Personalizing Gesture Recognition Using Hierarchical Bayesian Neural Networks

Ajjen Joshi[1] Soumya Ghosh[2] Margrit Betke[1] Stan Sclaroff[1] Hanspeter Pfister[3]
[1]Boston University [2]IBM T.J. Watson Research Center [3]Harvard University
[1]{ajjendj, betke, sclaroff}@bu.edu [2]ghoshso@us.ibm.com [3]pfister@seas.harvard.edu

Abstract

Building robust classifiers trained on data susceptible to group or subject-specific variations is a challenging pattern recognition problem. We develop hierarchical Bayesian neural networks to capture subject-specific variations and share statistical strength across subjects. Leveraging recent work on learning Bayesian neural networks, we build fast, scalable algorithms for inferring the posterior distribution over all network weights in the hierarchy. We also develop methods for adapting our model to new subjects when a small number of subject-specific personalization data is available. Finally, we investigate active learning algorithms for interactively labeling personalization data in resource-constrained scenarios. Focusing on the problem of gesture recognition where inter-subject variations are commonplace, we demonstrate the effectiveness of our proposed techniques. We test our framework on three widely used gesture recognition datasets, achieving personalization performance competitive with the state-of-the-art.

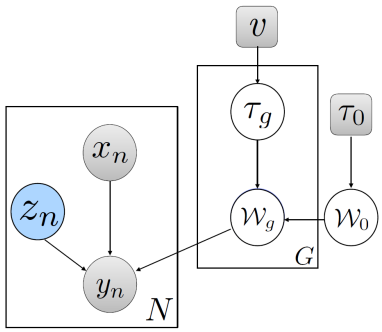

Figure 1. Graphical model representation of our hierarchical Bayesian model. Shaded nodes indicate observed random variables. We parameterize group-specific conditional distributions $p(y_n \mid z_n = g, f(x_n, \mathcal{W}_g))$, where $\mathcal{W}_g$ is the set of group-specific weights parameterizing a Bayesian neural network f. The class label, y_n, also depends on z_n, which indicates the group membership of data instance n. It is shaded blue to indicate that it is observed during training, but may be unobserved at test time.

1. Introduction

The problem of automatically recognizing human gestures has been an active area of computer vision and pattern recognition research. Gesture recognition enables natural and intuitive modes of interaction between human and computer, and therefore has numerous applications in a wide range of fields such as robotics, surveillance, and gaming. A generic gesture classifier, trained on examples of gestures pooled together from all subjects in the training set, is expected to be robust to variations with which gestures are performed by end-users. However, when the signal obtained from gestures performed by different users exhibit high variance, such systems have difficulty generalizing. Consider, for example, a vocabulary of gestures used by members of a household to control a smart-home device. Although each individual may perform the gestures consistently, it is likely that the gestures are performed with user-specific idiosyncrasies which may lead to large inter-subject variations in gesture performance. Designing systems robust to such variations is a challenging problem.

Personalizing gesture recognition systems using subject-specific training data provides a promising approach to alleviating such difficulties. In this paper, focusing on personalization, we build hierarchical Bayesian classifiers (Figure 1) that adapt to new subjects using subject-specific conditional distributions (Figure 2). Different from existing hierarchical Bayesian models, we parameterize the conditional distributions via multi-layered Bayesian neural networks. They allow us to learn potentially complex functional relationships between a subject's gestures and class labels from a modest number of training examples. Furthermore, by explicitly modeling uncertainty in weights, Bayesian neural networks are able to provide well calibrated estimates of posterior uncertainty along with predicted class labels. Leveraging recent progress on scalable stochastic variational inference, we develop algorithms for learning the posterior distribution over all network weights in the hierarchy. We further

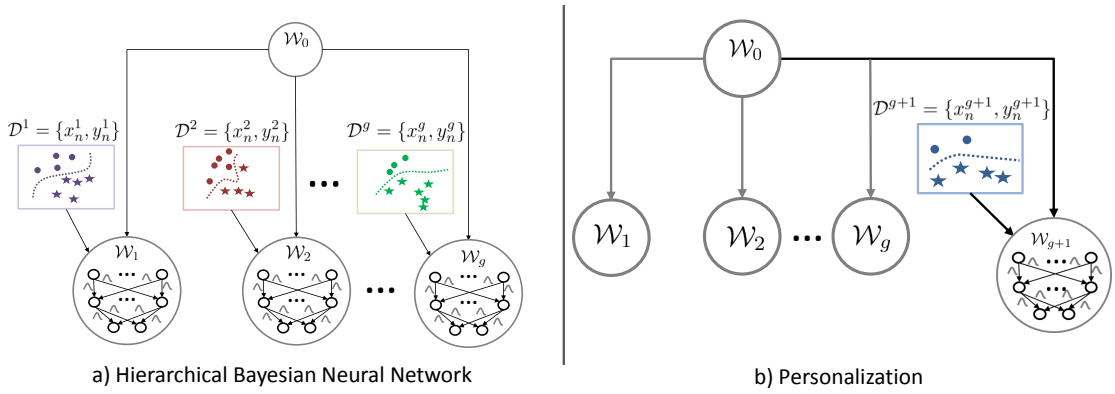

Figure 2. (a) Given gesture examples produced by g subjects, we train a classifier using a hierarchical framework, where $\mathcal{W}_g$ is the set of group-specific weights parameterizing a Bayesian neural network. The different shapes correspond to different gesture classes and the different colors represent the subjects who produced those examples. (b) Given few instances of training data from a new subject, we personalize our model to learn weights specific to the new subject.

use the inferred posterior to drive active learning algorithms that guide interactive labeling of personalization gestures given a small pool of unlabeled subject-specific gestures. We systematically test various aspects of the proposed models and algorithms on three challenging gesture recognition datasets — the MSRC-12 Kinect Gesture Dataset [10], the 2013 ChaLearn Gesture Challenge Dataset [8] and the NATOPS gesture dataset [29]. We find that even with relatively shallow two hidden layer networks, our approach is competitive with the *state-of-the-art* gesture personalization systems. We also empirically demonstrate that even with naive fully factorized variational inference, Bayesian neural networks provide uncertainty estimates that are useful for guiding active learning procedures.

In summary, we make three contributions in this paper. First, we develop hierarchical Bayesian neural networks for personalized gesture recognition in the presence of inter-subject variations. Second, we adapt reduced variance versions of stochastic variational inference for learning the posterior distribution over model parameters. Third, we utilize the inferred posterior to drive an active learning procedure that consistently improves over naive personalization. Our results demonstrate the effectiveness of the proposed models and algorithms for gesture recognition.

2. Related Work

Gesture recognition systems using various machine learning methods including nearest-neighbors based on dynamic time warped (DTW) distances [1], hidden Markov models (HMM) [30], hidden conditional random fields (HCRF) [28], random forests [16] and deep neural networks [23], have been proposed. Although related, our main focus is on the task of personalized classification of gestures.

Personalization approaches have been developed for speech [27], handwriting [7, 17], facial action unit recognition [6] and gestures [15]. Work on domain adaptation that either adapts model parameters [33] or feature representations [26] is closely related to these approaches. Our work draws on previous efforts in hierarchical Bayesian domain adaptation [9]. We extend this line of work by parameterizing group/domain-specific conditional distributions via more flexible Bayesian neural networks in place of simpler log-linear models.

A particular challenge faced by personalization systems is the small amounts of subject-specific data available for personalization. Yao et al. [34] tackled this by recasting the problem into one of selecting the best performing model from a portfolio of pre-trained models. Since no new learning occurs, the approach is very data efficient. However, they find it to be outperformed by baselines where the models are partially or fully re-trained given new personalization instances. We deal with data paucity by resorting to Bayesian neural networks. Pioneering work on Bayesian neural networks can be traced back to [5, 21, 22]. Recent progress in deep learning along with advances in scalable inference has reinvigorated interest in them. Hierarchical Bayesian neural networks have previously been proposed [13, 20]. However, they rely on expensive Markov chain Monte-Carlo inference and fail to scale to even moderate sized architectures. In contrast, we exploit stochastic variational methods [3, 32] that scale to both large architectures and large datasets. Previous work has developed such algorithms for Bayesian neural network [3] and Bayesian logistic regression [32] models. We introduce a stochastic variational formulation for hierarchical Bayesian neural networks. Further, we exploit the inferred posterior over weights to guide active learning [14] methods that significantly improve performance of the system in scenarios where labeling data is expensive.

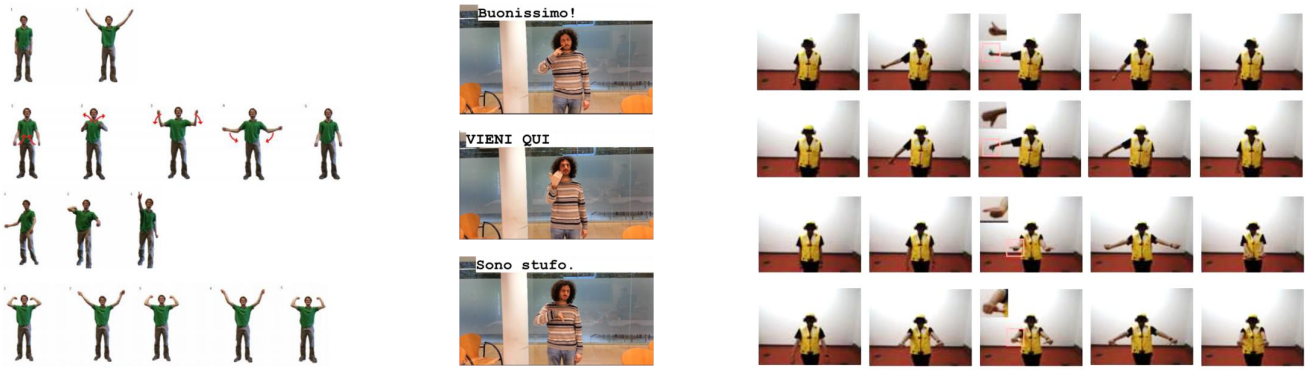

Figure 3. Examples of gestures from MSRC-12 dataset (left), ChaLearn 2013 dataset (middle) and the NATOPS dataset (right)

3. Hierarchical Bayesian Neural Networks

Given a dataset $\mathcal{D} = \{x_n, y_n\}_{n=1}^N$, containing N gesture $x_n \in \mathbb{R}^D$, and label $y_n \in \mathcal{Y}$ pairs, we aim to learn the functional mapping from gestures to class labels and to make class predictions for previously unseen gestures x_*. Further, we focus on the case where $\mathcal{D}$ is generated by G distinct subjects.

To preserve subject-specific effects we endow each subject with its own conditional distribution, allowing the gesture-label mapping to vary among subjects. The conditional distributions are parameterized via multi-layered feedforward neural networks, which enables the model to capture potentially complex mappings between gestures and labels. Assuming the distribution factorizes over data instances, we have,

$$p(\mathbf{y} \mid \mathcal{W}, \mathbf{z}, \mathbf{x}) = \prod_{n=1}^{N} \prod_{g=1}^{G} p(y_n \mid f(\mathcal{W}_g, x_n))^{\mathbf{1}[z_n=g]}. \quad (1)$$

Here, z_n is a G-dimensional categorical random variable indicating the subject membership of data instance n. We assume that the subject indicators $\mathbf{z} = \{z_n\}_{n=1}^N$ are observed during training. During testing we are able to reason about the class label y_* of a held-out feature x_* even when the corresponding subject membership z_* is unobserved. We wish to learn $\mathcal{W} = \{\mathcal{W}_1, \ldots, \mathcal{W}_G\}$, where $\mathcal{W}_g$ is the set of subject-specific weights parameterizing a neural network f whose hidden layers employ rectified linear activations and whose output layer is constrained to be linear. We note here that the function f can be any differentiable function.

We place factorized Gaussian priors on $\mathcal{W}_g$ with independent subject-specific variances to model our prior assumption that each subject's functional mapping is an independently corrupted version of a common latent mapping (parameterized by $\mathcal{W}_0$),

$$p(\mathcal{W}_g \mid \mathcal{W}_0, \tau_g) = \prod_{l=1}^{L} \prod_{i=1}^{V_{l-1}} \prod_{j=1}^{V_l} \mathcal{N}(w_{ij,l}^g \mid w_{ij,l}^0, \tau_g^{-1}). \quad (2)$$

We further place uninformative priors — zero mean Gaussians with a large fixed variance τ_0^{-1} on the weight means $\mathcal{W}_0$,

$$p(\mathcal{W}_0 \mid \tau_0) = \prod_{l=1}^{L} \prod_{i=1}^{V_{l-1}} \prod_{j=1}^{V_l} \mathcal{N}(w_{ij,l}^0 \mid 0, \tau_0^{-1}). \quad (3)$$

Here, V_l denotes the number of units in layer l and $l = 0$ corresponds to the input layer.

The subject specific variances τ_g^{-1} control the amount of deviation from the mean exhibited by the subject's gesture-label mapping. Specifying them manually can be difficult and authors in the past [34] have resorted to setting them via cross-validation. Although cross-validation procedures can be effective for simpler models, they are untenable here. Such a procedure would involve searching over G-dimensional continuous spaces, re-training the model for each parameter candidate. Instead, we place hyper-priors on the variances and infer them jointly with $\mathcal{W}$. The Gamma distribution is the conjugate prior over the precision of a Gaussian distribution and hence a popular choice [2]. However, recent work [11] has shown it to be unsuitable for specifying uninformative priors in hierarchical models. Following [11], we instead use the half-normal distribution with a large fixed variance v to specify uninformative priors over subject-specific standard deviations $\tau_g^{-1/2}$,

$$p(\gamma_g \mid v) = \mathcal{N}(\gamma_g \mid 0, v); \quad \tau_g^{-1/2} = |\gamma_g|, \quad (4)$$

where we have introduced an auxiliary variable γ_g and used the property, if $a \sim \mathcal{N}(0, \sigma^2)$, then $|a| \sim$ Half-Normal$(0, \sigma^2)$. It also immediately follows that $\tau_g^{-1} = \gamma_g^2$. In the next section, we will see that the auxiliary variable formulation simplifies inference. Finally, we model the observed class labels as categorically distributed random variables,

$$y_n \mid \mathcal{W}, x_n, z_n \sim \text{Cat}(y_n \mid \mathcal{S}(f(\mathcal{W}_{z_n}, x_n))), \quad (5)$$

where $\mathcal{S}(a) = \exp\{a\}/\sum_k \exp\{a_k\}$ is the softmax function that maps the real valued output of f to the probability simplex. We can summarize the joint distribution specified by the model as,

$$\begin{aligned} p(\mathcal{W}_0, \mathcal{W}, \mathcal{T}, \mathbf{y} \mid \mathbf{x}, \mathbf{z}, \tau_0, v) = p(\mathcal{W}_0 \mid \tau_0^{-1}) \\ \prod_{g=1}^{G} p(\gamma_g \mid v) p(\mathcal{W}_g \mid \mathcal{W}_0, \tau_g^{-1}) \\ \prod_{n=1}^{N} \prod_{g=1}^{G} p(y_n \mid f(\mathcal{W}_g, x_n))^{\mathbf{1}[z_n=g]}, \end{aligned} \tag{6}$$

where $\mathcal{T} = \{\gamma_1, \ldots, \gamma_G\}$. The hierarchical Bayesian neural network explicitly captures inter-subject variances by allowing the subject-specific conditional distribution of data from different subjects to systematically vary from each other. At the same time, they share statistical strength across subjects — samples observed for a particular subject not only provide information about that subject's distribution but also about other subject-specific distributions.

4. Scalable Learning and Inference

Learning our model involves inferring the posterior distribution $p(\mathcal{W}_0, \mathcal{W}, \mathcal{T} \mid \mathcal{D}, \mathbf{z}, \gamma_0, v)$ over model parameters. Unfortunately, the nonlinear activations employed by the networks in the hierarchy render this posterior intractable forcing us to resort to approximate inference techniques. Leveraging recent advances in scalable approximate Bayesian learning, we use variational inference to learn a tractable approximation to the posterior. We restrict the approximating family to the following form,

$$q(\mathcal{W}_0, \mathcal{W}, \mathcal{T} \mid \phi) = q(\mathcal{W}_0 | \phi_0) \prod_{g=1}^{G} q(\mathcal{W}_g | \phi_g) q(\gamma_g | \phi_{\gamma_g}), \tag{7}$$

where $\phi = \{\phi_0, \phi_1, \ldots, \phi_G, \phi_{\gamma_1}, \ldots, \phi_{\gamma_G}\}$ represents the variational free parameters. We approximate the weight posteriors with fully factorized Gaussian distributions,

$$\begin{aligned} q(\mathcal{W}_0 | \phi_0) = \prod_{l=1}^{L} \prod_{i=1}^{V_{l-1}} \prod_{j=1}^{V_l} \mathcal{N}(w_{ij,l}^0 \mid \mu_{ij,l}^0, \psi_{ij,l}^0), \\ q(\mathcal{W}_g | \phi_g) = \prod_{l=1}^{L} \prod_{i=1}^{V_{l-1}} \prod_{j=1}^{V_l} \mathcal{N}(w_{ij,l}^g \mid \mu_{ij,l}^g, \psi_{ij,l}^g). \end{aligned} \tag{8}$$

The auxiliary variable γ_g affects the model only through its absolute value $|\gamma_g|$. Thus, we can also restrict the posterior of γ_g to $q(\gamma_g | \phi_{\gamma_g}) = \mathcal{N}(\gamma_g \mid \mu_{\gamma_g}, \psi_{\gamma_g})$, a Gaussian family.

We optimize the variational parameters to minimize the Kullback-Leibler divergence $\mathrm{KL}(q||p)$ between the true posterior and the variational approximation by maximizing the expected lower bound (ELBO),

$$\begin{aligned} \mathcal{L}(\phi) = \mathbb{E}_{q_\phi}[\ln p(\mathcal{W}_0, \mathcal{W}, \mathcal{T}, \mathbf{y} \mid \mathbf{x}, \mathbf{z}, \gamma_0, v)] \\ - \mathbb{E}_{q_\phi}[\ln q(\mathcal{W}_0, \mathcal{W}, \mathcal{T} \mid \phi)], \end{aligned} \tag{9}$$

with respect to the variational free parameters ϕ.

The non-conjugacy between the neural network parameterized categorical distributions and the Gaussian priors cause the expectations in the ELBO to be intractable. This precludes the availability of traditional fixed point updates. Instead, following recent work [32, 3, 19, 24], we approximate the intractable expectations with unbiased Monte-Carlo estimates,

$$\begin{aligned} \hat{\mathcal{L}}(\phi) = \frac{1}{S} \sum_{s=1}^{S} \ln p(\mathcal{W}_0^s, \mathcal{W}^s, \mathcal{T}^s, \mathbf{y} \mid \mathbf{x}, \mathbf{z}, \gamma_0, v) \\ - \mathbb{E}_{q_\phi}[\ln q(\mathcal{W}_0, \mathcal{W}, \mathcal{T} \mid \phi)], \\ \mathcal{W}_0^s, \mathcal{W}^s, \mathcal{T}^s \sim q(\mathcal{W}_0, \mathcal{W}, \mathcal{T} \mid \phi). \end{aligned} \tag{10}$$

The gradient $\nabla_\phi \mathcal{L}(\phi)$ is then approximated with the noisy but unbiased estimate $\nabla_\phi \hat{\mathcal{L}}(\phi)$. Computing $\nabla_\phi \hat{\mathcal{L}}(\phi)$ requires gradients with respect to the means and variances of the Gaussian variational approximations. The non-centered parameterization proposed in [19], $w \sim \mathcal{N}(\mu, \psi) \Leftrightarrow \epsilon \sim \mathcal{N}(0,1), w = \mu + \psi^{1/2}\epsilon$, allows us to differentiate through the Monte-Carlo approximation,

$$\begin{aligned} \nabla_{\mu,\sigma} \mathbb{E}_{q_w}[g(w)] \Leftrightarrow & \nabla_{\mu,\psi} \mathbb{E}_{\mathcal{N}(\epsilon|0,1)}[g(\mu + \psi^{1/2}\epsilon)] \\ = & \mathbb{E}_{\mathcal{N}(\epsilon|0,1)}[\nabla_{\mu,\psi} g(\mu + \psi^{1/2}\epsilon)] \\ = & \frac{1}{S} \sum_s \nabla_{\mu,\psi} g(\mu + \psi^{1/2}\epsilon^s); \epsilon^s \sim \mathcal{N}(0,1), \end{aligned} \tag{11}$$

for any differentiable function g. With the unbiased gradient estimates in hand, Equation 9 can be optimized through stochastic gradient ascent [4].

4.1. Local Reparameterization

Although stochastic gradient ascent is guaranteed to asymptotically converge to a local optimum, its non asymptotic performance is contingent on the variance of the unbiased gradient estimates. While the gradient estimate in Equation 11 has been previously used to learn Bayesian neural networks [3], we find the variance of this estimator too high to effectively learn our hierarchical model.

To address this issue, we note that the weights in a layer only influence the ELBO ($\mathcal{L}(\phi)$) through the layer's pre-activations. Instead of estimating the ELBO by sampling the variational posterior on the weights one could instead sample the implied variational distribution on the considerably smaller number of pre-activations. This is the "local reparameterization trick" introduced in [18], where the authors show that the corresponding gradient estimates have provably lower variance. For factorized Gaussian variational posteriors over weights, the corresponding pre-activation distributions are also easy-to-compute factorized

Gaussians. The pre-activation b_{il}, of the i^{th} node of layer l is distributed as $\mathcal{N}(\mu_{w_{il}}^T a, \sigma_{w_{il}}^{2^T} a^2)$, where a is the input to layer l, $\mu_{w_{il}}$ and $\sigma_{w_{il}}^2$ are the means and variances of the variational posterior over weights associated with node i.

We find that local reparameterization provides significant computational cost savings, accuracy improvements and is crucial for effectively learning hierarchical Bayesian neural networks.

4.2. Predictions

Given a held-out gesture x_* from an observed subject z_*, the posterior predictive distribution over classes is given by,

$$\begin{aligned} & p(y_* \mid x_*, \mathcal{D}) \\ &= \int p(y_* | \mathcal{W}, z_*, x_*) p(\mathcal{W}_0, \mathcal{W}, \mathcal{T} \mid \mathcal{D}) d\mathcal{W}_0 d\mathcal{W} d\mathcal{T} \\ &\approx \int p(y_* \mid \mathcal{W}, z_*, x_*) q(\mathcal{W}_{z_*} \mid \hat{\phi}_{z_*}) d\mathcal{W}_{z_*}, \end{aligned} \tag{12}$$

where the approximation in the second line follows from the variational approximation and $\hat{\phi}_{z_*}$ denotes the optimal variational parameters. In our experiments, we evaluate the integral using a Monte-Carlo estimate.

Next, we consider the case when subject (z_*) and class (y_*) memberships are both unobserved and need to be inferred. Classifying x_* involves performing an additional inference of its subject membership. Since this inference needs to be performed at test time for each data instance, it is imperative that the inference be fast. To facilitate fast and accurate inference of the subject memberships, we use an inference network [25, 12] h_θ, another multi-layered fully connected neural network with weights θ and a G dimensional softmax output layer. We learn this inference network by utilizing all examples from the training set where z is observed. This inference network paramterizes the approximate posterior $q(z \mid x)$. Because z is observed during training, training of the subject inference network can occur independently of other variational parameters. At test time, inferring a distribution over the unknown subject memberships, $q(z_* \mid x_*, \hat{\theta}) = \text{Cat}(z_* \mid h_{\hat{\theta}}(x_*))$, simply involves a single forward pass through the network, where $\hat{\theta}$ denotes the estimated weights. Our use of an inference network is in sharp contrast to traditional mean field methods where each datapoint is assigned an independent variational parameter that is optimized via several iterations of expensive optimization, at test time. In the presence of a new subject, we add an output node to the subject inference network. However, we find that only updating the weights associated with the new node is sufficient and the network need not be retrained.

Marginalizing over the joint posterior predictive distribution, we get the predictive distribution over class labels:

$$\begin{aligned} & p(y_* \mid x_*, \mathcal{D}) = \sum_{z_*=1}^{G} p(y_*, z_* \mid x_*, \mathcal{D}) \\ &= \sum_{z_*=1}^{G} \int p(y_* | \mathcal{W}, z_*, x_*) p(\mathcal{W}_0, \mathcal{W}, z_*, \mathcal{T} \mid \mathcal{D}) d\mathcal{W}_0 d\mathcal{W} d\mathcal{T} \\ &\approx \sum_{z_*=1}^{G} q(z_* \mid x_*, \hat{\theta}) \int p(y_* \mid \mathcal{W}, z_*, x_*) q(\mathcal{W}_{z_*} \mid \hat{\phi}_{z_*}) d\mathcal{W}. \end{aligned} \tag{13}$$

The integral over $\mathcal{W}$ is estimated via a Monte-Carlo approximation, $p(y_* \mid x_*) \approx \sum_{z_*=1}^{G} q(z_* \mid x_*, \hat{\theta}) \frac{1}{T} \sum_t p(y_* \mid \mathcal{W}^t, z_*, x_*), \mathcal{W}^t \sim q(\mathcal{W} \mid \hat{\phi}_{z_*}, \hat{\theta})$.

5. Personalization

In this section, we focus on incorporating data from a new, previously unseen subject and adapting the model to the new subject. We call this process personalization and focus on the cases when a small number of data instances from the new subject are made available for training. Denoting gestures from new subject $G+1$ as $\mathcal{D}_{G+1}$, we learn a subject-specific model $\mathcal{W}_{G+1} \mid \mathcal{D}_{G+1}$. The learning can be performed efficiently by observing that $\{\mathcal{W}_g\}_{g=1}^{G+1}$ are conditionally independent given $\mathcal{W}_0$. Thus, given a model trained on $\mathcal{D}$, we only update $\mathcal{W}_{G+1}$ while keeping the estimates $\{\mathcal{W}_g\}_{g=1}^{G} \mid \mathcal{D}$ and $\mathcal{W}_0 \mid \mathcal{D}$ fixed. We could additionally update the posteriors $\{\mathcal{W}_g\}_{g=1}^{G} \mid \mathcal{D} \cup \mathcal{D}_{G+1}$ and $\mathcal{W}_0 \mid \mathcal{D} \cup \mathcal{D}_{G+1}$. However, typically only a small number of adaptation instances $\mathcal{D}_{G+1}$ are available — too few to have a sizeable effect on the posteriors $\{\mathcal{W}_g\}_{g=1}^{G} \mid \mathcal{D}$ and $\mathcal{W}_0 \mid \mathcal{D}$.

5.1. Active Learning

Collecting and labeling personalization gestures can be expensive. For example, consider a system designed to recognize specialized gestures such as those made by naval aircraft handlers onboard aircraft carriers. Not only is the process of collecting additional gestures likely to be challenging, labeling the gestures requires specialized domain knowledge and can be prohibitively expensive. To best utilize limited labeling resources, we next describe an active learning procedure to guide the selection of gestures to label, given a small pool of unlabeled adaptation examples.

Having access to the posterior distribution over weights, rather than just point estimates, allows us to use Bayesian active learning by disagreement (BALD) — a *state-of-the-art* active learning algorithm [14]. Given a pool of unlabeled gestures X_{pool} from subject g and a model trained on $\mathcal{D}$, BALD sequentially selects gestures x_l, such that,

$$x_l = \underset{x \in \text{X}_{pool}}{\text{argmax}} \, \mathbb{H}[y \mid x, \mathcal{D}] - \mathbb{E}_{\mathcal{W}_g \sim p(\mathcal{W}_g \mid \mathcal{D})} \mathbb{H}[y \mid x, \mathcal{W}_g], \tag{14}$$

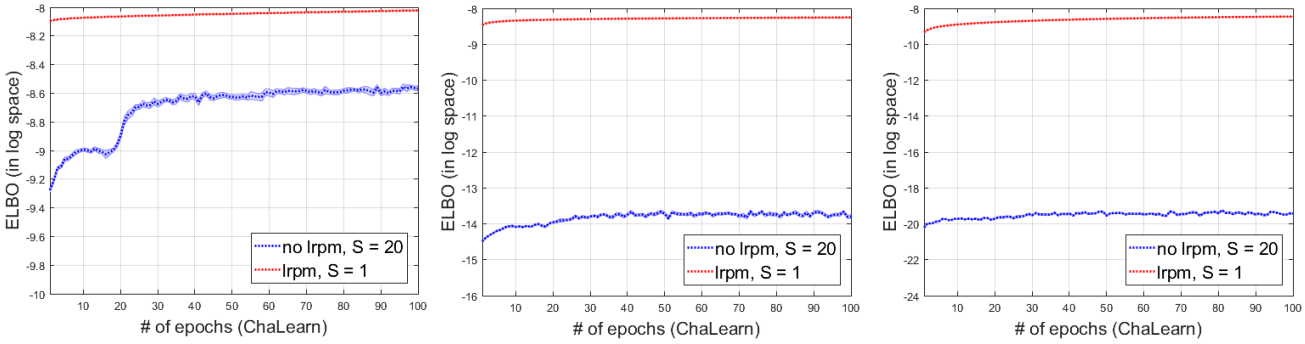

Figure 4. The mean logarithm of the expected lower bound (ELBO) versus the number of training epochs, for 15 random 75-25 splits of the ChaLearn dataset, when the model uses local reparameterization (lprm) and when it doesn't (no lrpm) for different HBNN architectures: HBNN with one hidden layer (left), HBNN with two hidden layers (middle) and HBNN with three hidden layers (right).

where $\mathbb{H}[t] = -\int p(t)\log p(t)dt$. As noted by Houlsby et al. [14], Eq. 14 lends itself to an intuitive explanation: BALD seeks a data instance x_l for which the model, averaging over all weights, is uncertain about y (high $\mathbb{H}[y \mid x, \mathcal{D}]$) but individual settings of the weights have high certainty in their predictions (low $\mathbb{E}_{\mathcal{W}_g \sim p(\mathcal{W}_g|\mathcal{D})}\mathbb{H}[y \mid x, \mathcal{W}_g]$) — i.e., when the posterior weights disagree the most. Approximation methods to efficiently evaluate Eq. 14 are available for certain classes of models, but do not extend to our multi-class classification problem. We therefore resort to a Monte-Carlo approach. We empirically found that, even with a modest number of samples, the approximations significantly improve upon selecting gestures uniformly at random.

6. Experimental Results

We used three datasets to test our framework, all of which contain skeletal data of the subjects performing the gestures. The MSRC-12 Kinect Gesture Dataset contains 12 different gestures performed by 30 different subjects for a total of ∼4900 gesture instances (Figure 3 left). The gestures were recorded using the Microsoft Kinect.

The 2013 Chalearn Gesture Challenge dataset contains examples of 20 gestures collected from 36 different subjects. Like Yao et al. [34], we experimented with the Training and Validation data containing ∼11000 samples. The gestures in the dataset, recorded using the Microsoft Kinect, represent common communication signals used in the Italian language (Figure 3 middle).

The NATOPS dataset [29] consists of 24 unique aircraft handling signals performed by 20 different subjects, where each gesture has been performed 20 times by all subjects (Figure 3 right). A 12-dimensional vector of body features (angular joint velocities for the right and left elbows and wrists), as well as an 8 dimensional vector of hand features (probability values for hand shapes for the left and right hands) collected by Song et al. [29] are provided as features for all frames of all videos in the dataset.

For controlled comparisons with previous work [34], we used identical feature representations — raw x,y,z world coordinates for 20 body joints in the MSRC-12 and Chalearn datasets. For NATOPS, we used the 20 dimensional features made available in [29], per frame. We extracted frames by sampling uniformly in time and concatenated the per-frame features to produce 600-dimensional input feature vectors for all three datasets. This allowed us to use a common model architecture for the three different datasets. In our experiments, we trained a Hierarchical Bayesian Neural Network with varying number of hidden layers, each with 400 activation nodes. We set the hyper-parameters v to 100 and τ_0^{-1} to 1000 and used RMSprop [31] to optimize the ELBO.

6.1. Benefits of Local Reparameterization

To investigate the effectiveness of the locally re-parameterized ELBO gradients, we trained an HBNN with 1, 2 and 3 hidden layers, each layer with 400 activation nodes, for 100 epochs replicated over 15 random 75/25 splits of the ChaLearn dataset. Fig. 4 displays the ELBO evolution over the course of training with and without local reparameterization (lprm). We found that for all three architectures, the models using locally re-parameterized gradients made better progress, achieving higher expected lower bounds with the gap in performance increasing with depth.

6.2. Gesture Recognition

Next, we demonstrate the flexibility afforded by parameterizing the group-specific conditional distributions with Bayesian neural networks. For all datasets, we trained a HBNN with two hidden layers with 400 units each and benchmark against two strong baselines: a multinomial regression version of our hierarchical Bayesian framework (HBMR), and a two hidden layer non-hierarchical Bayesian neural network that pools data from all subjects into a sin-

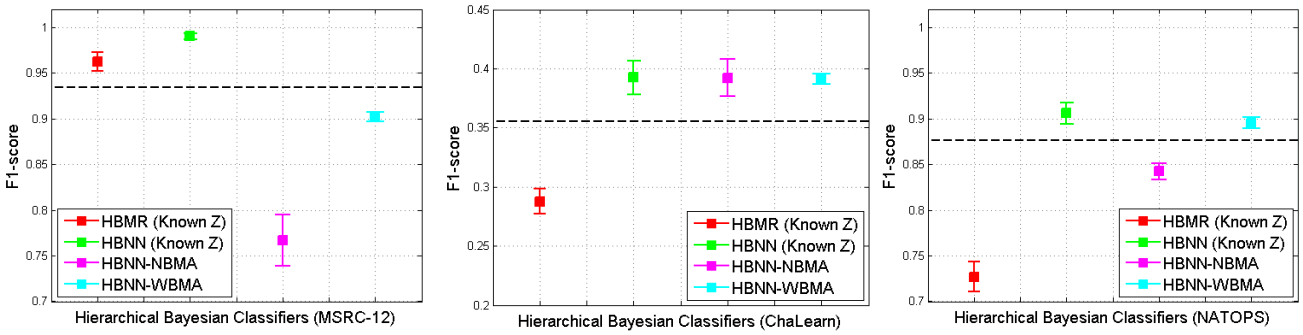

Figure 5. The mean F1-scores for different versions of our Hierarchical Bayesian gesture classifier. For all three datasets (MSRC-12 dataset (left) and Chalearn 2013 dataset (middle) and NATOPS dataset (right)), we trained a Hierarchical Bayesian Multinomial Regression classifier (HBMR) and a Hierarchical Bayesian Neural Network (HBNN) and used them to predict the class labels of the test data. For HBNN, when group membership of the test data is known, we used the weights belonging to the corresponding group to make a prediction (HBNN (Known Z)). When group membership of the test data is unknown, we present results obtained with Naive Bayesian Model Averaging (HBNN-NBMA) and Weighted Bayesian Model Averaging (HBNN-WBMA). We compared our results with a baseline BNN trained with data from all subjects pooled into one group, whose mean is depicted in the figures as a dashed black line.

gle group. We trained all models for 50 epochs on 5 random 75/25 replications of the data. Fig. 5 presents the corresponding results. First, focusing on the case when subject memberships are known (HBNN-Known Z and HBMR-Known Z), we found that the non-linear HBNN models significantly improved upon their (conditionally) linear counterparts HBMR models across the three datasets. HBNNs also outperformed the non-hierarchical Bayesian neural networks on all three datasets clearly demonstrating the benefits of employing subject-specific models over pooled ones. Interestingly, HBMR only outperformed the non-hierarchical Bayesian neural network on the MSRC dataset. This suggests that compared to capturing complex non-linear relationships between gestures and labels, modeling subject-specific idiosyncrasies is less important for the NATOPS and Chalearn datasets. Further comparisons with existing gesture recognition systems are available in the supplement.

Unknown Subject Memberships. We studied the effectiveness of our proposed subject membership inference network. When the membership of a test gesture is unknown we compared two methods for predicting its class label — naive Bayesian model averaging (HBNN-NBMA) where we uniformly averaged the posterior predictive distributions of all subjects and, weighted Bayesian model averaging (HBNN-WBMA), where the weights were determined by the subject membership inference network. On the MSRC-12 and NATOPS datasets, we found that HBNN-WBMA significantly outperformed HBNN-NBMA. On ChaLearn, both methods performed similarly but HBNN-WBMA exhibited lower variance across splits. Together, these results demonstrate that the use of a recognition network is helpful when subject-memberships are not known at test time.

We note that *apriori* knowledge of the subject-membership of a gesture leads to better predictive performance on all but the ChaLearn dataset. The ChaLearn dataset is more challenging due to less rigidly defined gestures. This results in more variability in gestures and weakens our assumption that each subject performs a given gesture consistently and differently from other individuals. This may explain why knowing the subject memberships does not translate into significant performance improvements.

6.3. Personalization

Finally, we present experiments demonstrating the personalization ability of HBNN models. Given a limited number of training instances from the new subject, we learned model parameters tuned to the subject. For all datasets, we used a leave-one-subject-out cross validation scheme, where we personalized models pre-trained on $G-1$ subjects and used a pool of seven (fifteen for NATOPS) randomly selected gestures per class from the test subject for personalization. Both pre-trained and personalized models contained two layers, with 400 units each, and were trained for 50 epochs. We considered two schemes for incorporating gestures from the personalization pool: RAND, where data from the training pool of the test subject was added uniformly at random, and BALD where data from the training pool was selected using uncertainty-based sampling (Eq. 14). For each test subject, we repeated the experiment five times, randomly selecting the pool of personalization gestures in each replicate.

We benchmarked these methods against a strong non-personalized baseline — a non-hierarchical BNN (with two 400-unit hidden layers) trained with data from all subjects

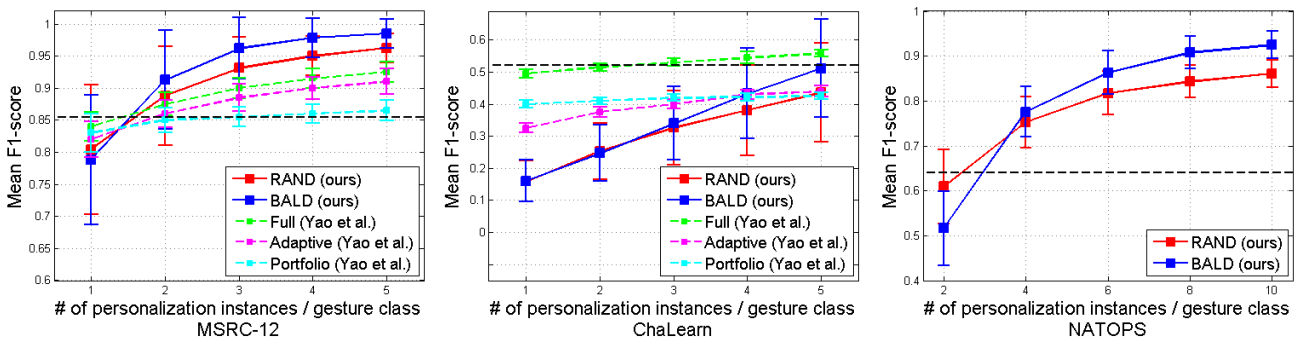

Figure 6. The mean F1-scores for different personalization schemes plotted against the number of personalization instances per gesture. We observe that personalization using BALD outperforms personalization using RAND when the number of personalization instances is greater than 1 for the MSRC-12 dataset (left), 3 for the ChaLearn 2013 dataset (middle) and 4 for the NATOPS dataset (right). Our results also compare favorably with the personalization methods presented by Yao et al. [34], who reported their results for the MSRC-12 and ChaLearn 2013 datasets. We compare the personalization results with a baseline BNN trained with all training data pooled into one group, whose mean is depicted in the figures as a dashed black line.

except the test (personalization) subject pooled together. The results in Fig. 6 show that with as few as two and three gesture examples per subject, HBNN outperformed the baseline on MSRC and NATOPS. On ChaLearn, BALD with five gesture examples per class performed as well as the non-personalized baseline.

It may be surprising to note that personalization baseline on ChaLearn (Fig. 6) resulted in higher F1 scores than the non-personalized baseline presented in Fig. 5. However, the baseline in Fig 5 corresponds to a model trained on samples from all subjects but with the training set size limited to 75% while the model in Fig 6 was trained on 35 out of 36 subjects corresponding to 97% of the dataset. For the ChaLearn data intra-subject variability in gestures dwarfs inter-subject variations. Thus, observing more of the dataset as opposed to gestures from the same subject leads to better performance. This is also why HBNNs need more (4) personalization examples for ChaLearn than the other datasets.

Comparing BALD with RAND, we found that BALD improves personalization performance on all three datasets, when the number of training instances exceeded one, three and four for MSRC, NATOPS and ChaLearn datasets. This is an interesting result which suggests that even our naive mean field approximation provides predictive uncertainty estimates of sufficient fidelity that lead to BALD's uncertainty based sampling outperforming RAND's uniform at random sampling. Moreover, our experiments suggest that when labeling resources are limited, BALD based active learning is an attractive option for building personalized classification systems. We do note that BALD and RAND perform similarly when very few personalization instances are available. This may be due to the uncertainty estimates being poor in the very few personalization instances regime.

We compared our approach to the existing *state-of-the-art* in gesture personalization [34] on MSRC and ChaLearn datasets (Fig. 6). Yao et al. [34] presented three personalization methods: *full personalization*, which refers to fully re-training random forest classifiers given personalization data, *adaptive personalization*, which refers to adapting the parameters of pre-trained random forests given personalization data, and a *portfolio* approach, where a library of random forest classifiers are pre-trained and the best performing portfolio member is used to classify data from a new subject. We observe that on MSRC, both RAND and BALD outperformed all of the competing methods when the number of personalization instances per gesture class is greater than two. On ChaLearn, BALD outperformed portfolio and adaptive schemes and is within noise of full personalization after observing five personalization instances.

7. Conclusions

We developed a personalized gesture recognition system using a hierarchical Bayesian neural network and described algorithms for performing posterior inference. We illustrated the benefits of the hierarchical model over baselines that ignore subject-specific gesture variations and demonstrated the scalability of the model's capacity to learn complex feature-label mappings. Finally, we used the inferred posterior distributions over weights to guide active learning procedures for personalizing pre-trained models to new users. Our posterior driven active learning algorithm consistently outperformed selecting gestures at random. Further extensions of this work may include expanding this formulation to simultaneously localize as well as classify gestures from an input stream, as well as testing this framework on personalization challenges in other domains.

Acknowledgments. The authors wish to thank Jessica Hodgins, Leonid Sigal, Scott Watson, Jamie Robertson, and Michael Holton. This work was supported in part by Disney Research and NSF grants 1551572 and 1337866.

References

[1] J. Alon, V. Athitsos, Q. Yuan, and S. Sclaroff. Simultaneous localization and recognition of dynamic hand gestures. In *Seventh IEEE Workshops on Application of Computer Vision, 2005. WACV/MOTIONS'05*, volume 2, pages 254–260. IEEE, 2005.

[2] C. M. Bishop. Pattern recognition. *Machine Learning*, 2006.

[3] C. Blundell, J. Cornebise, K. Kavukcuoglu, and D. Wierstra. Weight uncertainty in neural network. In *Proceedings of the 32nd International Conference on Machine Learning (ICML-15)*, pages 1613–1622, 2015.

[4] L. Bottou. Stochastic gradient learning in neural networks. *Proceedings of Neuro-Nimes*, 91(8), 1991.

[5] W. L. Buntine and A. S. Weigend. Bayesian backpropagation. *Complex systems*, 5(6):603–643, 1991.

[6] W.-S. Chu, F. De la Torre, and J. F. Cohn. Selective transfer machine for personalized facial action unit detection. In *Computer Vision and Pattern Recognition (CVPR), 2013 IEEE Conference on*, pages 3515–3522. IEEE, 2013.

[7] S. D. Connell and A. K. Jain. Writer adaptation for online handwriting recognition. *Pattern Analysis and Machine Intelligence, IEEE transactions on*, 24(3):329–346, 2002.

[8] S. Escalera, J. Gonzàlez, X. Baró, M. Reyes, O. Lopes, I. Guyon, V. Athitsos, and H. Escalante. Multi-modal gesture recognition challenge 2013: Dataset and results. In *Proceedings of the 15th ACM on International conference on multimodal interaction*, pages 445–452. ACM, 2013.

[9] J. R. Finkel and C. D. Manning. Hierarchical bayesian domain adaptation. In *Proceedings of Human Language Technologies: The 2009 Annual Conference of the North American Chapter of the Association for Computational Linguistics*, pages 602–610. Association for Computational Linguistics, 2009.

[10] S. Fothergill, H. Mentis, P. Kohli, and S. Nowozin. Instructing people for training gestural interactive systems. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, pages 1737–1746. ACM, 2012.

[11] A. Gelman and J. Hill. *Data Analysis Using Regression and Multilevel/Hierarchical models*. Cambridge University Press, 2006.

[12] S. J. Gershman and N. D. Goodman. Amortized inference in probabilistic reasoning. In *Proceedings of the 36th Annual Conference of the Cognitive Science Society*, 2014.

[13] M. Ghosh, T. Maiti, D. Kim, S. Chakraborty, and A. Tewari. Hierarchical Bayesian neural networks: an application to a prostate cancer study. *Journal of the American Statistical Association*, 99(467):601–608, 2004.

[14] N. Houlsby, F. Huszár, Z. Ghahramani, and M. Lengyel. Bayesian active learning for classification and preference learning. *arXiv preprint arXiv:1112.5745*, 2011.

[15] A. Joshi, S. Ghosh, M. Betke, and H. Pfister. Hierarchical Bayesian neural networks for personalized classification. In *Neural Information Processing Systems Workshop on Bayesian Deep Learning*, 2016.

[16] A. Joshi, C. Monnier, M. Betke, and S. Sclaroff. A random forest approach to segmenting and classifying gestures. In *2015 IEEE International Conference on Automatic Face & Gesture Recognition and Workshops (FG 2015)*. IEEE, 2015.

[17] W. Kienzle and K. Chellapilla. Personalized handwriting recognition via biased regularization. In *Proceedings of the 23rd International Conference on Machine Learning*, pages 457–464. ACM, 2006.

[18] D. P. Kingma, T. Salimans, and M. Welling. Variational dropout and the local reparameterization trick. In *Advances in Neural Information Processing Systems*, 2015.

[19] D. P. Kingma and M. Welling. Stochastic gradient VB and the variational auto-encoder. In *Second International Conference on Learning Representations, ICLR*, 2014.

[20] Y. Liang and A. G. Kelemen. Hierarchical Bayesian neural network for gene expression temporal patterns. *Statistical Applications in Genetics and Molecular Biology*, 3(1):1–23.

[21] D. J. MacKay. A practical Bayesian framework for backpropagation networks. *Neural computation*, 4(3):448–472, 1992.

[22] R. M. Neal. *Bayesian learning for neural networks*, volume 118. Springer Science & Business Media, 2012.

[23] N. Neverova, C. Wolf, G. W. Taylor, and F. Nebout. Multi-scale deep learning for gesture detection and localization. In *Computer Vision-ECCV 2014 Workshops*, pages 474–490. Springer, 2014.

[24] R. Ranganath, S. Gerrish, and D. M. Blei. Black box variational inference. In *AISTATS*, pages 814–822, 2014.

[25] D. J. Rezende, S. Mohamed, and D. Wierstra. Stochastic backpropagation and approximate inference in deep generative models. In *Proceedings of The 31st International Conference on Machine Learning*, pages 1278–1286, 2014.

[26] K. Saenko, B. Kulis, M. Fritz, and T. Darrell. Adapting visual category models to new domains. In *Computer Vision–ECCV 2010*, pages 213–226. Springer, 2010.

[27] K. Shinoda and C.-H. Lee. A structural Bayes approach to speaker adaptation. *Speech and Audio Processing, IEEE Transactions on*, 9(3):276–287, 2001.

[28] Y. Song, D. Demirdjian, and R. Davis. Multi-signal gesture recognition using temporal smoothing hidden conditional random fields. In *2011 IEEE International Conference on Automatic Face & Gesture Recognition and Workshops (FG 2011)*, pages 388–393. IEEE, 2011.

[29] Y. Song, D. Demirdjian, and R. Davis. Tracking body and hands for gesture recognition: Natops aircraft handling signals database. In *Automatic Face & Gesture Recognition and Workshops (FG 2011), 2011 IEEE International Conference on*, pages 500–506. IEEE, 2011.

[30] T. Starner and A. Pentland. Real-time american sign language recognition from video using hidden markov models. In *Motion-Based Recognition*, pages 227–243. Springer, 1997.

[31] T. Tieleman and G. Hinton. Lecture 6.5-rmsprop: Divide the gradient by a running average of its recent magnitude. *COURSERA: Neural networks for machine learning*, 2012.

[32] M. Titsias and M. Lázaro-gredilla. Doubly stochastic variational Bayes for non-conjugate inference. In *Proceedings*

of the 31st International Conference on Machine Learning (ICML-14), pages 1971–1979, 2014.

[33] J. Yang, R. Yan, and A. G. Hauptmann. Cross-domain video concept detection using adaptive SVMs. In *Proceedings of the 15th International Conference on Multimedia*, pages 188–197. ACM, 2007.

[34] A. Yao, L. Van Gool, and P. Kohli. Gesture recognition portfolios for personalization. In *Computer Vision and Pattern Recognition (CVPR), 2014 IEEE Conference on*, pages 1923–1930. IEEE, 2014.