this example and the HS $F = \mathbf{1}_{|u|\times 1}$ is meaningless. Note that our result is mathematically correct and is due to not modeling any other identity than that of $\mathbf{x}_l$.

We do not want to explicitly represent negative examples. This is because we want to model how the person looks and do not want to waste our resources on modeling other people. However, we need to introduce some notion of dissimilarity. For instance, if the vertex $\mathbf{x}_l$ cannot be reached from another vertex $\mathbf{x}_t$ in a small number of random jumps, these vertices may not be similar. To achieve this behavior, we introduce a special *sink* vertex $\mathbf{x}_0$. This vertex absorbs all random walks that reach it and is connected to all unlabeled vertices $i \in u$ by weighted edges $w_{i0} = \gamma$, where γ is a tunable parameter. Therefore, not all random walks get absorbed by the labeled vertex $\mathbf{x}_l$. The probability of being absorbed depends on the structure of W, γ, and the starting point of the random walk. Similarly to the HS (Equation 3), the absorption probabilities $\mathbf{f} \in [0,1]^{|u|\times 1}$ can be computed in a closed form [13]:

$$\mathbf{f} = (L_{uu} + \gamma I_u)^{-1} W_{ul}, \tag{4}$$

where I_u is a $|u|\times|u|$ identity matrix. Our identity inference method is outlined in Algorithm 2.

Algorithm 2 Identity inference.

Input:
Representative faces u_{t+1}
Observed face $\mathbf{x}_t$
Generalization radius R
Recognition threshold ε

if $(d(\mathbf{x}_t, \mathbf{x}_l) \leq R)$ **then**
$u \leftarrow u_{t+1}$
$v \leftarrow u \cup \{l\}$
$W \leftarrow |v| \times |v|$ similarity matrix of faces v
$D \leftarrow |v| \times |v|$ diagonal matrix s.t. $D_{ii} = \sum_j W_{ij}$
$L \leftarrow D - W$
Compute the probability that the faces v are positives:
$\mathbf{f} \leftarrow (L_{uu} + \gamma I_u)^{-1} W_{ul}$
$j \leftarrow \arg\min_{i\in u} d(\mathbf{x}_t, \mathbf{x}_i)$
$\hat{y}_t \leftarrow \mathbb{1}\{f_j > \varepsilon\}$
else
$\hat{y}_t \leftarrow 0$
end if

Output:
Identity $\hat{y}_t$ of the face $\mathbf{x}_t$

C. Algorithm

Our solution is an online learning algorithm. At time t, we quantize the face $\mathbf{x}_t$ (Algorithm 1) and then infer its identity (Algorithm 2). We refer to our technique as *online manifold tracking (OMT)* because it approximately tracks the manifold of faces and then utilizes it to build a better face recognizer. An illustration of the tracked manifold is shown in Figure 1.

Each step of our algorithm consumes $O(k^3)$ time because online k-center clustering takes $O(k)$ time and the harmonic solution can be computed in $O(k^3)$ time, by solving k linear equations. As a result, the time complexity of our algorithm is independent of time t. Note that when the similarity matrix W is $O(k)$ sparse, the time complexity of computing the HS

on W is $O(k^2)$. In addition, many fast approximate solutions exist.

D. Parameterization

Our method has several tunable parameters. In this section, we discuss how to set these parameters and explain how they affect the behavior of our algorithm. In Section IV, we show that many of these parameters do not have to be set perfectly to achieve good performance.

The generalization radius R controls extrapolation to unlabeled data. Larger values of R result in farther extrapolation. Technically, the radius R determines the maximum TPR and FPR of our method. Note that these are the same as the TPR and FPR of the classifier $f_R^{\mathrm{nn}}(\mathbf{x})$ (Equation 1) with the same radius R. In practice, R should be set to the minimum value such that the maximum TPR and FPR are high and relatively low, respectively. In Section IV-D, we conduct an experiment that shows how the generalization radius R impacts learning.

The maximum number of representative faces k trades off the error of the cover (Equation 2) for the computational cost of inference. In general, as k increases, the error of the cover decreases and the time complexity of our approach increases, cubically with k. In Section IV-E, we conduct an experiment that illustrates these trends.

The main parameter that controls the TPR and FPR of our algorithm is the recognition threshold ε (Algorithm 2). Both the TPR and FPR increase as ε decreases. So the ROC curve for our method can be generated by varying ε. We adopt this methodology in the experimental section.

Graph-based inference algorithms [12] tend to be sensitive to the choice of the graph and our method is not an exception. In our domain, the *similarity* of faces $\mathbf{x}_i$ and $\mathbf{x}_j$ is computed as:

$$w_{ij} = \exp\left[-d^2(\mathbf{x}_i, \mathbf{x}_j)/(2\sigma^2)\right], \tag{5}$$

where σ is the *heat parameter* and $d(\mathbf{x}_i, \mathbf{x}_j)$ is the distance of the faces. The *distance* is defined as $d(\mathbf{x}_i, \mathbf{x}_j) = \|\mathbf{x}_i - \mathbf{x}_j\|_2$, where $\mathbf{x}_i$ and $\mathbf{x}_j$ denote pixel intensities in 96×96 images of faces. The intensities are rescaled such that $\max_{\mathbf{x}} \|\mathbf{x}\|_2 = 1$. So the maximum distance between any two faces is two. The distance between consecutive faces in our datasets is usually less than 0.1. We set the heat parameter σ to 0.03 and so this distance is about 3σ. Our setting is motivated by a statistical rule that events that are 3σ away from the mean are unlikely. We experimented with other values σ, both 0.025 and 0.035, and all trends in our experiments remained the same.

The similarity to the sink is $\gamma = \exp[-3^2/2]$. This setting can be interpreted as follows. When two faces are closer than 3σ, the probability that a random transition between the faces terminates in the sink $\mathbf{x}_0$ is less than:

$$\frac{\gamma}{\gamma + w_{ij}} \leq \frac{\gamma}{\gamma + \exp[-3^2/2]} = \frac{1}{2}. \tag{6}$$

On the other hand, when two faces are more than 4σ and 5σ away, the probability of terminating in the sink $\mathbf{x}_0$ is at least:

$$\frac{\gamma}{\gamma + w_{ij}} \geq \frac{\gamma}{\gamma + \exp[-4^2/2]} \approx 0.9707 \tag{7}$$

and:

$$\frac{\gamma}{\gamma + w_{ij}} \geq \frac{\gamma}{\gamma + \exp[-5^2/2]} \approx 0.9997, \tag{8}$$

respectively. In other words, faces that are more distant than 3σ are likely to be perceived as different, and the probability of being different increases exponentially with their distance.

IV. Experiments

We evaluate our method on video recordings of 43 people (Section IV-A). Our experimental results support two claims. First, we show that online learning from a single labeled face and unlabeled faces performs better than supervised learning (Section IV-C). Second, we demonstrate that our approach is complementary to learning with better features. In particular, we show that OMT with Fisherfaces outperforms both OMT and Fisherfaces when used separately. Finally, we conduct a comprehensive sensitivity analysis of our method and discuss how to parameterize it (Sections IV-D and IV-E). We observe that OMT performs robustly even when its parameters are not set optimally.

A. Dataset

The VidTIMIT dataset [14] is comprised of video and the corresponding audio recordings of 43 people (Figure 2a) that recite short sentences. The dataset was recorded in 3 sessions. The delay between Sessions 1 and 2 is 7 days, and the delay between Sessions 2 and 3 is 6 days. Each person is asked to recite ten sentences: 6 in Session 1, 2 in Session 2, and 2 in Session 3. The recording was done in an office environment using a broadcast quality digital video camera. The video of each person is a sequence of 512 x 384 images. The average length of the video is 1062 images. The primary variations in our data are in facial expressions and time, since the dataset is comprised of three separate recordings.

Faces in the images are detected by OpenCV [15], turned into grayscale, resized to 96×96 pixels, cropped, and finally we equalize their histograms. We label one image per person (Figure 2b).

B. Methodology

All experiments are conducted on 43 video traces from the VidTIMIT dataset (Section IV-A). In each video, one person recites 10 sentences and no other person appears. This setting does not seem challenging because the identity in each video frame can be predicted by tracking the face from the labeled image. To make the videos more realistic, we add outliers to them. In particular, after each frame in the video (Figure 2c), we insert a randomly selected image from the remaining 42 videos (Figure 2d). The new videos are challenging because a half of the frames are negatives, people that do not belong to the modeled class. Moreover, two consecutive faces never belong to the same person, and therefore face recognition by tracking would perform poorly in this setting. However, note that the videos are still temporarily smooth in the sense that two consecutive positives are similar. OMT can identify this pattern and learns from it.

(a) People in the dataset.

(b) One labeled face per person.

(c) A sequence of faces in one video.

(d) A noisy sequence of faces. The odd faces belong to the original video (Figure 2c) and the even ones are chosen randomly from the videos of the remaining 42 people.

Fig. 2: Images and faces in the VidTIMIT dataset.

The quality of solutions is measured by their TPR and FPR at various operating points on the ROC curve. The operating points of the NN classifier are obtained by varying the radius R (Equation 1). The operating points of OMT are computed by varying the recognition threshold ε (Algorithm 2). In each video, we compute the TPR and FPR, and then average them over all videos. The generalization radius R and the number of representative faces k in OMT are by default 0.3 and 300, respectively. The sensitivity to the setting of these parameters is studied in Sections IV-D and IV-E.

C. *Quality of solutions*

In the first experiment, we compare our algorithm to three baselines. The first baseline is a 1-NN classifier (Equation 1) and we compare to it to illustrate the benefit of learning from unlabeled faces. The second baseline is a 5-NN classifier and it shows how much labeled data are needed to learn as good predictor as using our algorithm. The last baseline is a 1-NN

Fig. 3: Representative faces learned by OMT for Person 1, 15, 22, and 42. The four leftmost faces are the labeled examples.

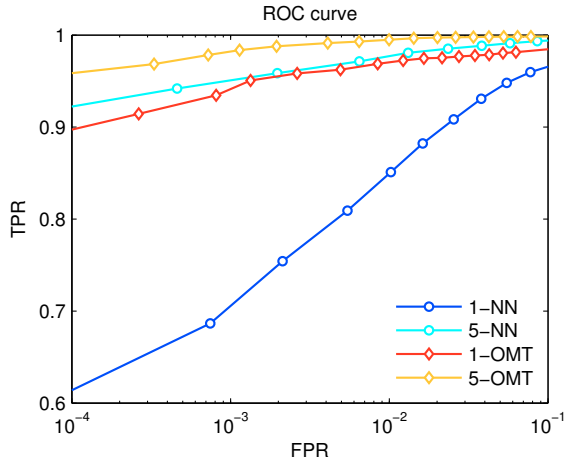

Fig. 4: Comparison of the NN and OMT recognizers that are trained from 1 and 5 labeled faces.

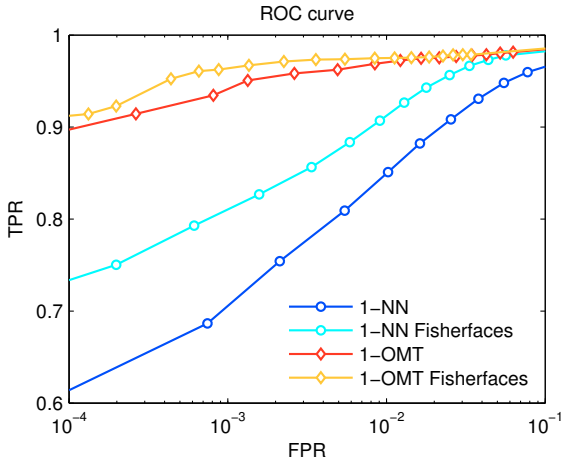

Fig. 5: Comparison of the NN and OMT recognizers that are trained on pixel intensities and projections on 64 Fisherfaces.

classifier in the space of 64 Fisherfaces. The Fisherfaces are computed from 43 labeled faces (Figure 2b), one per person. Note that the within-class scatter matrix in our problem is all zeros. Therefore, it cannot be used in Fisherfaces (Section II) and we substitute it for an identity matrix. We experimented with various numbers of Fisherfaces and 64 yields the largest area under the ROC curve in Figure 4. Our results are shown in Figures 4 and 5. We observe four major trends.

First, OMT learns a pretty accurate predictor. The TPR of OMT at 10^{-4} FPR is 0.89. In other words, OMT recognizes people most of the time at nearly zero false positives. A few examples of correctly identified faces are shown in Figure 3. Many faces are quite different from the original labeled face. Each video is processed in 45 seconds on average. Therefore, an average face is recognized in $45/(2{\cdot}1062) \approx 0.02$ second, essentially in real time.

Second, OMT performs significantly better than the 1-NN baseline. The TPR of OMT at 10^{-4} FPR is 0.89, 50% higher than that of the baseline. Note that both OMT and the 1-NN classifier are trained using the same amount of labeled data. So our comparison demonstrates the benefit of learning from unlabeled data. Finally, we plot the ROC curve for the 5-NN classifier (Figure 4) and note that it is similar to that of OMT. As a result, we may conclude that OMT learns the equivalent of 5 labeled faces.

Third, OMT performs much better than the 1-NN baseline on Fisherfaces. At low FPRs, the improvement in the TPR is in double digits. In contrast, note that most holistic methods outperform Fisherfaces and eigenfaces only in the low single digits [1].

Fourth, OMT improves with more labeled faces and better features, similarly to other face recognition algorithms. For instance, Figure 4 shows that the NN baseline improves a lot when the number of labeled faces increases from 1 to 5. The TPR of OMT, which is already in the low nineties, increases in this case by about 5%, and is higher than the new baseline at all FPRs. Figure 5 shows that the 1-NN baseline improves when the original feature space is substituted for Fisherfaces. The TPR of OMT increases in this case by 2% at low FPRs, and is higher than the new baseline at all FPRs.

D. Generalization radius R

In the second experiment, we study how the generalization radius R affects the behavior of OMT. Our results are shown in Figure 6. We observe several trends.

At all FPRs, the TPR for $R = 0.3$ is higher than the TPR for $R = 0.25$. This trend can be explained as follows. About 8% of positives are farther from the labeled example $\mathbf{x}_l$ than 0.25. Because the TPR for $R = 0.3$ is always higher than the

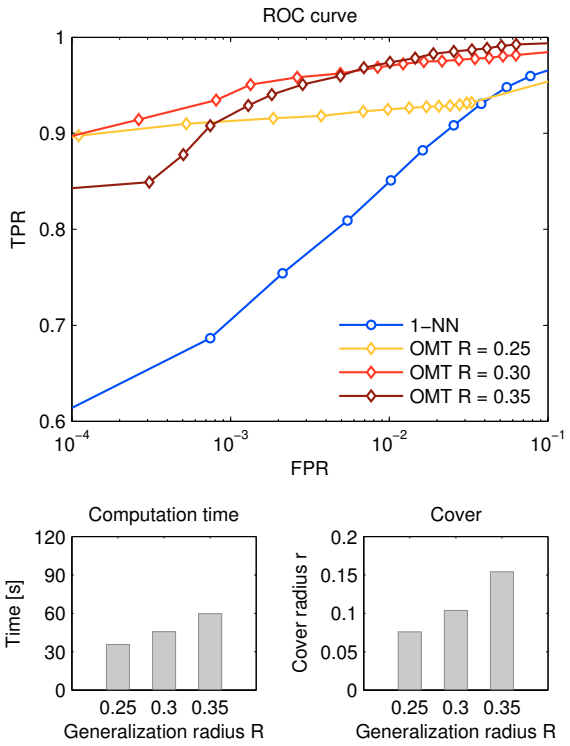

Fig. 6: Varying the generalization radius R in OMT. For each value R, we report the ROC curve, the computation time, and the cover radius r.

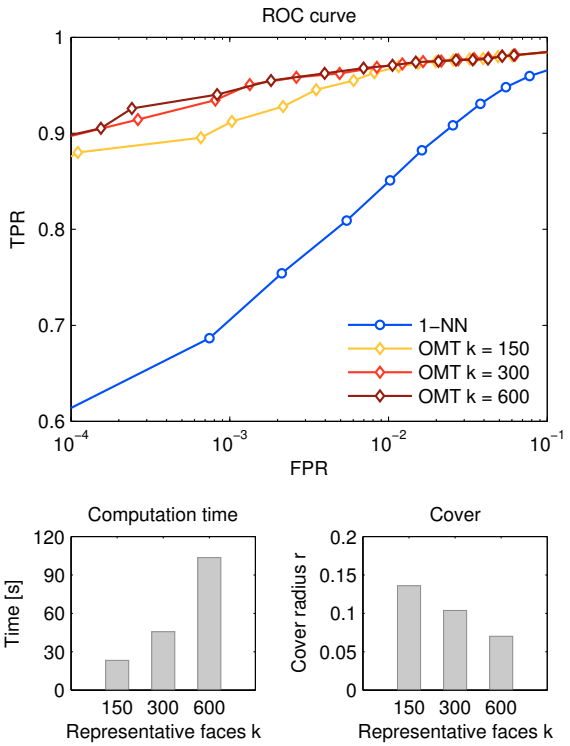

Fig. 7: Varying the number of representative faces k in OMT. For each value k, we report the ROC curve, the computation time, and the cover radius r.

TPR for $R = 0.25$, many of these positives can be classified correctly at nearly zero false positives. So the generalization radius of $R = 0.25$ is too restrictive.

At low FPRs, the TPR for $R = 0.3$ is higher than the TPR for $R = 0.35$. This trend can be explained as follows. Barely 2% of positives are farther from the labeled example $\mathbf{x}_l$ than 0.3. As a result, the potential increase in true positives when the radius R increases beyond 0.3 is small, and in our results it is outweighed by the increase in false negatives. Therefore, $R = 0.3$ yields a higher TPR at low FPRs. Nevertheless, we note that OMT performs acceptably well for all tested values of R.

Finally, note that as the generalization radius R increases, the cover radius r and computation time increase. The radius r increases since the covered space, $\mathbf{x}_t$ such that $d(\mathbf{x}_t, \mathbf{x}_l) \leq R$, increases but the number of faces k that cover it remains constant. The computation time increases because more faces satisfy $d(\mathbf{x}_t, \mathbf{x}_l) \leq R$, and must be quantized and classified.

E. Number of representative faces k

In the last experiment, we study how the behavior of OMT changes based on the number of representative faces k. Our results are reported in Figure 7. We observe several trends.

As the number of representative faces k increases, both the accuracy of inference and computation time increase, and the cover radius r decreases. The radius r decreases because the covered space, $\mathbf{x}_t$ such that $d(\mathbf{x}_t, \mathbf{x}_l) \leq R$, remains the same but the number of faces k that cover it increases. As a result, the accuracy also increases. The computation time grows less than quadratically with k, significantly slower than suggested by the analysis of our method (Section III-C). The reason for this trend is that our feature vectors $\mathbf{x}_t$ are long, 96^2 entries, and their quantization dominates the computational cost. The amortized per-step cost of online k-center clustering is $O(k)$, which is in line with the observed trend.

We recommend that the number of representative faces k be chosen as high as the computational resources allow. The more variable the face and environment, the larger the value of k. Finally, note that as few as 150 representative faces are sufficient to learn interesting patterns.

V. Related work

In this section, we review related work on face recognition and online semi-supervised learning.

A. Face recognition

The state-of-the-art in face recognition [2] advanced so far that face recognition is available in consumer products. Face recognition from a single image per person is still considered to be a hard problem and we study a variation of this problem [1]. According to Tan *et al.* [1], we propose a *holistic method* because we identify faces based on the whole image, and do not extract local features. Existing holistic methods [1] either employ some form of PCA to extract features [16] or enlarge the training set, for instance by novel views of the face [17].

The novel views are generated by transformations, which are learned from a separate training set that comprises all views.

We take a very different approach in this paper. This is the first work that shows how to learn computationally efficiently a non-parametric model of a face from a stream of unlabeled data and a single labeled face. Our method can be viewed as learning novel views of the face from unlabeled data. Unlike Beymer and Poggio [17], the method does not have an offline training phase and can learn concepts that are hard to model, such as aging or growing a mustache. The main disadvantage of our method is that it is data driven. Therefore, it may need a large amount of unlabeled data to learn. Such data may not be available in all domains.

Note that our approach is complementary to learning from more sophisticated features. In Section IV-C, we apply OMT to Fisherfaces and demonstrate that the new approach yields better results than each method separately.

B. Online semi-supervised learning

In machine learning, online learning from partially labeled data is known as *online semi-supervised learning*. This problem has been formulated and solved in various ways, such as boosting, regularization of support vector machines (SVMs), and learning on graphs. *Online semi-supervised boosting* [18] is a variation of boosting, in which unlabeled data are labeled greedily using the data adjacency graph and then employed in the standard boosting fashion. *Online manifold regularization of SVMs* [19] regularizes a max-margin classifier by the data adjacency graph. *Online semi-supervised learning on a graph* [10] incrementally compresses the data adjacency graph and then infers labels of unlabeled examples based on this graph.

All of the above methods assume that at least two classes of examples are labeled and cannot be easily extended to our setting. Valko *et al.* [10] and Balcan *et al.* [11] studied face recognition on similarity graphs from multiple labeled faces. This is the first work that studies face recognition on a graph from a single labeled image.

VI. Conclusions

In this paper, we present online manifold tracking (OMT), a new online face recognition algorithm which is suitable for environments with minimal human supervision. In comparison to existing methods, which learn discriminative features, OMT relies on unlabeled data as the main form of feedback. We evaluate our method on a dataset of 43 people and show that it produces superior results. In addition, we demonstrate that OMT is complementary to learning with better features, such as Fisherfaces. Finally, we discuss how to parameterize our method and show that it is robust to a small perturbation of its parameters.

In this work, OMT is presented as a holistic method, where the whole face is treated as an input. In our future work, we plan to extend OMT to local facial features, such as the nose, eyes, and mouth. In the single-image-per-person setting, it is accepted that local methods outperform holistic methods [1]. We strongly believe that we can improve these methods even further by online adaptation, perhaps based on similarities in consecutive video frames.

VII. Acknowledgements

This research work was supported by Ministry of Higher Education and Research, Nord-Pas de Calais Regional Council and FEDER through the "contrat de projets état region 2007–2013", and by PASCAL2 European Network of Excellence. The research leading to these results has also received funding from the European Community's Seventh Framework Programme (FP7/2007-2013) under grant agreement n° 270327 (project CompLACS).

References

[1] X. Tan, S. Chen, Z.-H. Zhou, and F. Zhang, "Face recognition from a single image per person: A survey," *Pattern Recognition*, vol. 39, no. 9, pp. 1725–1745, 2006.

[2] W.-Y. Zhao, R. Chellappa, P. Phillips, and A. Rosenfeld, "Face recognition: A literature survey," *ACM Computing Surveys*, vol. 35, no. 4, pp. 399–458, 2003.

[3] X. He, S. Yan, Y. Hu, P. Niyogi, and H. Zhang, "Face recognition using Laplacianfaces," *IEEE Transactions Pattern Analysis and Machine Intelligence*, vol. 27, no. 3, pp. 328–340, 2005.

[4] P. Belhumeur, J. Hespanha, and D. Kriegman, "Eigenfaces vs. Fisherfaces: Recognition using class specific linear projection," *IEEE Transactions Pattern Analysis and Machine Intelligence*, vol. 19, no. 7, pp. 711–720, 1997.

[5] M. Turk and A. Pentland, "Face recognition using eigenfaces," in *IEEE Conference on Computer Vision and Pattern Recognition*, 1991, pp. 586–591.

[6] D. Tax, "One-class classification," Ph.D. dissertation, Tu Delft, 2001.

[7] R. Gray and D. Neuhoff, "Quantization," *IEEE Transactions on Information Theory*, vol. 44, no. 6, pp. 2325–2383, 1998.

[8] M. Charikar, C. Chekuri, T. Feder, and R. Motwani, "Incremental clustering and dynamic information retrieval," in *Proceedings of the 29th Annual ACM Symposium on Theory of Computing*, 1997, pp. 626–635.

[9] A. Beygelzimer, S. Kakade, and J. Langford, "Cover trees for nearest neighbor," in *Proceedings of the 23rd International Conference on Machine Learning*, 2006, pp. 97–104.

[10] M. Valko, B. Kveton, L. Huang, and D. Ting, "Online semi-supervised learning on quantized graphs," in *Proceedings of the 26th Conference on Uncertainty in Artificial Intelligence*, 2010.

[11] M.-F. Balcan, A. Blum, P. P. Choi, J. Lafferty, B. Pantano, M. R. Rwebangira, and X. Zhu, "Person identification in webcam images: An application of semi-supervised learning," in *ICML 2005 Workshop on Learning with Partially Classified Training Data*, 2005.

[12] X. Zhu, Z. Ghahramani, and J. Lafferty, "Semi-supervised learning using Gaussian fields and harmonic functions," in *Proceedings of the 20th International Conference on Machine Learning*, 2003, pp. 912–919.

[13] B. Kveton, M. Valko, A. Rahimi, and L. Huang, "Semi-supervised learning with max-margin graph cuts," in *Proceedings of the 13th International Conference on Artificial Intelligence and Statistics*, 2010, pp. 421–428.

[14] C. Sanderson and B. Lovell, "Multi-region probabilistic histograms for robust and scalable identity inference," in *Proceedings of the 3rd International Conferences on Advances in Biometrics*, 2009, pp. 199–208.

[15] G. Bradski, "The OpenCV Library," *Dr. Dobb's Journal of Software Tools*, 2000.

[16] J. Wu and Z.-H. Zhou, "Face recognition with one training image per person," *Pattern Recognition Letters*, vol. 49, no. 14, pp. 1711–1719, 2002.

[17] D. Beymer and T. Poggio, "Face recognition from one example view," in *Proceedings of the 5th International Conference on Computer Vision*, 1995, pp. 500–507.

[18] H. Grabner, C. Leistner, and H. Bischof, "Semi-supervised on-line boosting for robust tracking," in *Proceedings of the 10th European Conference on Computer Vision*, 2008, pp. 234–247.

[19] A. Goldberg, M. Li, and X. Zhu, "Online manifold regularization: A new learning setting and empirical study," in *Proceeding of European Conference on Machine Learning and Principles and Practice of Knowledge Discovery in Databases*, 2008.