From Face Recognition to Models of Identity: A Bayesian Approach to Learning about Unknown Identities from Unsupervised Data

Daniel Coelho de Castro[1,⋆] and Sebastian Nowozin[2]

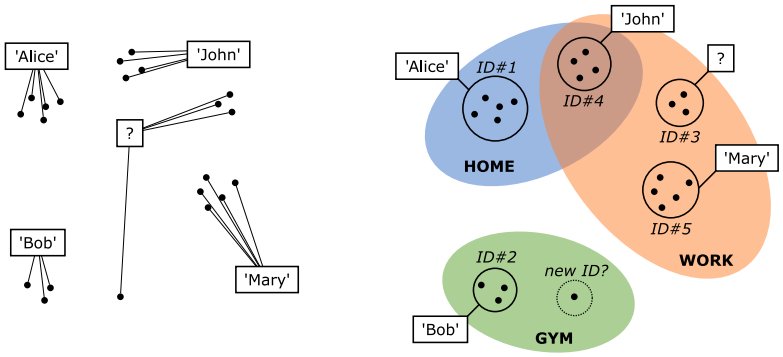
[1] Imperial College London, UK
dc315@imperial.ac.uk
[2] Microsoft Research, Cambridge, UK
Sebastian.Nowozin@microsoft.com

Abstract. Current face recognition systems robustly recognize identities across a wide variety of imaging conditions. In these systems recognition is performed via classification into known identities obtained from supervised identity annotations. There are two problems with this current paradigm: (1) current systems are unable to benefit from unlabelled data which may be available in large quantities; and (2) current systems equate successful recognition with labelling a given input image. Humans, on the other hand, regularly perform identification of individuals completely unsupervised, recognising the identity of someone they have seen before even without being able to name that individual. How can we go beyond the current classification paradigm towards a more human understanding of identities? We propose an integrated Bayesian model that coherently reasons about the observed images, identities, partial knowledge about names, and the situational context of each observation. While our model achieves good recognition performance against known identities, it can also discover new identities from unsupervised data and learns to associate identities with different contexts depending on which identities tend to be observed together. In addition, the proposed semi-supervised component is able to handle not only acquaintances, whose names are known, but also unlabelled familiar faces and complete strangers in a unified framework.

1 Introduction

For the following discussion, we decompose the usual face identification task into two sub-problems: *recognition* and *tagging*. Here we understand recognition as the unsupervised task of matching an observed face to a cluster of previously seen faces with similar appearance (disregarding variations in pose, illumination etc.), which we refer to as an *identity*. Humans routinely operate at this level of abstraction to recognise familiar faces: even when people's names are not known, we can still tell them apart. Tagging, on the other hand, refers to putting names to faces, i.e. associating string literals to known identities.

⋆ Work done during an internship at Microsoft Research.



(a) Standard face recognition (b) Context-aware model of identities

Fig. 1: Face recognition settings. Points represent face observations and boxes are name labels.

Humans tend to create an inductive mental model of facial appearance for each person we meet, which we then query at new encounters to be able to recognise them. This is opposed to a transductive approach, attempting to match faces to specific instances from a memorised gallery of past face observations—which is how identification systems are often implemented [16].

An alternative way to represent faces, aligned with our inductive recognition, is via *generative* face models, which explicitly separate latent identity content, tied across all pictures of a same individual, from nuisance factors such as pose, expression and illumination [15,21,18]. While mostly limited to linear projections from pixel space (or mixtures thereof), the probabilistic framework applied in these works allowed tackling a variety of face recognition tasks, such as closed- and open-set identification, verification and clustering.

A further important aspect of social interactions is that, as an individual continues to observe faces every day, they encounter some people much more often than others, and the total number of distinct identities ever met tends to increase virtually without bounds. Additionally, we argue that human face recognition does not happen in an isolated environment, but situational contexts (e.g. 'home', 'work', 'gym') constitute strong cues for the groups of people a person expects to meet (Fig. 1b).

With regards to tagging, in daily life we very rarely obtain named face observations: acquaintances normally introduce themselves only once, and not repeatedly whenever they are in our field of view. In other words, humans are naturally capable of semi-supervised learning, generalising sparse name annotations to all observations of the corresponding individuals, while additionally reconciling naming conflicts due to noise and uncertainty.

In contrast, standard computational face identification is *fully supervised* (see Fig. 1a), relying on vast labelled databases of high-quality images [1]. Although many supervised methods achieve astonishing accuracy on challenging benchmarks (e.g. [26,25]) and are successfully employed in practical biometric applications, this setting has arguably limited analogy to human social experience.

Expanding on the generative perspective, we introduce a unified Bayesian model which reflects all the above considerations on identity distributions, context-awareness and labelling (Fig. 1b). Our nonparametric identity model effectively represents an unbounded population of identities, while taking contextual co-occurrence relations into account and exploiting modern deep face representations to overcome limitations of previous linear generative models. Our main contributions in this work are twofold:

1. We propose an unsupervised face recognition model which can explicitly reason about people it has never seen; and
2. We attach to it a novel robust label model enabling it to predict names by learning from both named and unnamed faces.

Related Work

Other face recognition methods (even those formulated in a Bayesian framework) [32,33,9,27,17], often limit themselves to point estimates of parameters and predictions, occasionally including ad-hoc confidence metrics. A distinct advantage of our approach is that it is probabilistic end-to-end, and thus naturally provides predictions with principled, quantifiable uncertainties. Moreover, we employ modern Bayesian modelling tools—namely hierarchical nonparametrics—which enable dynamically adapting model complexity while faithfully reflecting the real-world assumptions laid out above.

Secondly, although automatic face tagging is a very common task, each problem setting can impose wildly different assumptions and constraints. Typical application domains involve the annotation of personal photo galleries [32,33,3,12], multimedia (e.g. TV) [27,17] or security/surveillance [16]. Our work focuses on egocentric human-like face recognition, a setting which seems largely unexplored, as most of the work using first-person footage appears to revolve around other tasks like object and activity recognition, face detection, and tracking [4]. As we explained previously, the dynamic, *online* nature of first-person social experience brings a number of specific modelling challenges for face recognition.

Finally, while there is substantial prior work on using contexts to assist face recognition, we emphasize that much (perhaps most) of it is effectively complementary to our unified framework. Notions of *global* context such as timestamp, geolocation and image background [30,33,3,9] can readily be used to inform our current context model (Section 2.1). In addition, we can naturally augment the proposed face model (Section 2.3) to leverage further *individual* context features, e.g. clothing and speech [33,3,27,17]. Integration of these additional factors opens exciting avenues for future research.

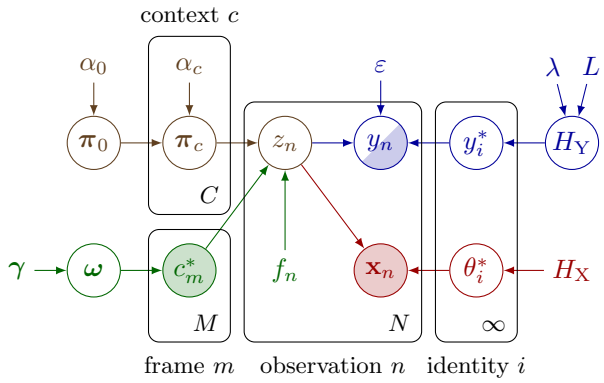


Fig. 2: Overview of the proposed generative model, encompassing the **context model**, **identity model**, **face model** and **label model**. Unfilled nodes represent latent variables, shaded nodes are observed, the half-shaded node is observed only for a subset of the indices and uncircled nodes denote fixed hyperparameters. $\boldsymbol{\pi}_0$ and $(\boldsymbol{\pi}_c)_{c=1}^C$ are the global and context-wise identity probabilities, $\boldsymbol{\omega}$ denotes the context probabilities, $(c_m^*)_{m=1}^M$ are the frame-wise context labels, indexed by the frame numbers $(f_n)_{n=1}^N$, $(z_n)_{n=1}^N$ are the latent identity indicators, $(\mathbf{x}_n)_{n=1}^N$ are the face observations and $(y_n)_{n=1}^N$ are the respective name annotations, $(\theta_i^*)_{i=1}^\infty$ are the parameters of the face model and $(y_i^*)_{i=1}^\infty$ are the identities' name labels. See text for descriptions of the remaining symbols.

2 A Model of Identities

In this section, we describe in isolation each of the building blocks of the proposed approach to facial identity recognition: the context model, the identity model and the face model. We assume data is collected in the form of camera *frames* (either photographs or a video stills), numbered 1 to M, and faces are cropped with some face detection system and grouped by frame number indicators, $f_n \in \{1, \ldots, M\}$. The diagram in Fig. 2 illustrates the full proposed graphical model, including the label model detailed in Section 3.

2.1 Context Model

In our identity recognition scenario, we imagine the user moving between contexts throughout the day (e.g. home–work–gym...). Since humans naturally use situational context as a strong prior on the groups of people we expect to encounter in each situation, we incorporate context-awareness in our model of identities to mimic human-like face recognition.

The context model we propose involves a categorical variable $c_n \in \{1, \ldots, C\}$ for each observation, where C is some fixed number of distinct contexts.[3] Cru-

[3] See footnote 4.

cially, we assume that all observations in frame m, $\mathcal{F}_m = \{n : f_n = m\}$, share the same context, c^*_m (i.e. $\forall n, c_n = c^*_{f_n}$).

We define the identity indicators to be independent given the context of the corresponding frames (see Section 2.2, below). However, since the contexts are tied by frame, marginalising over the contexts captures identity co-occurrence relations. In turn, these allow the model to make more confident predictions about people who tend to be seen together in the same environment.

This formalisation of contexts as discrete semantic labels is closely related to the place recognition model in [30], used there to disambiguate predictions for object detection. It has also been demonstrated that explicit incorporation of a context variable can greatly improve clustering with mixture models [19].

Finally, we assume the context indicators c^*_m are independently distributed according to probabilities $\boldsymbol{\omega}$, which themselves follow a Dirichlet prior:

$$\boldsymbol{\omega} \sim \mathrm{Dir}(\boldsymbol{\gamma}) \tag{1}$$

$$c^*_m \mid \boldsymbol{\omega} \sim \mathrm{Cat}(\boldsymbol{\omega})\,, \qquad m = 1, \ldots, M\,, \tag{2}$$

where M is the total number of frames. In our simulation experiments, we use a symmetric Dirichlet prior, setting $\boldsymbol{\gamma} = (\gamma_0/C, \ldots, \gamma_0/C)$.

2.2 Identity Model

In the daily-life scenario described in Section 1, an increasing number of unique identities will tend to appear as more faces are observed. This number is expected to grow much more slowly than the number of observations, and can be considered unbounded in practice (we do not expect a user to run out of new people to meet). Moreover, we can expect some people to be encountered much more often than others. Since a Dirichlet process (DP) [10] displays properties that mirror all of the above phenomena [28], it is a sound choice for modelling the distribution of identities.

Furthermore, the assumption that all people can potentially be encountered in any context, but with different probabilities, is perfectly captured by a hierarchical Dirichlet process (HDP) [29]. Making use of the context model, we define one DP *per context* c, each with concentration parameter α_c and sharing the same *global* DP as a base measure.[4] This hierarchical construction thus produces context-specific distributions over a common set of identities.

We consider that each of the N face detections is associated to a latent identity indicator variable, z_n. We can write the generative process as

$$\boldsymbol{\pi}_0 \sim \mathrm{GEM}(\alpha_0) \tag{3}$$

$$\boldsymbol{\pi}_c \mid \boldsymbol{\pi}_0 \sim \mathrm{DP}(\alpha_c, \boldsymbol{\pi}_0)\,, \qquad c = 1, \ldots, C \tag{4}$$

$$z_n \mid f_n = m, \mathbf{c}^*, (\boldsymbol{\pi}_c)_c \sim \mathrm{Cat}(\boldsymbol{\pi}_{c^*_m})\,, \qquad n = 1, \ldots, N\,, \tag{5}$$

[4] One could further allow an unbounded number of latent contexts by incorporating a nonparametric context distribution, resulting in a structure akin to the nested DP [23,5] or the dual DP described in [31]. More details in the online supplement, Sec. A.

where $\mathrm{GEM}(\alpha_0)$ is the DP stick-breaking distribution, $\pi_{0i} = \beta_i \prod_{j=1}^{i-1}(1-\beta_j)$, with $\beta_i \sim \mathrm{Beta}(1, \alpha_0)$ and $i = 1, \ldots, \infty$. Here, $\boldsymbol{\pi}_0$ is the global identity distribution and $(\boldsymbol{\pi}_c)_{c=1}^{C}$ are the context-specific identity distributions.

Although the full generative model involves infinite-dimensional objects, DP-based models present simple finite-dimensional marginals. In particular, the posterior predictive probability of encountering a known identity i is

$$p(z_{N+1} = i \mid c_{N+1} = c, \mathbf{z}, \mathbf{c}^*, \boldsymbol{\pi}_0) = \frac{\alpha_c \pi_{0i} + N_{ci}}{\alpha_c + N_{c\cdot}}\,, \tag{6}$$

where N_{ci} is the number of observations assigned to context c and identity i and $N_{c\cdot}$ is the total number of observations in context c.

Finally, such a nonparametric model is well suited for an open-set identification task, as it can elegantly estimate the prior probability of encountering an unknown identity:

$$p(z_{N+1} = I+1 \mid c_{N+1} = c, \mathbf{z}, \mathbf{c}^*, \boldsymbol{\pi}_0) = \frac{\alpha_c \pi'_0}{\alpha_c + N_{c\cdot}}\,, \tag{7}$$

where I is the current number of distinct known identities and $\pi'_0 = \sum_{i=I+1}^{\infty} \pi_{0i}$ denotes the global probability of sampling a new identity.

2.3 Face Model

In face recognition applications, it is typically more convenient and meaningful to extract a compact representation of face features than to work directly in a high-dimensional pixel space.

We assume that the observed features of the n^{th} face, $\mathbf{x}_n$, arise from a parametric family of distributions, F_{X}. The parameters of this distribution, θ_i^*, drawn from a prior, H_{X}, are unique for each identity and are shared across all face feature observations of the same person:

$$\theta_i^* \sim H_{\mathrm{X}}\,, \qquad i = 1, \ldots, \infty \tag{8}$$

$$\mathbf{x}_n \mid z_n, \boldsymbol{\theta}^* \sim F_{\mathrm{X}}(\theta_{z_n}^*)\,, \qquad n = 1, \ldots, N\,. \tag{9}$$

As a consequence, the marginal distribution of faces is given by a *mixture model*: $p(\mathbf{x}_n \mid c_n = c, \boldsymbol{\theta}^*, \boldsymbol{\pi}_c) = \sum_{i=1}^{\infty} \pi_{ci} F_{\mathrm{X}}(\mathbf{x}_n \mid \theta_i^*)$.

In the experiments reported in this paper, we used the 128-dimensional embeddings produced by OpenFace, a publicly available, state-of-the-art neural network for face recognition [2], implementing FaceNet's architecture and methodology [25]. In practice, this could easily be swapped for other face embeddings (e.g. DeepFace [26]) without affecting the remainder of the model. We chose isotropic Gaussian mixture components for the face features (F_{X}), with an empirical Gaussian–inverse gamma prior for their means and variances (H_{X}).

3 Robust Semi-Supervised Label Model

We expect to work with only a small number of labelled observations manually provided by the user. Since the final goal is to identify any observed face, our probabilistic model needs to incorporate a semi-supervised aspect, generalising the sparse given labels to unlabelled instances. Throughout this section, the terms 'identity' and 'cluster' will be used interchangeably.

One of the cornerstones of semi-supervised learning (SSL) is the premise that clustered items tend to belong to the same class [8, §1.2.2]. Building on this *cluster assumption*, mixture models, such as ours, have been successfully applied to SSL tasks [6]. We illustrate in Fig. 3 our proposed label model detailed below, comparing it qualitatively to nearest-neighbour classification on a toy example.

With the motivation above, we attach a label variable (a *name*) to each cluster (identity), here denoted y_i^*. This notation suggests that there is a single true label $\tilde{y}_n = y_{z_n}^*$ for each observation n, analogously to the observation parameters: $\theta_n = \theta_{z_n}^*$. Finally, the observed labels, y_n, are potentially corrupted through some noise process, F_{Y}. Let $\mathcal{L}$ denote the set of indices of the labelled data. The complete generative process is presented below:

$$H_{\mathrm{Y}} \sim \mathrm{DP}(\lambda, L) \tag{10}$$

$$y_i^* \mid H_{\mathrm{Y}} \sim H_{\mathrm{Y}}\,, \qquad i = 1, \ldots, \infty \tag{11}$$

$$y_n \mid z_n, \mathbf{y}^* \sim F_{\mathrm{Y}}(y_{z_n}^*)\,, \qquad n \in \mathcal{L}\,. \tag{12}$$

As mentioned previously, a related model for mixture model-based SSL with noisy labels was proposed in [6]. Instead of considering an explicit noise model for the class labels, the authors of that work model directly the conditional label distribution for each cluster. Our setting here is more general: we assume not only an unbounded number of clusters, but also of possible labels.

3.1 Label Prior

We assume that the number of distinct labels will tend to increase without bounds as more data is observed. Therefore, we adopt a further nonparametric prior on the cluster-wide labels:

$$H_{\mathrm{Y}} \sim \mathrm{DP}(\lambda, L)\,, \tag{13}$$

where L is some base probability distribution over the countable but unbounded label space (e.g. strings).[5] We briefly discuss the choice of L further below.

All concrete knowledge we have about the random label prior H_{Y} comes from the set of observed labels, $\mathbf{y}_{\mathcal{L}}$. Crucially, if we marginalise out H_{Y}, the predictive label distribution is simply [28]

$$y_{I+1}^* \mid \mathbf{y}^* \sim \frac{1}{\lambda + I}\left(\lambda L + \sum_{\ell \in \mathcal{Y}} J_\ell \delta_\ell\right), \tag{14}$$

[5] One could instead consider a Pitman–Yor process if power-law behaviour seems more appropriate than the DP's exponential tails [20].

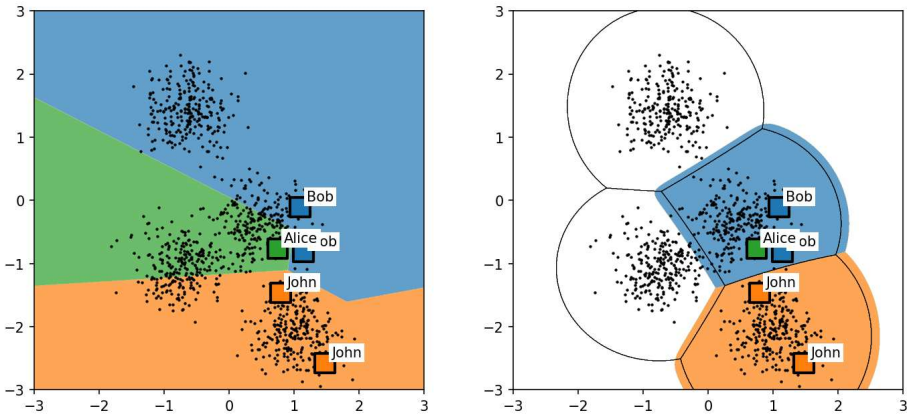


Fig. 3: Hard label predictions of the proposed semi-supervised label model (right) and nearest-neighbour classification (left). Points represent unlabelled face observations, squares are labelled and the black contours on the right show identity boundaries. The proposed label model produces more *natural* boundaries, assigning the 'unknown' label (white) to unlabelled clusters and regions distant from any observed cluster, while also accommodating label noise ('Bob' → 'Alice') without the spurious boundaries introduced by NN.

which we will denote $\widehat{H_{\mathrm{Y}}}(y^*_{I+1} \mid \mathbf{y}^*)$. Here, $\mathcal{Y}$ is the set of distinct known labels among $\mathbf{y}_{\mathcal{L}}$ and $J_\ell = |\{i : y^*_i = \ell\}|$, the number of components with label ℓ (note that $\sum_\ell J_\ell = I$).

In addition to allowing multiple clusters to have repeated labels, this formulation allows us to reason about *unseen* labels. For instance, some of the learned clusters may have no labelled training points assigned to them, and the true (unobserved) labels of those clusters may never have been encountered among the training labels. Another situation in which unseen labels come into play is with points away from any clusters, for which the identity model would allocate a new cluster with high probability. In both cases, this model gives us a principled estimate of the probability of assigning a special 'unknown' label.

The base measure L may be defined over a rudimentary language model. For this work, we adopted a geometric/negative binomial model for the string length $|\ell|$, with characters drawn uniformly from an alphabet of size K:

$$L_{\phi,K}(\ell) = \mathrm{Geom}(|\ell|; \tfrac{1}{\phi})\,\mathrm{Unif}(\ell; K^{|\ell|}) = \frac{1}{\phi - 1}\left(\frac{\phi - 1}{\phi K}\right)^{|\ell|}, \tag{15}$$

where ϕ is the expected string length.

3.2 Label Likelihood

In the simplest case, we could consider $F_{\mathrm{Y}}(\cdot) = \delta_\cdot$, i.e. noiseless labels. Although straightforward to interpret and implement, this could make inference highly

unstable whenever there would be conflicting labels for an identity. Moreover, in our application, the labels would be provided provided by a human user who may not have perfect knowledge of the target person's true name or its spelling, for example.

Therefore, we incorporate a label noise model, which can gracefully handle conflicts and mislabelling. We assume observed labels are noisy completely at random (NCAR) [11, §II-C], with a fixed error rate ε:[6]

$$\widehat{F_{\mathrm{Y}}}(\ell \mid y_i^*; \mathbf{y}^*) = \begin{cases} 1-\varepsilon\,, & \ell = y_i^* \\ \varepsilon \frac{\widehat{H_{\mathrm{Y}}}(\ell \mid \mathbf{y}^*)}{1-\widehat{H_{\mathrm{Y}}}(y_i^* \mid \mathbf{y}^*)}\,, & \ell \neq y_i^* \end{cases}. \tag{16}$$

Intuitively, an observed label, y_n, agrees with its identity's assigned label, $y^*_{z_n}$, with probability $1-\varepsilon$. Otherwise, it is assumed to come from a modified label distribution, in which we restrict and renormalise $\widehat{H_{\mathrm{Y}}}$ to exclude $y^*_{z_n}$. Here we use $\widehat{H_{\mathrm{Y}}}$ in the error distribution instead of L to reflect that a user is likely to mistake a person's name for another known name, rather than for a completely random string.

3.3 Label Prediction

For label prediction, we are only concerned with the true, noiseless labels, $\tilde{y}_n$. The predictive distribution for a single new sample is given by

$$\begin{aligned} &p(\tilde{y}_{N+1} = \ell \mid \mathbf{x}_{N+1}, \mathbf{z}, \mathbf{c}^*, \mathbf{y}^*, \boldsymbol{\theta}^*, \boldsymbol{\pi}_0) \\ &\quad = \sum_{i \leq I : y_i^* = \ell} p(z_{N+1} = i \mid \mathbf{x}_{N+1}, \mathbf{z}, \mathbf{c}^*, \boldsymbol{\theta}^*, \boldsymbol{\pi}_0) \\ &\qquad + \widehat{H_{\mathrm{Y}}}(y^*_{I+1} = \ell \mid \mathbf{y}^*)\, p(z_{N+1} = I+1 \mid \mathbf{x}_{N+1}, \mathbf{z}, \mathbf{c}^*, \boldsymbol{\theta}^*, \boldsymbol{\pi}_0)\,. \end{aligned} \tag{17}$$

The sum in the first term is the probability of the sample being assigned to any of the existing identities that have label ℓ, while the last term is the probability of instantiating a new identity with that label.

4 Evaluation

One of the main strengths of the proposed model is that it creates a single rich representation of the known world, which can then be queried from various angles to obtain distinct insights. In this spirit, we designed three experimental setups to assess different properties of the model: detecting whether a person has been seen before (outlier detection), recognising faces as different identities in a sequence of frames (clustering, unsupervised) and correctly naming observed faces by generalising sparse user annotations (semi-supervised learning).

[6] The 'true' label likelihood $F_{\mathrm{Y}}(\ell \mid y_i^*)$ is random due to its dependence on the unobserved prior H_{Y}. We thus define $\widehat{F_{\mathrm{Y}}}$ as its posterior expectation given the known identity labels $\mathbf{y}^*$. See supplementary material, Sec. B, for details.

In all experiments, we used celebrity photographs from the Labelled Faces in the Wild (LFW) database [13].[7] We have implemented inference via Gibbs Markov chain Monte Carlo (MCMC) sampling, whose conditional distributions can be found in the supplementary material (Sec. C), and we run multiple chains with randomised initial conditions to better estimate the variability in the posterior distribution. For all metrics evaluated on our model, we report the estimated 95% highest posterior density (HPD) credible intervals over pooled samples from 8 independent Gibbs chains, unless stated otherwise.

4.1 Experiment 1: Unknown Person Detection

In our first set of experiments, we study the model's ability to determine whether or not a person has been seen before. This key feature of the proposed model is evaluated based on the probability of an observed face not corresponding to any of the known identities, as given by Eq. (7). In order to evaluate purely the detection of unrecognised faces, we constrained the model to a single context ($C = 1$) and set aside the label model ($\mathcal{L} = \emptyset$).

This task is closely related to outlier/anomaly detection. In particular, our proposed approach mirrors one of its common formulations, involving a mixture of a 'normal' distribution, typically fitted to some training data, and a flatter 'anomalous' distribution[8] [7, §7.1.3].

We selected the 19 celebrities with at least 40 pictures available in LFW and randomly split them in two groups: 10 known and 9 unknown people. We used 27 images of each of the *known* people as training data and a disjoint test set of 13 images of each of the *known* and *unknown* people. We therefore have a binary classification setting with well-balanced classes at test time. Here, we ran our Gibbs sampler for 500 steps, discarding the first 100 burn-in iterations and thinning by a factor of 10, resulting in 320 pooled samples.

In Fig. 4a, we visualise the agreements between maximum *a posteriori* (MAP) identity predictions for test images:

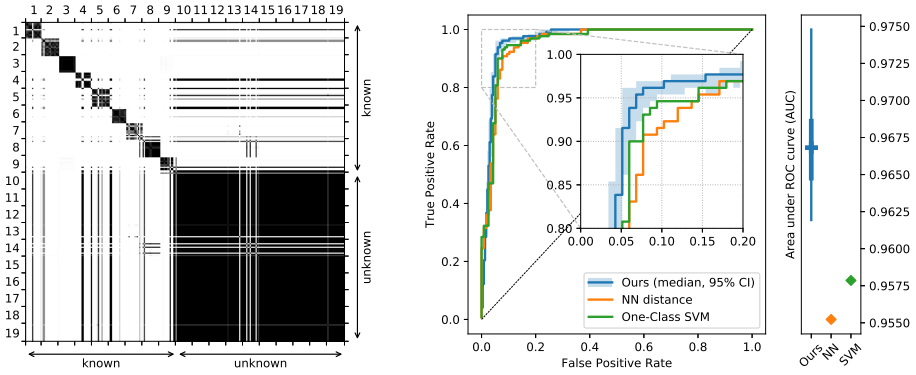
$$\hat{z}_n = \arg\max_i p(z_n = i \mid \mathbf{x}_n, \mathbf{z}, \mathbf{c}^*, \boldsymbol{\pi}_0, \boldsymbol{\theta}^*)\,, \tag{18}$$

where i ranges from 1 to $I + 1$, the latter indicating an *unknown* identity, absent from the training set, and n indexes the test instances. Despite occasional ambiguous cases, the proposed model seems able to consistently group together all unknown faces, while successfully distinguishing between known identities.

As a simple baseline detector for comparison, we consider a threshold on the distance to the nearest neighbour (NN) in the face feature space [7, §5.1]. We also evaluate the decision function of a one-class SVM [24], using an RBF kernel with $\gamma = 10$, chosen via leave-one-person-out cross-validation on the training set (roughly equivalent to thresholding the training data's kernel density estimate with bandwidth $1/\sqrt{2\gamma} \approx 0.22$). We compare the effectiveness of both detection approaches using ROC curve analysis.

[7] Available at: http://vis-www.cs.umass.edu/lfw/

[8] The predictive distribution of $\mathbf{x}_n$ for new identities is a wide Student's t.



(a) Association matrix, counting agreements in the MAP identity predictions (including *unknown*). Ticks delimit ground-truth identities.

(b) ROC analysis for unknown person detection compared with baselines. AUC is shown with median and 50% and 95% HPD intervals.

Fig. 4: Results of the unknown person detection experiment on test images

Figure 4b shows that, while all methods are highly effective at detecting unknown faces, scoring 95%+ AUC, ours consistently outperforms, by a small margin, both the NN baseline and the purpose-designed one-class SVM. Taking the MAP prediction, our model achieves $[92.3\%, 94.3\%]$ detection accuracy.

4.2 Experiment 2: Identity Discovery

We then investigate the clustering properties of the model in a purely unsupervised setting, when only context is provided. We evaluate the consistency of the estimated partitions of images into identities with the ground truth in terms of the adjusted Rand index [22,14].

Using simulations, besides having an endless source of data with ground-truth context and identity labels, we have full control over several important aspects of experimental setup, such as sequence lengths, rates of encounters, numbers of distinct contexts and people and amount of provided labels. Below we describe the simulation algorithm used in our experiments and illustrated in Fig. 5.

In our experiments we aim to simulate two important aspects of real-world identity recognition settings: 1. *Context*: knowing the context (e.g. location or time) makes it more likely for us to observe a particular subset of people; and 2. *Temporal consistency*: identities will not appear and disappear at random but instead be present for a longer duration.

To reproduce contexts, we simulate a single session of a user meeting new people. To this end we first create a number of fixed contexts and then assign identities uniformly at random to each context. For these experiments, we defined three contexts: ‘home’, ‘work’ and ‘gym’. At any time, the user knows its own

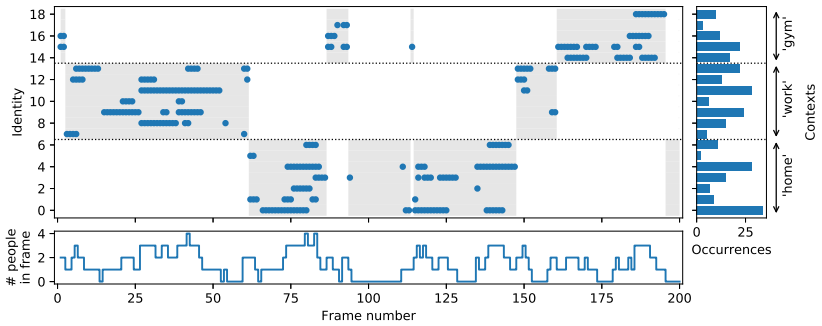


Fig. 5: The simulation used in Experiment 2, showing identities coming in and out of the camera frame. Identities are shown grouped by their context (far right), and shading indicates identities present in the user's current context.

context and over time transitions between contexts. Independently at each frame, the user may switch context with a small probability.

To simulate temporal consistency, each person in the current context enters and leaves the camera frame as an independent binary Markov chain. As shown in Fig. 5 this naturally produces grouped observations. The image that is observed for each 'detected' face is sampled from the person's pictures available in the database. We sample these images without replacement and in cycles, to avoid observing the same image consecutively.

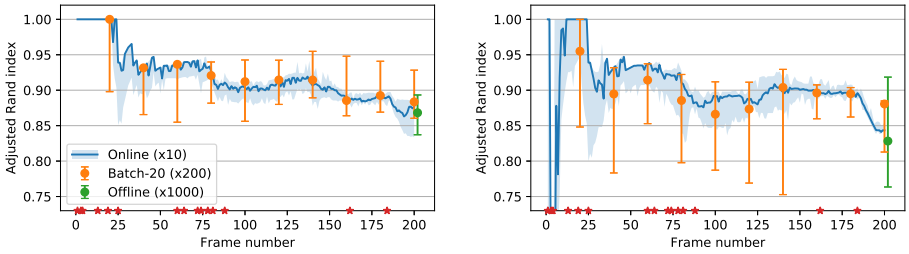
For this set of experiments, we consider three practical scenarios:

- *Online:* data is processed on a frame-by-frame basis, i.e. we extend the training set after each frame and run the Gibbs sampler for 10 full iterations
- *Batch:* same as above, but enqueue data for 20 frames before extending the training set and updating the model for 200 steps
- *Offline:* assume entire sequence is available at once and iterate for 1000 steps

In the interest of fairness, the number of steps for each protocol was selected to give them roughly the same overall computation budget (ca. 200 000 frame-wise steps). In addition, we also study the impact on recognition performance of disabling the context model, by setting $C = 1$ and $c^*_m = 1, \forall m$.

We show the results of this experiment in Fig. 6. Clearly it is expected that, as more identities are met over time, the problem grows more challenging and clustering performance tends to decrease. Another general observation is that online processing produced much lower variance than batch or offline in both cases. The incremental availability of training data therefore seems to lead to more coherent states of the model.

Now, comparing Figs. 6a and 6b, it is evident that context-awareness not only reduces variance but also shows marginal improvements over the context-oblivious variant. Thus, without hurting recognition performance, the addition of a context model enables the *prediction* of context at test time, which may be useful for downstream user-experience systems.



(a) With contexts (b) Without contexts

Fig. 6: Identity clustering consistency. Markers on the horizontal axis (★) indicate when new people are met for the first time.

4.3 Experiment 3: Semi-Supervised Labelling

In our final set of experiments, we aimed to validate the application of the proposed label model for semi-supervised learning with sparse labels.

In the context of face identification, we may define three groups of people:

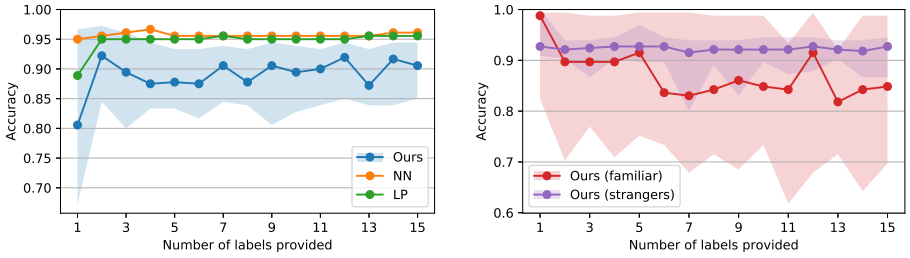
- *Acquainted:* known identity with known name
- *Familiar:* known identity with unknown name
- *Stranger:* unknown identity

We thus selected the 34 LFW celebrities with more than 30 pictures, and split them roughly equally in these three categories at random. From the *acquainted* and *familiar* groups, we randomly picked 15 of their images for training and 15 for testing, and we used 15 pictures of each *stranger* at test time only. We evaluated the label prediction accuracy as we varied the number of labelled training images provided for each acquaintance, from 1 to 15.

For baseline comparison, we evaluate nearest-neighbour classification (NN) and label propagation (LP) [34], a similarity graph-based semi-supervised algorithm. We computed the LP edge weights with the same kernel as the SVM in Section 4.1. Recall that the face embedding network was trained with a triplet loss to explicitly optimise Euclidean distances for classification [2]. As both NN and LP are distance-based, they are therefore expected to hold an advantage over our model for classifying labelled identities.

Figure 7a shows the label prediction results for the labelled identities (acquaintances). In this setting, NN and LP performed nearly identically and slightly better than ours, likely due to the favourable embedding structure. Moreover, all methods predictably become more accurate as more supervision is introduced in the training data.

More importantly, the key distinctive capabilities of our model are demonstrated in Fig. 7b. As already discussed in Section 4.1, the proposed model is capable of detecting complete strangers, and here we see that it correctly predicts that their name is unknown. Furthermore, our model can acknowledge that



(a) Acquaintances (b) Familiar and strangers

Fig. 7: Label prediction accuracy. Note that NN and LP effectively have null accuracy for the *familiar* and *strangers* groups, as they cannot predict 'unknown'.

familiar faces belong to different people, whose names may not be known. Neither of these functionalities is provided by the baselines, as they are limited to the closed-set identification task.

5 Conclusion

In this work, we introduced a fully Bayesian treatment of the face identification problem. Each component of our proposed approach was motivated from human intuition about face recognition and tagging in daily social interactions. Our principled identity model can contemplate an unbounded population of identities, accounting for context-specific probabilities of meeting them.

We demonstrated that the proposed identity model can accurately detect when a face is unfamiliar, and is able to incrementally learn to differentiate between new people as they are met in a streaming data scenario. Lastly, we verified that our approach to dealing with sparse name annotations can handle not only acquaintances, whose names are known, but also familiar faces and complete strangers in a unified manner—a functionality unavailable in conventional (semi-) supervised identification methods.

Here we considered a fully supervised context structure. As mentioned in Section 1, one could imagine an unsupervised approach involving global visual or non-visual signals to drive context inference (e.g. global image features, time or GPS coordinates), in addition to extensions to the face model with individual context information (e.g. clothing, speech). Yet another interesting research direction is to explicitly consider time dependence, e.g. by endowing the sequence of latent contexts with a hidden Markov model-like structure [30].

Acknowledgement. This work was partly supported by CAPES, Brazil (BEX 1500/2015-05).

References

1. Labeled faces in the wild: A survey. In: Kawulok, M., Celebi, M.E., Smolka, B. (eds.) Advances in Face Detection and Facial Image Analysis, pp. 189–248. Springer (2016). https://doi.org/10.1007/978-3-319-25958-1
2. Amos, B., Ludwiczuk, B., Satyanarayanan, M.: OpenFace: A general-purpose face recognition library with mobile applications. Tech. Rep. CMU-CS-16-118, CMU School of Computer Science (2016)
3. Anguelov, D., Lee, K.c., Gökturk, S.B., Sumengen, B.: Contextual identity recognition in personal photo albums. In: CVPR 2007. pp. 1–7 (2007). https://doi.org/10.1109/CVPR.2007.383057
4. Betancourt, A., Morerio, P., Regazzoni, C.S., Rauterberg, M.: The evolution of first person vision methods: A survey. IEEE Transactions on Circuits and Systems for Video Technology **25**(5), 744–760 (may 2015). https://doi.org/10.1109/TCSVT.2015.2409731
5. Blei, D.M., Griffiths, T.L., Jordan, M.I.: The nested chinese restaurant process and bayesian nonparametric inference of topic hierarchies. Journal of the ACM **57**(2) (jan 2010). https://doi.org/10.1145/1667053.1667056
6. Bouveyron, C., Girard, S.: Robust supervised classification with mixture models: Learning from data with uncertain labels. Pattern Recognition **42**(11), 2649–2658 (2009). https://doi.org/10.1016/j.patcog.2009.03.027
7. Chandola, V., Banerjee, A., Kumar, V.: Anomaly detection: A survey. ACM Computing Surveys **41**(3), 1–58 (jul 2009). https://doi.org/10.1145/1541880.1541882
8. Chapelle, O., Schölkopf, B., Zien, A. (eds.): Semi-Supervised Learning. MIT Press (2006)
9. Choi, J.Y., De Neve, W., Ro, Y.M., Plataniotis, K.: Automatic face annotation in personal photo collections using context-based unsupervised clustering and face information fusion. IEEE Transactions on Circuits and Systems for Video Technology **20**(10), 1292–1309 (oct 2010). https://doi.org/10.1109/TCSVT.2010.2058470
10. Ferguson, T.S.: A bayesian analysis of some nonparametric problems. The Annals of Statistics **1**(2), 209–230 (1973), http://www.jstor.org/stable/2958008
11. Frénay, B., Verleysen, M.: Classification in the presence of label noise: A survey. IEEE Transactions on Neural Networks and Learning Systems **25**(5), 845–869 (2014). https://doi.org/10.1109/TNNLS.2013.2292894
12. Gallagher, A.C., Chen, T.: Using context to recognize people in consumer images. IPSJ Transactions on Computer Vision and Applications **1**, 115–126 (2009). https://doi.org/10.2197/ipsjtcva.1.115
13. Huang, G.B., Ramesh, M., Berg, T., Learned-Miller, E.: Labeled Faces in the Wild: A database for studying face recognition in unconstrained environments. Tech. rep., University of Massachusetts Amherst (2007)
14. Hubert, L., Arabie, P.: Comparing partitions. Journal of Classification **2**(1), 193–218 (dec 1985). https://doi.org/10.1007/BF01908075
15. Ioffe, S.: Probabilistic linear discriminant analysis. In: Computer Vision – ECCV 2006. vol. 3954 LNCS, pp. 531–542 (2006). https://doi.org/10.1007/11744085_41
16. Jafri, R., Arabnia, H.R.: A survey of face recognition techniques. Journal of Information Processing Systems **5**(2), 41–68 (2009). https://doi.org/10.3745/JIPS.2009.5.2.041
17. Le, N., Bredin, H., Sargent, G., India, M., Lopez-Otero, P., Barras, C., Guinaudeau, C., Gravier, G., da Fonseca, G.B., Freire, I.L., Patrocínio, Jr, Z., Guimarães, S.J.F., Martí, G., Morros, J.R., Hernando, J., Docio-Fernandez, L., Garcia-Mateo, C.,

Meignier, S., Odobez, J.M.: Towards large scale multimedia indexing: A case study on person discovery in broadcast news. In: CBMI 2017. pp. 18:1–18:6. ACM (2017). https://doi.org/10.1145/3095713.3095732

18. Li, P., Fu, Y., Mohammed, U., Elder, J.H., Prince, S.J.D.: Probabilistic models for inference about identity. IEEE Transactions on Pattern Analysis and Machine Intelligence **34**(1), 144–157 (2012). https://doi.org/10.1109/TPAMI.2011.104
19. Perdikis, S., Leeb, R., Chavarriaga, R., Millán, J.d.R.: Context-aware learning for finite mixture models (2015), http://arxiv.org/abs/1507.08272
20. Pitman, J., Yor, M.: The two-parameter Poisson–Dirichlet distribution derived from a stable subordinator. The Annals of Probability **25**(2), 855–900 (1997), http://www.jstor.org/stable/2959614
21. Prince, S.J., Elder, J.H.: Probabilistic linear discriminant analysis for inferences about identity. In: Proceedings of the 11th IEEE International Conference on Computer Vision (ICCV 2007). IEEE (2007). https://doi.org/10.1109/ICCV.2007.4409052
22. Rand, W.M.: Objective criteria for the evaluation of clustering methods. Journal of the American Statistical Association **66**(336), 846–850 (dec 1971). https://doi.org/10.1080/01621459.1971.10482356
23. Rodríguez, A., Dunson, D.B., Gelfand, A.E.: The nested dirichlet process. Journal of the American Statistical Association **103**(483), 1131–1154 (sep 2008). https://doi.org/10.1198/016214508000000553
24. Schölkopf, B., Platt, J.C., Shawe-Taylor, J., Smola, A.J., Williamson, R.C.: Estimating the support of a high-dimensional distribution. Neural Computation **13**(7), 1443–1471 (jul 2001). https://doi.org/10.1162/089976601750264965
25. Schroff, F., Kalenichenko, D., Philbin, J.: FaceNet: A unified embedding for face recognition and clustering. In: Proceedings of the 2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR 2015). pp. 815–823. IEEE (jun 2015). https://doi.org/10.1109/CVPR.2015.7298682
26. Taigman, Y., Yang, M., Ranzato, M., Wolf, L.: DeepFace: Closing the gap to human-level performance in face verification. In: Proceedings of the 2014 IEEE Conference on Computer Vision and Pattern Recognition (CVPR 2014). pp. 1701–1708. IEEE (jun 2014). https://doi.org/10.1109/CVPR.2014.220
27. Tapaswi, M., Bäuml, M., Stiefelhagen, R.: "Knock! Knock! Who is it?" Probabilistic person identification in TV-series. In: CVPR 2012. pp. 2658–2665 (2012). https://doi.org/10.1109/CVPR.2012.6247986
28. Teh, Y.W.: Dirichlet process. In: Sammut, C., Webb, G.I. (eds.) Encyclopedia of Machine Learning, pp. 280–287. Springer US (2010). https://doi.org/10.1007/978-0-387-30164-8_219
29. Teh, Y.W., Jordan, M.I., Beal, M.J., Blei, D.M.: Hierarchical Dirichlet processes. Journal of the American Statistical Association **101**(476), 1566–1581 (dec 2006). https://doi.org/10.1198/016214506000000302
30. Torralba, A., Murphy, K.P., Freeman, W.T., Rubin, M.A.: Context-based vision system for place and object recognition. In: Proceedings Ninth IEEE International Conference on Computer Vision (ICCV 2003). vol. 1, pp. 273–280 (2003). https://doi.org/10.1109/ICCV.2003.1238354
31. Wang, X., Ma, X., Grimson, W.E.L.: Unsupervised activity perception in crowded and complicated scenes using hierarchical bayesian models. IEEE Transactions on Pattern Analysis and Machine Intelligence **31**(3), 539–555 (2009). https://doi.org/10.1109/TPAMI.2008.87

32. Zhang, L., Chen, L., Li, M., Zhang, H.: Automated annotation of human faces in family albums. In: MULTIMEDIA '03. pp. 355–358. ACM Press (2003). https://doi.org/10.1145/957013.957090
33. Zhao, M., Teo, Y.W., Liu, S., Chua, T.S., Jain, R.: Automatic person annotation of family photo album. In: CIVR 2006. pp. 163–172. Springer, Berlin, Heidelberg (2006). https://doi.org/10.1007/11788034_17
34. Zhu, X., Ghahramani, Z.: Learning from labeled and unlabeled data with label propagation. Tech. Rep. CMU-CALD-02-107, Carnegie Mellon University (2002)