Contextual Face Recognition with a Nested-Hierarchical Nonparametric Identity Model

Daniel C. Castro[*]
Imperial College London, UK
dc315@imperial.ac.uk

Sebastian Nowozin
Microsoft Research Cambridge, UK
Sebastian.Nowozin@microsoft.com

Abstract

Current face recognition systems typically operate via classification into known identities obtained from supervised identity annotations. There are two problems with this paradigm: (1) current systems are unable to benefit from often abundant unlabelled data; and (2) they 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? In previous work, we proposed an integrated Bayesian model that coherently reasons about the observed images, identities, partial knowledge about names, and the situational context of each observation. Here, we propose extensions of the contextual component of this model, enabling unsupervised discovery of an unbounded number of contexts for improved face recognition.

1 Introduction

Face identification can be decomposed 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.

An 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. 1).

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.

Recently, we introduced a unified Bayesian model which reflects all the above considerations on identity distributions, context-awareness and labelling (Fig. 1) (Castro and Nowozin, 2018). Our nonparametric identity model effectively represents an unbounded population of identities, while taking contextual co-occurrence relations and sparse noisy labels into account.

[*] Work partly done during an internship at Microsoft Research.

NeurIPS 2018 Workshop on All of Bayesian Nonparametrics (BNP@NeurIPS 2018), Montréal, Canada.

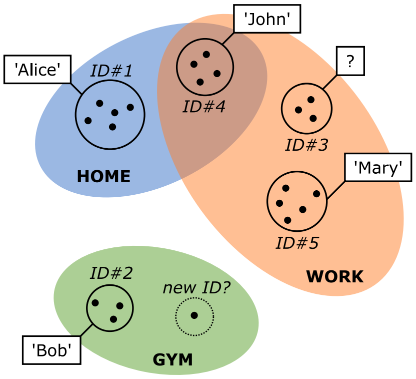

Figure 1: Context-aware model of identities (Castro and Nowozin, 2018)

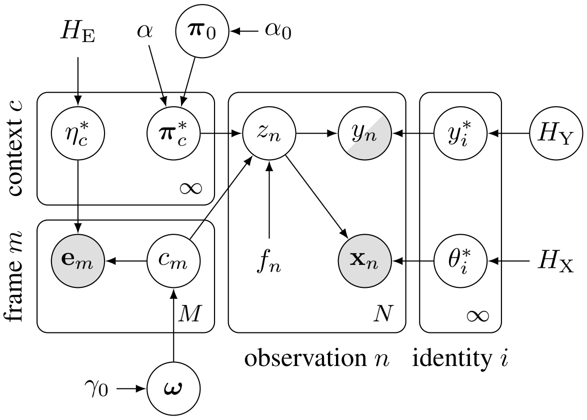

Figure 2: Overview of the proposed probabilistic model

In this preliminary work, we extend the referred context model in two ways: we explore its limit with an unbounded number of contexts, uncovering a rich nonparametric structure, and we lay the foundations for incorporating environmental cues (such as timestamps and geographical locations of frames) in our model to improve unsupervised context discovery and prediction.

2 Background

We begin by reviewing the face identification framework presented in Castro and Nowozin (2018), consisting of four main components: a context model (which we extend in Section 3), an identity model, a face model, and a semi-supervised label model.

2.1 Context Model

Data is assumed to be collected in frames, i.e. photo album or video stills, which are run through some off-the-shelf face detector. This produces N observations, grouped into the M frames via an indicator $f_n = m$ for each observation n in frame m. Context is therefore naturally shared among all face detections in each frame. We model context as a discrete latent variable, representing categories of situations in which a subject may find herself: e.g. home, work, gym.

We assume the context indicators $c_m \in \{1, \dots, C\}$ for each frame m, where C is some fixed number of distinct contexts, are independently distributed according to probabilities $\boldsymbol{\omega}$, which themselves follow a Dirichlet prior:

$$\boldsymbol{\omega} \sim \mathrm{Dir}(\boldsymbol{\gamma})\,, \qquad\qquad c_m \,|\, \boldsymbol{\omega} \sim \mathrm{Cat}(\boldsymbol{\omega})\,, \quad m = 1, \dots, M\,, \tag{1}$$

where M is the total number of frames. In our simulation experiments in Castro and Nowozin (2018), we used a symmetric Dirichlet prior, setting $\boldsymbol{\gamma} = (\gamma_0/C, \dots, \gamma_0/C)$.

2.2 Identity Model

In a daily-life scenario, an increasing number of unique identities will tend to appear as more faces are observed, i.e. we do not expect a user to run out of new people to meet. Moreover, some people are likely to be encountered much more often than others. Since a Dirichlet process (DP) (Ferguson, 1973) displays properties that mirror all of the above phenomena (Teh, 2010), 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) (Teh et al., 2006). 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. This hierarchical construction thus produces context-specific distributions over a common set of identities. Such a nonparametric model is additionally well suited for an open-set identification task, as it can elegantly estimate the prior probability of encountering an unknown identity.

To each of the N face detections is associated a latent identity indicator variable, z_n. Letting $\boldsymbol{\pi}_0$ denote the global identity distribution and $(\boldsymbol{\pi}_c^*)_{c=1}^C$ the context-specific identity distributions, we can write the generative process as

$$\begin{aligned}
\boldsymbol{\pi}_0 &\sim \mathrm{GEM}(\alpha_0)\,, && && (2)\\
\boldsymbol{\pi}_c^* \mid \boldsymbol{\pi}_0 &\sim \mathrm{HGEM}(\alpha_c, \boldsymbol{\pi}_0)\,, && c = 1, \ldots, C\,, && (3)\\
z_n \mid f_n = m, \mathbf{c}, (\boldsymbol{\pi}_c^*)_c &\sim \mathrm{Cat}(\boldsymbol{\pi}_{c_m}^*)\,, && n = 1, \ldots, N\,, && (4)
\end{aligned}$$

where $\mathrm{GEM}(\alpha_0)$ is the DP stick-breaking distribution, $\pi_{0i} = \beta_{0i} \prod_{j=1}^{i-1}(1 - \beta_{0j})$, with $\beta_{0i} \sim \mathrm{Beta}(1, \alpha_0)$ and $i = 1, \ldots, \infty$ (Sethuraman, 1994; Pitman, 2006). We additionally define a *hierarchical GEM distribution*, $\mathrm{HGEM}(\alpha, \boldsymbol{\pi}_0)$, such that $\pi_{ci}^* = \beta_{ci} \prod_{j=1}^{i-1}(1 - \beta_{cj})$, with $\beta_{ci} \sim \mathrm{Beta}(\alpha_c \pi_{0i}, \alpha_c(1 - \sum_{j=1}^{i} \pi_{0j}))$ (Teh et al., 2006, Eq. (21)).

2.3 Face Model

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:

$$\theta_i^* \sim H_{\mathrm{X}}\,, \quad i = 1, \ldots, \infty\,, \qquad\qquad \mathbf{x}_n \mid z_n, \boldsymbol{\theta}^* \sim F_{\mathrm{X}}(\theta_{z_n}^*)\,, \quad n = 1, \ldots, N\,. \tag{5}$$

As a consequence, the marginal distribution of faces is given by an *infinite mixture model* (Antoniak, 1974): $p(\mathbf{x}_n \mid c_n = c, \boldsymbol{\theta}^*, \boldsymbol{\pi}_c^*) = \sum_{i=1}^{\infty} \pi_{ci}^* F_{\mathrm{X}}(\mathbf{x}_n \mid \theta_i^*)$.

In 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. For the experiments reported in Castro and Nowozin (2018), we used embeddings produced by a pre-trained neural network (Amos et al., 2016). 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}).

2.4 Label Model

We expect to work with only a small number of user-labelled observations. Building on the *cluster assumption* for semi-supervised learning (Chapelle et al., 2006, Sec. 1.2.2), we attach a label variable (a *name*) to each cluster (identity), here denoted y_i^*. Since the number of distinct labels will tend to increase without bounds as more data is observed, we adopt a further nonparametric prior on these identity-wide labels, H_{Y},[2] using some base probability distribution L over the countable but unbounded label space (e.g. strings). In Castro and Nowozin (2018) we defined L over a rudimentary language model. Lastly, the observed labels, y_n, are assumed potentially corrupted through some noise process, F_{Y}. Let $\mathcal{L}$ denote the set of indices of the labelled data. We then have

$$\begin{aligned}
H_{\mathrm{Y}} &\sim \mathrm{DP}(\lambda, L)\,, && (6)\\
y_i^* \mid H_{\mathrm{Y}} &\sim H_{\mathrm{Y}}\,, \qquad\qquad i = 1, \ldots, \infty\,, && (7)\\
y_n \mid z_n, \mathbf{y}^*, H_{\mathrm{Y}} &\sim F_{\mathrm{Y}}(y_{z_n}^*; H_{\mathrm{Y}})\,, \quad n \in \mathcal{L}\,. && (8)
\end{aligned}$$

All concrete knowledge we have about the random label prior H_{Y} comes from the set of observed labels, $\mathbf{y}_{\mathcal{L}}$. Crucially, we can easily marginalise out H_{Y} (Teh, 2010), obtaining a tractable predictive label distribution, $\widehat{H_{\mathrm{Y}}}(y_{I+1}^* \mid \mathbf{y}^*)$.

According to the proposed noise model, an observed label, y_n, agrees with its identity's assigned label, $y_{z_n}^*$, with a fixed probability. Otherwise, it is assumed to come from a modified label distribution, in which we delete $y_{z_n}^*$ from $\widehat{H_{\mathrm{Y}}}$ and renormalise it. 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 an arbitrary random string.

[2] One could instead consider a Pitman–Yor process if power-law behaviour seems more appropriate than the DP's exponential tails (Pitman and Yor, 1997).

3 Extended Context Model

The context framework employed in Castro and Nowozin (2018) assumes a finite collection of pre-specified contexts and is fully supervised: an explicit context label is observed with each frame. This simplified scenario was adopted as a proof of concept, yet is admittedly unrealistic.

3.1 Unbounded Contexts

As reviewed in Section 2.1, the original context model had a *finite* $\mathrm{Dir}(\frac{\gamma_0}{C}, \dots, \frac{\gamma_0}{C})$ prior. A natural extension of such model is to take its limit as $C \to \infty$, while tying the values of all context-wise concentration hyperparameters ($\alpha_c = \alpha, \forall c$), which results in a Dirichlet process (Neal, 2000). In particular, up to a reordering of the contexts, the prior on context proportions, $\boldsymbol{\omega}$, becomes $\mathrm{GEM}(\gamma_0)$.

This transformation has interesting theoretical and practical implications: the resulting structure is a *nested-hierarchical Dirichlet process*.[3] As before, at the top level we have the global identity distribution, G_0, over face parameters and labels, and the context-specific identity distributions, $(G_c^*)_{c=1}^C$, follow a DP with G_0 as a base measure:

$$G_0 \mid H_\mathrm{Y} \sim \mathrm{DP}(\alpha_0, H_\mathrm{X} \otimes H_\mathrm{Y})\,, \tag{9}$$

$$G_c^* \mid G_0 \sim \mathrm{DP}(\alpha, G_0)\,, \qquad c = 1, \dots, \infty\,, \tag{10}$$

a prototypical example of a hierarchical DP (HDP) (Teh et al., 2006). If $G_0 = \sum_{i=1}^\infty \pi_{0i} \delta_{(\theta_i^*, y_i^*)}$, we can write $G_c^* = \sum_{i=1}^\infty \pi_{ci}^* \delta_{(\theta_i^*, y_i^*)}$.

Now, the nonparametric distribution of contexts implies wrapping the bottom level of the HDP, Eq. (10), as base for another DP, to form a nested DP (Blei et al., 2010; Rodríguez et al., 2008):

$$Q \mid G_0 \sim \mathrm{DP}(\gamma_0, \mathrm{DP}(\alpha, G_0))\,, \tag{11}$$

$$G_m \mid Q \sim Q = \textstyle\sum_{c=1}^\infty \omega_c \delta_{G_c^*}\,, \qquad m = 1, \dots, M\,. \tag{12}$$

This construction inherits desirable properties from both elements: the hierarchy ensures that all frame-wise identity distributions, $(G_m)_{m=1}^M$, have the same support, and nesting produces clusters of frames with shared identity weights (i.e. contexts).

3.2 Environmental Cues

While a purely identity-driven unsupervised context model may be able to disentangle co-occurrence patterns given enough data, we believe that environmental cues—such as timestamp and GPS coordinates of an acquired frame, if available—could considerably facilitate context discovery and prediction, in turn improving inference about identities.

Let us define $\mathbf{e}_m$ as the environmental measurements available for frame m, F_E a likelihood family parametrised by η_m, and H_E a prior distribution for such parameters. Plugging $\mathrm{DP}(\alpha, G_0) \otimes H_\mathrm{E}$ as base measure for the nested DP Q in Eq. (11), we can write

$$\eta_c^* \sim H_\mathrm{E}\,, \quad c = 1, \dots, \infty\,, \qquad \mathbf{e}_m \mid c_m, \boldsymbol{\eta}^* \sim F_\mathrm{E}(\eta_{c_m}^*)\,, \quad m = 1, \dots, M\,. \tag{13}$$

Some preliminary ideas for a spatial model include a 'geodetic' Fisher distribution or a tangential Gaussian (Straub et al., 2015), while a temporal model would have to accommodate recurring and occasional contexts, potentially adopting a Cox process formalism (Cox, 1955).

4 Conclusion

In this work, we reviewed the fully Bayesian treatment introduced in Castro and Nowozin (2018) 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, such that our principled identity model can contemplate context-specific probabilities of meeting an unbounded population.

We further proposed a nonparametric extension of the context model enabling unbounded context discovery, and discussed some of its theoretical implications in terms of nested-hierarchical nonparametric structures. Finally, we briefly examined how available environmental cues could be integrated into the model to replace the simplified supervised setting.

[3] This is related to the dual-HDP described in Wang et al. (2009) and the single-entity model of Agrawal et al. (2013), for example, although these works tended to focus on textual topic modelling.

Acknowledgments

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

References

Agrawal, P., Tekumalla, L. S., and Bhattacharya, I. (2013). Nested hierarchical Dirichlet process for nonparametric entity-topic analysis. In *Machine Learning and Knowledge Discovery in Databases – ECML PKDD 2013*, volume 8189 of *LNCS*, pages 564–579. Springer, Berlin, Heidelberg.

Amos, B., Ludwiczuk, B., and Satyanarayanan, M. (2016). OpenFace: A general-purpose face recognition library with mobile applications. Technical Report CMU-CS-16-118, CMU School of Computer Science.

Antoniak, C. E. (1974). Mixtures of Dirichlet processes with applications to Bayesian nonparametric problems. *The Annals of Statistics*, 2(6):1152–1174.

Blei, D. M., Griffiths, T. L., and Jordan, M. I. (2010). The nested Chinese restaurant process and Bayesian nonparametric inference of topic hierarchies. *Journal of the ACM*, 57(2):7.

Castro, D. C. and Nowozin, S. (2018). From face recognition to models of identity: A Bayesian approach to learning about unknown identities from unsupervised data. In *Computer Vision – ECCV 2018*, volume 11206 of *LNCS*, pages 745–761. Springer. Extended version with supplement: arXiv:1807.07872.

Chapelle, O., Schölkopf, B., and Zien, A., editors (2006). *Semi-Supervised Learning*. MIT Press.

Cox, D. R. (1955). Some statistical methods connected with series of events. *Journal of the Royal Statistical Society: Series B (Methodological)*, 17(2):129–164.

Ferguson, T. S. (1973). A Bayesian analysis of some nonparametric problems. *The Annals of Statistics*, 1(2):209–230.

Neal, R. M. (2000). Markov chain sampling methods for Dirichlet process mixture models. *Journal of Computational and Graphical Statistics*, 9(2):249–265.

Pitman, J. (2006). *Combinatorial Stochastic Processes*, volume 1875 of *Lecture Notes in Mathematics*. Springer-Verlag, Berlin. Lectures from the 32nd Summer School on Probability Theory held in Saint-Flour, July 7–24, 2002, with a foreword by Jean Picard.

Pitman, J. and Yor, M. (1997). The two-parameter Poisson–Dirichlet distribution derived from a stable subordinator. *The Annals of Probability*, 25(2):855–900.

Rodríguez, A., Dunson, D. B., and Gelfand, A. E. (2008). The nested Dirichlet process. *Journal of the American Statistical Association*, 103(483):1131–1154.

Sethuraman, J. (1994). A constructive definition of Dirichlet priors. *Statistica Sinica*, 4(2):639–650.

Straub, J., Chang, J., Freifeld, O., and Fisher III, J. W. (2015). A Dirichlet process mixture model for spherical data. In *Proceedings of the 18th International Conference on Artificial Intelligence and Statistics (AISTATS 2015)*, volume 38 of *PMLR*, pages 930–938. PMLR.

Teh, Y. W. (2010). Dirichlet process. In Sammut, C. and Webb, G. I., editors, *Encyclopedia of Machine Learning*, pages 280–287. Springer US.

Teh, Y. W., Jordan, M. I., Beal, M. J., and Blei, D. M. (2006). Hierarchical Dirichlet processes. *Journal of the American Statistical Association*, 101(476):1566–1581.

Wang, X., Ma, X., and Grimson, W. E. L. (2009). Unsupervised activity perception in crowded and complicated scenes using hierarchical Bayesian models. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 31(3):539–555.