INFINITE HIDDEN MARKOV MODELS AND ISA FEATURES FOR UNUSUAL-EVENT DETECTION IN VIDEO

Iulian Pruteanu-Malinici and Lawrence Carin

Department of Electrical and Computer Engineering, Duke University, Durham, NC, 27705

Email: {ip6, lcarin}@ee.duke.edu

ABSTRACT

We address the problem of unusual-event detection in a video sequence. Invariant subspace analysis (ISA) is used to extract features from the video, and the time-evolving properties of these features are modeled via an infinite hidden Markov model (iHMM), which is trained using "normal"/"typical" video data. The iHMM automatically determines the proper number of HMM states, and it retains a full posterior density function on all model parameters. Anomalies (unusual events) are detected subsequently if a low likelihood is observed when associated sequential features are submitted to the trained iHMM. A hierarchical Dirichlet process (HDP) framework is employed in the formulation of the iHMM. The evaluation of posterior distributions for the iHMM is achieved in two ways: via MCMC and using a variational Bayes (VB) formulation.

Index Terms—Hidden Markov models, Dirichlet process, Variational Bayes

1. INTRODUCTION

The automatic detection of infrequent events is a problem that has recently attracted considerable attention. Such events are often referred to as being unusual, abnormal or rare [1, 2]. Anomaly detection, in the context of computer vision, is the process whereby a baseline of normal behavior is established with deviation from this norm triggering an alert.

We utilize the infinite hidden Markov model (iHMM) framework [3] to model the sequential characteristics of typical video, and employ invariant subspaces as features based on an invariant-subspace analysis (ISA) [4]. Motivated by the desire to handle complex scenes, feature extraction is not performed on elements (e.g., moving objects) within the scene, but on the entire scene; for complex environments, involving multiple and overlapping moving entities, feature extraction linked directly to such objects is often difficult, and therefore this step is avoided here entirely. In our work there is also no initial step of background subtraction or removal.

To address the complexity of time-series data, as in video, parametric techniques such as state-space models and differential-equation models have been employed [5, 6]. The approach outlined in this paper uses a semi-parametric HMM formulation, for which the form of the model (the HMM) is specified, but the number of underlying states is addressed non-parametrically.

Approaches utilizing HMMs have been shown to be effective in video detection, but are hampered by having to choose model architectures with appropriate complexities [7]. Typically, state decomposition has been performed in an *ad-hoc* manner [8], requiring trial and error for manually selecting the model structure, (*e.g.*, number of states). Concerning the iHMM considered here, there are in principle an infinite number of parameters in the transition matrix and observation matrix [9], although in practice the posterior density function on the number of states is peaked about a finite number of states characteristic of the data. We therefore avoid the problem of selecting a fixed number of HMM states. The iHMM was first introduced in [3], and to our knowledge this paper represents its first use in video analysis. Moreover, in [3] the inference was performed using MCMC, where here (we believe for the first time) variational Bayesian [10] inference is also considered.

2. INVARIANT SUBSPACE ANALYSIS

A traditional approach for video-based feature extraction is to use linear transformations, for which a given feature is computed as the inner product of the input data with a particular basis. A problem with linear features is their lack of invariance with respect to spatial shifts or phase changes. Kohonen [11] developed the concept of invariant-feature subspaces as an abstract approach to representing features. This concept states that one may consider an invariant feature as a linear subspace in a feature space. The value of the invariant feature is given by the norm of the projection of the given data point on that subspace. A feature subspace can be represented by a set of orthogonal basis vectors $\boldsymbol{w}_t$, $t = 1,...,d$, where d is the dimension of the subspace (in our case $d = 4$). Therefore the value of the invariant feature in any given subspace is given by

$$s^{inv} = \sqrt{\sum_{t=1}^{d} \left\langle \boldsymbol{w}_t, \boldsymbol{x} \right\rangle^2}, \qquad (1)$$

where $\boldsymbol{x} = (x_1, x_2, ..., x_n)^T$ is a vector of observed variables (in our case an image defined by a vector of pixel gray levels) and $\left\langle \boldsymbol{w}_t, \boldsymbol{x} \right\rangle$ denotes an inner product. This is equivalent to computing the distance between the input vector $\boldsymbol{x}$ and a general linear combination of the basis vectors $\boldsymbol{w}_t$ of the feature subspace.

Learning the independent feature subspace representation can be achieved by gradient ascent of the log-likelihood of the data

$$\log L(\boldsymbol{x} \mid \boldsymbol{w}) = \sum_{k=1}^{Q} \sum_{j=1}^{J} \log p\left(\sum_{t \in S_j} \left\langle \boldsymbol{w}_t, \boldsymbol{x}_k \right\rangle^2 \right) + Q \cdot \log \left| \det \boldsymbol{W} \right|, \qquad (2)$$

where Q is the size of the data (number of images in the video sequence $\{\boldsymbol{x}_k\}_{k=1}^{Q}$), J is the number of independent feature subspaces, $\{S_j\}_{j=1}^{J}$ represents the set of indices of the $\boldsymbol{w}_t$'s

belonging to the subspace of index j, and $\boldsymbol{W}$ is a matrix containing the filters $\boldsymbol{w}_t$'s as its columns. For the probability density we choose a multidimensional version of the exponential distribution

$$\log p(s^{inv\,2}) = -\alpha\left[\, s^{inv\,2}\,\right]^{\frac{1}{2}} + \beta\,, \tag{3}$$

where the scaling constant α and the normalization constant β are determined so as to give a probability density compatible with the constraint of unit variance of the s_j^{inv}.

A stochastic gradient ascent of the log-likelihood is obtained as

$$\Delta \boldsymbol{w}_q \propto \boldsymbol{x}_k \left\langle \boldsymbol{w}_q, \boldsymbol{x}_k \right\rangle g\Big(\sum_{t \in S_{j(q)}} \left\langle \boldsymbol{w}_t, \boldsymbol{x}_k \right\rangle^2\Big)\,, \tag{4}$$

where $j(q)$ is the index of the subspace to which $\boldsymbol{w}_q$ belongs to, and $g(u) = u^{-\frac{1}{2}}$. After every step of (4) the $\boldsymbol{w}_q$'s are orthonormalized; for a variety of methods to perform this, see [12].

In Section IV we compare the effectiveness of the ISA against two popular methods for video feature extraction: shift-invariant wavelets (SIW) [13] and independent component analysis (ICA) [13]. It is important to note that each of these feature-extraction techniques process the entire scene under test directly; there is no initial step associated with extracting (for example) moving entities. This is avoided because of its difficulty and inaccuracy for complex scenes involving multiple overlapping moving entities (with potentially unanticipated shapes).

3. INFINITE HIDDEN MARKOV MODEL

Consider N groups of data, denoted $\{\{x_{ji}\}_{i=1}^{n_j}\}_{j=1}^{N}$; the statistics of each group of data is modeled via a mixture model. As is usually done, we here assume a Gaussian mixture model (GMM). A Dirichlet process (DP) model [9] is used to non-parametrically learn a GMM separately for each of the N data sets. The data-group-dependent DP encourages clustering; data within the same cluster, or mixture component, are shared when learning the associated mixture-component parameters, and the appropriate sharing/clustering mechanism is determined by the algorithm. In a hierarchical DP (HDP) [9] the base distributions of each of the N DPs are drawn from a shared DP, and this encourages appropriate sharing of data between the N data sets.

To construct an HDP, a probability measure G_0 is drawn from a DP with precision parameter γ and base distribution H

$$G_0 \mid \gamma, H \sim DP(\gamma, H)\,. \tag{5}$$

The distribution G_0 serves as the base distribution for the N DPs associated with the N data sets:

$$G_j \mid \alpha, G_0 \sim DP(\alpha, G_0)\,, \tag{6}$$

where α is the data-set-dependent precision parameter, which is here kept as the same for all data sets, for simplicity. Since G_0 is discrete with probability one, it is guaranteed that all G_j will use the same set of mixture components defined in G_0, but in different proportions.

One may summarize the following properties of the HDP. For each of the N data sets considered, we learn an associated GMM. The parameters of the mixture components are shared across the N data sets, but the mixture weights are distinct. By the construction of DP, each of the mixture components has, in principle, an infinite number of components. This suggests the use of such a model in an HMM, for which the state-dependent observation probabilities are modeled by single Gaussians. One of these GMMs is used to represent the probability of the next observation, given the current HMM state occupied [9]. Since each GMM has an infinite number of components, each now linked to an associated HMM state, the HMM in principle has an infinite number of states. The sharing of state-dependent parameters across the mixture components is critical to realizing the appropriate form of the HMM. While in principle there are an infinite number of states, as constituted by the HDP prior, the available training data yields a posterior that is distributed around a finite number of states. The iHMM does not select one – fixed – number of states, but maintains a full posterior.

As in conventional HMMs, the data associated with the iHMM are the observations $\boldsymbol{O} = \{o_1, o_2, ..., o_T\}$. If the previous (hidden) state visited is s_{t-1}, then the probability of observation o_t corresponds to a unique one of the aforementioned GMMs, jointly representing the probability of transiting from state s_{t-1} and then observing o_t; the associated mixture weights correspond to the probability of transitioning to the respective next state s_t (each mixture component representative of a particular state) and the associated Gaussian defines the probability of the observing o_t. The iHMM may be summarized in the following hierarchical manner

$$\begin{aligned}
&\boldsymbol{\beta} \mid \gamma \sim Stick(\gamma)\\
&\boldsymbol{\pi}_{s_t} \mid \alpha, \boldsymbol{\beta} \sim DP(\alpha, \boldsymbol{\beta})\\
&s_t \mid \boldsymbol{\pi}_{s_{t-1}} \sim Mult(\boldsymbol{\pi}_{s_{t-1}})\\
&\theta_k^* \mid H \sim H\\
&o_t \mid s_t, (\theta_k)_{k=1}^{\infty} \sim F(\theta_{s_t})
\end{aligned}\,, \tag{7}$$

where each observation is generated independently as $o_t \sim F(\theta_{s_t})$, $F(\cdot)$ is a Gaussian density and the random weight variables $\boldsymbol{\beta} = (\beta_1, \beta_2, ..., \beta_k)$ use a Beta distribution to partition a unit-length *stick* as follows

$$\beta_k = \beta_k' \prod_{l=1}^{k-1} (1 - \beta_l') \qquad \beta_k' \sim Beta(1, \gamma) \quad for \quad k = 1, 2, ... \tag{8}$$

The evaluation of posterior distributions for the iHMM was achieved in [14] using a Markov chain Monte Carlo (MCMC) method; we here consider MCMC inference as well as an approximate and efficient variational Bayes (VB) formulation.

From Bayes' rule, we have

$$p(\Phi \mid \boldsymbol{O}, \Psi) = \frac{p(\boldsymbol{O} \mid \Phi)\, p(\Phi \mid \Psi)}{\int p(\boldsymbol{O} \mid \Phi)\, p(\Phi \mid \Psi)\, d\Phi}\,, \tag{9}$$

where $\Phi = \{\boldsymbol{z}, \boldsymbol{m}, \boldsymbol{\beta}\}$ are hidden variables of interest ($\{\{z_{ji}\}_{i=1}^{n}\}_{j=1}^{N}$ denote the mixture component associated with x_{ji} and $\{\{m_{jk}\}_{k=1,2,...}\}_{j=1}^{N}$ denote the number of clusters using the same mixture component k within a group j) and $\Psi = \{\alpha, \gamma\}$ are hyper-parameters that determine the distribution of the model

parameters (below we discuss the placement of hyperpriors [14] on α and γ). Instead of directly estimating $p(\Phi \mid \boldsymbol{O}, \Psi)$, variational methods seek a distribution $q(\Phi)$ to approximate the true posterior distribution $p(\Phi \mid \boldsymbol{O}, \Psi)$. Consider the log marginal likelihood

$$\log p(\boldsymbol{O} \mid \Psi) = L(q(\Phi)) + D_{KL}(q(\Phi) \,\|\, p(\Phi \mid \boldsymbol{O}, \Psi)), \tag{10}$$

where

$$L(q(\Phi)) = \int q(\Phi) \log \frac{p(\boldsymbol{O} \mid \Phi) p(\Phi \mid \Psi)}{q(\Phi)} d\Phi, \tag{11}$$

and

$$D_{KL}(q(\Phi) \,\|\, p(\Phi \mid \boldsymbol{O}, \Psi)) = \int q(\Phi) \log \frac{\mathrm{q}(\Phi)}{\mathrm{p}(\Phi \mid \boldsymbol{O}, \Psi)} d\Phi. \tag{12}$$

$D_{KL}(q(\Phi) \,\|\, p(\Phi \mid \boldsymbol{O}, \Psi))$ is the KL divergence between the approximate $q(\Phi)$ and the true posterior $p(\Phi \mid \boldsymbol{O}, \Psi)$. The approximation of the true posterior $p(\Phi \mid \boldsymbol{O}, \Psi)$ using $q(\Phi)$ is realized by minimizing $D_{KL}(q(\Phi) \,\|\, p(\Phi \mid \boldsymbol{O}, \Psi))$.

For computational convenience, $q(\Phi)$ is expressed in a factorized form, with the same functional form as the priors $p(\Phi \mid \Psi)$ and each parameter represented by its own conjugate prior. For the iHMM model, we assume

$$q(\Phi) = q(\boldsymbol{z}, \boldsymbol{m}, \boldsymbol{\beta}) = q(\boldsymbol{z}) q(\boldsymbol{m}) q(\boldsymbol{\beta}), \tag{13}$$

where $q(\boldsymbol{z})$, $q(\boldsymbol{m})$ and $q(\boldsymbol{\beta})$ have the following forms:

$$q(z_{ji} = k \mid \boldsymbol{z}^{-ji}, \boldsymbol{m}, \boldsymbol{\beta}) \propto \propto \begin{cases} (n_{jk}^{-ji} + \alpha\beta_k) f_k^{-x_{ji}}(x_{ji}) \dfrac{\alpha\beta_{s_{t+1}} + n_{ks_{t+1}}}{\alpha + \sum_{k'=1}^{K} n_{kk'}}, & if \quad k \in (1, \ldots, K) \\ \alpha\beta_{\mathrm{u}} f_{k^{new}}^{-x_{ji}}(x_{ji}) \beta_l, & if \quad k = k^{new} \end{cases} \tag{14}$$

$$q(m_{jk} = m \mid \boldsymbol{z}, \boldsymbol{m}^{-jk}, \boldsymbol{\beta}) = \frac{\Gamma(\alpha\beta_k)}{\Gamma(\alpha\beta_k + n_{jk})} s(n_{jk}, m) (\alpha\beta_k)^m \tag{15}$$

$$q(\beta_1, \ldots, \beta_K, \beta_u \mid \boldsymbol{z}, \boldsymbol{m}) \propto Dir(\sum_j m_{j1}, \ldots, \sum_j m_{jK}, \gamma) \tag{16}$$

Once we learn the hyperparameters of these variational distributions from the data, we obtain the approximation of $p(\Phi \mid \boldsymbol{O}, \Psi)$ by $q(\Phi)$. The joint distribution of Φ and observations $\boldsymbol{O}$ is given as

$$\begin{aligned} p(\boldsymbol{O}, \Phi) &= p(\boldsymbol{O}, \boldsymbol{z}, \boldsymbol{m}, \boldsymbol{\beta} \mid \Psi) \\ &= p(\boldsymbol{\beta}) p(\boldsymbol{m}) p(\boldsymbol{z} \mid \boldsymbol{m}, \boldsymbol{\beta}, o), \end{aligned} \tag{17}$$

where priors $p(\boldsymbol{z})$, $p(\boldsymbol{m})$ and $p(\boldsymbol{\beta})$ are given in (14), (15) and (16) respectively. All parameters $\Psi = \{\alpha, \gamma\}$ in these prior distributions are assumed to be set.

We substitute (13) and (17) into (11) to yield

$$\begin{aligned} L(q) = \int q(\boldsymbol{z}) q(\boldsymbol{m}) q(\boldsymbol{\beta}) \cdot \{ &\log p(\boldsymbol{\beta}) + \log p(\boldsymbol{m}) + \\ &+ \log p(\boldsymbol{z} \mid \boldsymbol{m}, \boldsymbol{\beta}, o) - \log q(\boldsymbol{z}) - \log q(\boldsymbol{m}) - \log q(\boldsymbol{\beta}) \}. \end{aligned} \tag{18}$$

The optimization of the lower bound in (18) is realized by taking functional derivatives with respect to each of the $q(\cdot)$ distributions while fixing the other q distributions and setting $\frac{\partial L(q)}{\partial q(\cdot)} = 0$ to find the distribution $q(\cdot)$ that increases L [15]. The local maximum of the lower bound $L(q)$ is achieved by iteratively updating the parameters of the variational distributions $q(\cdot)$. We terminate the algorithm when the change in $L(q)$ is negligibly small.

4. EXPERIMENTAL RESULTS

Performance of the iHMM has been considered for video data collected in an outdoor environment using a Canon VB-C50iR network camera. The video was collected at 8~10 frames per second with 320 x 240 pixel resolution. The data contains 35,000 images recorded at different moments in time, with each video sequence comprising 20 consecutive images. The iHMM was trained on the "normal" events; unusual events, in our case are defined by the presence of trucks and bike-riders (the area under test is typically characterized by walking individuals and cars), are detected through the low likelihood resulting from the trained iHMM.

We divide each frame of a video sequence into $V = 16$ spatial blocks $\{\boldsymbol{B}_{1,1}, \boldsymbol{B}_{1,2}, \ldots, \boldsymbol{B}_{4,4}\}$ and each block corresponds to an 80 x 60 pixel image area (the use of blocks reduces feature vector dimensionality, and the 16 associated iHMMs are linked to specific characteristics of the local scene within the block).

We first compared the effectiveness of the ISA against shift-invariant wavelets [13] and an independent component analysis (ICA) [13]. For this evaluation, we considered the features associated with one particular block, $\boldsymbol{B}_{3,1}$; see Fig. 2. The training data were 1,000 video sequences of pedestrians. Eight video sequences, (Table 1) corresponding to "normal" events were then used to test, for each of which the log-likelihood was computed using the corresponding iHMM model.

Table 1 shows the average classification results, for typical behavior, in the form of the log-likelihood using the three feature-extraction methods. The ISA features yield more stable log-likelihoods: small changes (shifts) in the input image cause larger variations in the distribution of the likelihood of both ICA features and the shift-invariant wavelets. Another advantage of using ISA features is that their dimension is much smaller than that of the shift-invariant wavelets (40 as opposed to 1,200); a high dimension of shift-invariant wavelets causes implementation problems (not enough training data). For these reasons, we use the ISA features for our detection problem. The results in Table 1 were computed using VB, with MCMC yielding comparable results.

	1	2	3	4	5	6	7	8
ICA	-28.2	-24.3	-27.5	-18.9	-20.8	-29.2	-31.2	-22.7
SIW	-18.7	-20.2	-18.4	-20.8	-19.5	-19.4	-21.3	-20.7
ISA	-19.4	-19.2	-20.1	-20	-19.7	-19.9	-20.3	-20.1

Table 1. Comparison of independent component analysis, shift-invariant wavelets, and invariant subspace (ISA) analysis.

In the results presented here, the posterior density function of both the MCMC and VB solutions indicated that from three to ten states were required of the HMM, depending on the complexity of the associated video block. We re-emphasize that the iHMM does not employ a single HMM but rather a full posterior distribution on the model structure.

We evaluated the iHMM for unusual-event detection in a video sequence, using both MCMC and VB implementations. In these representative examples we used $n = 20$ video sequences to test the iHMM performance. For each block in our testing data, we evaluated the log-likelihood of being generated from that respective block's trained iHMM model. For example,

classification results for block $B_{3,2}$ for the first video sequence are shown in Fig. 1. We use a histogram to graphically display the log-likelihoods corresponding to both normal and abnormal sequences from testing data considered.

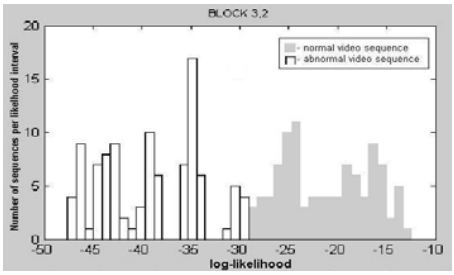

Fig. 1. Classification results for block $B_{3,2}$

Figure 2 presents snapshots from three example and representative sequences that highlight the effectiveness of the MCMC-based method. In each example, we display three frames and the corresponding log-likelihood plots as explained above; only the middle two blocks of each image are presented here. Example one contains normal walking behavior, example two an abnormal bike-rider, and example three an abnormal truck (from top to bottom in Fig. 2). The log-likelihoods corresponding to abnormal events are bolded. We also evaluated the effectiveness of the VB implementation on the same three sequences. The truck is detected as being abnormal via VB, whereas the first two examples (normal walking and the bike-rider) are both detected as being normal, despite the fact that the latter constitutes abnormal behavior. The VB version causes higher likelihoods corresponding to normal behavior (getting closer to the abnormal-event likelihoods), making abnormal-event detection more difficult.

In our analysis of a large set of video, the MCMC method works better than the VB version because the latter is an approximate inference algorithm. However, the computation of MCMC is expensive; it requires roughly 10 hours of CPU in non-optimized Matlab™ on a Pentium IV PC with a 2.1 GHz CPU to *train* the model, while VB requires less than 15 minutes. For both models, the *testing* is fast (a few seconds in non-optimized Matlab™).

5. CONCLUSION AND DISCUSSION

We have considered an algorithm that combines the advantages of the iHMM framework with the independent subspace analysis for anomaly detection. This framework is well suited for cases in which collecting sufficient unusual data is impractical or cannot be defined in advance.

Two inference tools are considered for the iHMM: an MCMC solution based on a Gibbs sampler, and a VB approach via maximizing a variational lower bound. We compared MCMC and VB implementations and showed that VB provides an efficient alternative to MCMC (being less expensive to train) but demonstrated that VB sometimes incorrectly detects abnormal scenarios in testing.

The proposed algorithm has several limitations that we intend to address in future work. First, we will consider developing models to utilize the information between adjacent blocks. In addition, based on the efficiency of the VB framework and the performance of the MCMC algorithm, we will consider combining the two methods (MCMC-VB) in future work.

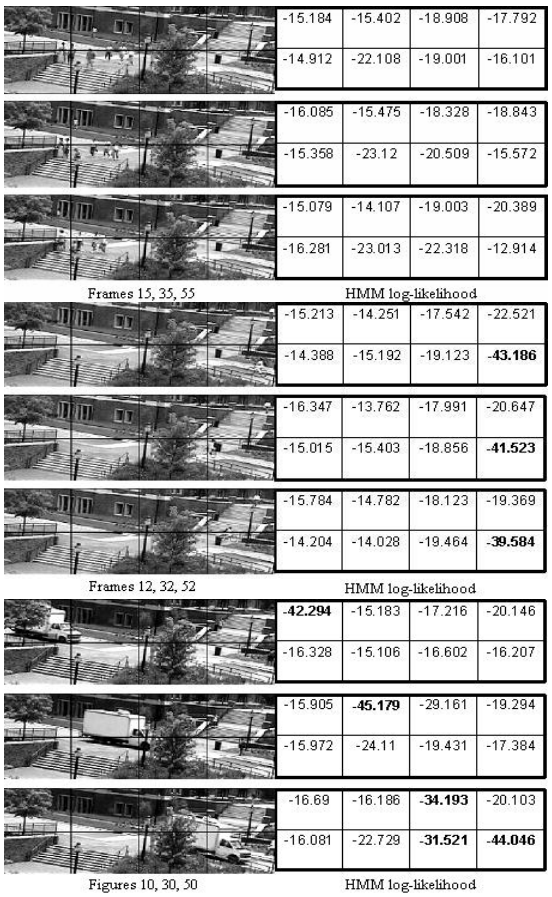

Fig. 2. Video sequence classification using MCMC inference

6. REFERENCES

[1] C. Stauffer and W. Eric, "*Learning patterns of activity using real-time tracking*", IEEE Trans. on PAMI, August 2000.

[2] H. Zhong, "*Detecting unusual activity in video*", IEEE CVPR, 2004.

[3] M. Beal *et al.*, "*The infinite hidden Markov model*", Advances in Neural Information Processing Systems, Cambridge, MIT Press, 2002.

[4] A. Hyvarinen *et al.*, "*Emergence of phase and shift-invariant features by decomposition of natural images*", Natural Computation, 2000.

[5] A. Dubey *et al.*, "*Clustering protein sequence and structure space with infinite Gaussian mixture models*", Pacific Symposium, 2004.

[6] M. Medvedovic and S. Sivaganesan, "*Bayesian infinite mixture model based clustering of gene profiles*", Bioinformatics, 2002.

[7] A. Kale and A. Rajagopalan, "*Gait-based recognition of humans using continuous HMMs*", IEE, FGR, 2002.

[8] P. Runkle *et al.*, "*Hidden Markov models for multi-aspect target classification*", IEEE Transactions on Signal Processing, July 1999.

[9] T. Ferguson, "*A Bayesian analysis of some nonparametric problems*", Annals of Statistics, vol. 1, no. 2, 1973.

[10] D. MacKay, "*Ensemble Learning for Hidden Markov Models*," technical report, Dept. of Physics, Univ. of Cambridge 1997.

[11] T. Kohonen, "*Emergence of invariant-feature detectors in the adaptive-subspace self-organizing map*", Biological Cybernetics, 1996.

[12] A. Hyvarinen *et al.*, "*A fast fixed-point algorithm for independent component analysis*", Neural Computation, 1997.

[13] S. Mallat, "*A wavelet tour of signal processing*", Academic Pres, 1999.

[14] Y. Teh *et al.*, "*Sharing clusters among elated groups: Hierarchical Dirichlet processes*", U. C. Berkeley Statistics, Technical report, 2004.

[15] M. Beal, "*Variational algorithms for approximate Bayesian inference*", Ph.D. dissertation, Gatsby Computational Neuroscience Unit, 2003.