Action Classification using a Discriminative Non-Parametric Hidden Markov Model

Natraj Raman, S.J.Maybank, and Dell Zhang
Department of Computer Science and Information Systems, Birkbeck, University of London

ABSTRACT

We classify human actions occurring in videos, using the skeletal joint positions extracted from a depth image sequence as features. Each action class is represented by a non-parametric Hidden Markov Model (NP-HMM) and the model parameters are learnt in a *discriminative* way. Specifically, we use a Bayesian framework based on Hierarchical Dirichlet Process (HDP) to automatically infer the cardinality of hidden states and formulate a discriminative function based on distance between Gaussian distributions to improve classification performance. We use elliptical slice sampling to efficiently sample parameters from the complex posterior distribution induced by our discriminative likelihood function. We illustrate our classification results for action class models trained using this technique.

Keywords: action classification, depth image, HDP-HMM, discriminative, elliptical slice sampling.

1. INTRODUCTION

Recognizing actions in videos has applications in diverse areas such as search and navigation of video sequences, smart surveillance and natural user interfaces for human-computer interaction. We consider the action classification problem in which the joint positions of a human in each video frame are available. A straight forward mechanism for modeling this time-series data is a Hidden Markov Model (HMM) that has a discrete hidden state at each time step and a density function for the observations conditional on the hidden state.

In classical parametric HMMs, the number of hidden states must be fixed in advance. In many applications this number is unknown a-priori and different numbers of states are tried during training. This model selection based approach treats the model complexity in an ad hoc manner. Instead, the Hierarchical Dirichlet Process HMM (HDP-HMM) provides a Bayesian framework for learning the number of states automatically from data, avoiding any prior assumptions about the cardinality of the states. It uses a set of Dirichlet Processes, one corresponding to each state, to link the Markovian dynamics [1,2].

Given observations X and their corresponding class labels Y, typically the HMM model parameters θ are sampled from the data likelihood $p(X \mid Y, \theta)p(Y|\theta)$ i.e. the parameters are selected to best *explain* the training examples. Instead, using the predictive likelihood $p(Y \mid X, \theta)$ during training often result in improved classification accuracy [3,4]. As observed in [5], we can decouple the parameters for the predictive distribution and the input data distribution by introducing new set of parameters $\theta'$and sample from $p(Y \mid X, \theta)\ p(X| \theta')\ p(\theta, \theta')$. This way of learning the parameters usually performs better because it allows compensating differences between the distribution specified by the model and the true distribution of the data. This discriminative likelihood can use negative examples from other classes when learning the parameters for a specific class and offers the flexibility of using unlabelled training observations.

Note that it is analytically intractable to sample posterior parameters from $p(Y \mid X, \theta)$. Hence we formulate this distribution in terms of data likelihood and distances between the observation density functions of the various classes. Intuitively, if the parameters of a class are distinct from the parameters of another class, when test examples not belonging to a class is evaluated, its pdf value will be lower and would result in overall improved classification accuracy.

We propose a mechanism to train HDP-HMM models that are optimized for classification. While learning the HDP-HMM parameters for a class, we also consider the parameters of other classes and sample parameters that maximize a *discriminative likelihood* function. Since it is not feasible to sample posterior parameters directly from this distribution, we use slice sampling [6] based techniques. Specifically, we place a Gaussian prior on the observation density parameters and use elliptical slice sampling [7] to efficiently sample the posterior. We are not aware of a discriminatively trained non-parametric HMM used in the literature.

2. RELATED WORK

A survey of research in human activity analysis can be found in [8]. The various techniques used for analyzing actions occurring in depth videos are discussed briefly in [9]. In particular, [10] reports good results for classifying actions by associating each joint position with a local occupancy feature and characterizing actions using a subset of these joint positions called an “actionlet”. The temporal structure is represented by a Fourier temporal pyramid.

Of late, there has been wider interest in the application of Bayesian non-parametrics for video analysis. In [11], a beta process driven HMM is used to obtain a temporal segmentation of each video into coherent behaviors for unsupervised activity discovery. In [12], a hierarchical non-parametric Bayesian model is used to segment common actions and cluster them into behaviors. A HDP-HMM based method that jointly segments and classifies actions with the ability to discover new classes as they occur is discussed in [13].

Discriminative training of HMMs is very popular in the automatic speech recognition (ASR) community. [14] provides an overview of the various discriminative learning criteria used in ASR and the procedures used to optimize these criteria. A thorough discussion of the various discriminative learning criteria such as MMI, MCE etc. and the issues associated with the optimization techniques can be found in [4].

3. DISCRIMINATIVE HDP-HMM

We assume that we are given the 3D joint positions of humans performing various actions in a video sequence. These joint positions are typically noisy functions of time. We model each action class using a HDP-HMM and learn the model parameters in a discriminative way by taking into account the parameters of the other action classes.

3.1Model Overview

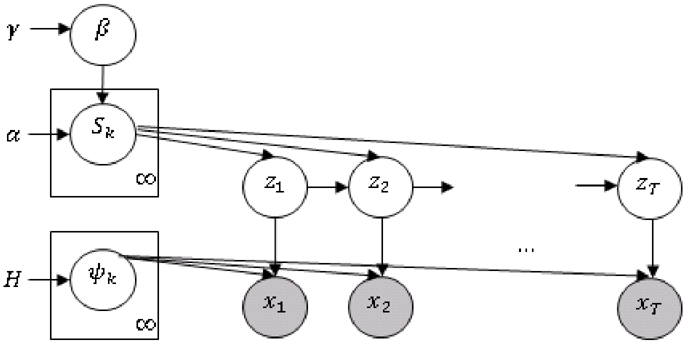

Figure 1. HDP-HMM model.

A HMM consists of two levels - an observation sequence $\{x_t\}_{t=1}^{T}, x_t \in \mathbb{R}^d$, a corresponding hidden state sequence $\{z_t\}_{t=1}^{T}, z_t \in \{ 1,2 \dots K\}$ that follows a Markov chain $z_t \perp z_{1:t-2} \mid z_{t-1}$ and $x_t \perp z_{1:t-1}, x_{1:t-1} | \, z_t$. Transitions between the states are parameterized as $\{S_{k,l}\}_{k=0,l=1}^{K}$, $S_{k,l} = P(z_t = l|z_{t-1} = k)$, $S_{0,l} = P(z_1 = l)$. The observation distribution is parameterized as $P(x_t|z_t = k) \sim F(\psi_k)$ where $\psi_k$are the natural parameters of the family F of distributions. Here, the observations are generated by Gaussian mixtures, with one Gaussian for each state and $\psi_k = (\mu_k, \Sigma_k)$. Let H be the prior corresponding to ψ. Let the state transitions have a Dirichlet prior with shared parameters ensuring that the transitions out of different states are coupled i.e. $\beta \sim Dir(\frac{\gamma}{K} \dots \frac{\gamma}{K})$ and $S_k \sim Dir(\alpha\beta_1 \dots \alpha\beta_K)$. Here α and γ are hyper parameters.

For a non-parametric HMM, the number K of states is unbounded and we can use a Hierarchical Dirichlet Process (HDP) to describe the priors. Given a base distribution G_0 and a concentration parameter $\alpha \in \mathbb{R}_+$, draws from a Dirichlet Process $DP(\alpha, G_0)$ will return random distributions containing values drawn from G_0 with α controlling the variability around it. HDP is a set of DPs coupled through a base distribution, with the base distribution itself being drawn from a DP. In the HMM context, each row in the state transition matrix is a DP that is tied to another top level DP i.e. $S_k \sim DP(\alpha, \beta)$ and $\beta \sim DP(\gamma, H)$. The model is shown in figure 1 and further details can be found in [1,2].

We are given i.i.d training data $X = \{x^n\}_{n=1}^N, Y = \{y^n\}_{n=1}^N$, where $x^n = x_1^n \dots x_T^n$ is an observation sequence and $y^n \in \{1 \dots C\}$ its corresponding action class. An observation $x_t \in \mathbb{R}^d$ at a time-step t consists of joint positions $\{P_i\}_{i=1}^{20} \in \mathbb{R}^3$. The HDP-HMM model parameters for a specific class are $\theta^c = \{\beta, S, \psi\}$ and the set of all model parameters is $\theta = \{\theta^c\}_{c=1}^C$.

3.2 Discriminative Likelihood Function

In order to improve classification accuracy, our intention is to draw posterior parameters from the distribution $p(Y \mid X, \theta)\, p(X| \theta')\, p(\theta, \theta')$ where θ' is a new set of parameters identical to θ. Following [5], this can be written as

$$p(\theta, \theta'|X, Y) \propto [p(Y|X, \theta)\, p(\theta)\, \delta(\theta, \theta')] * [p(X|\theta')\, p(\theta')] \tag{1}$$

Applying Gibbs sampling, let us sample θ' first and then sample θ given θ'. We can write $p(x) = \sum_c p(x, c)$ through inference and it is straight forward to sample posterior θ' from (2) since it involves only the data likelihood.

$$p(\theta'|X) \propto \prod_{n=1}^{N} \sum_c p\left(x^n \middle| c, \theta'^c\right) * p(c) * \; p(\theta') \tag{2}$$

It is analytically intractable to sample θ directly from the predictive likelihood. Hence we reformulate this distribution and sample parameters for each class one at a time, given the parameters of other classes as

$$p\left(\theta^c \middle| X, Y, \theta^{\setminus c}, \theta'\right) \propto \left[\prod_{n:y^n=c} p(x^n|c, \theta^c)\; p(\theta^c)\right] * \; p(\theta^c\,|\theta^{\setminus c}) \; * \; \delta(\theta^c, \theta'^c) \tag{3}$$

The third term on the right hand side of (3) controls the extent to which the predictive likelihood parameters differ from the data distribution parameters. We set $\delta(\theta^c, \theta'^c) = e^{-\xi\,|\theta^c - \theta'^c|}$. If ξ is large, then θ^c and θ'^c are likely to be similar in value.

The second term on the right hand side of (3) contributes to the *discriminative likelihood*. It encourages the parameters of a given class to repel away from the parameters of the other classes. Typically the HMM observation density parameters $\mu_{1\dots K}, \Sigma_{1\dots K}$ make the most significant contribution to discrimination. We write

$$p(\theta^c\,|\theta^{\setminus c}) \;\propto\; \prod_{k} \prod_{c' \in \setminus c} \prod_{k'} D\left(\mathcal{N}(\mu_k^c, \Sigma_k^c), \mathcal{N}\left(\mu_{k'}^{c'}, \Sigma_{k'}^{c'}\right)\right) \tag{4}$$

Here $D(.)$ is any measure such as Hellinger or Bhattacharya distance that computes distance between two normal distributions. Intuitively, if the total distance between the observation densities of a class and the densities of all other classes is large, then the parameters are sufficiently separated and will be discriminative enough.

4. SAMPLING PROCEDURE

For faster mixing, we block sample the state sequences using the HMM forward-backward algorithm i.e. given an observation sequence $x_{1:T}$, the state transition probabilities S and the observation density parameters ψ, we sample the state sequence $z_{1:T}$ at one go. In order to use the forward-backward algorithm, we use the weak limit approximation [2] to the Dirichlet process. We set the number of states K to a large value and during training fewer than K states are learnt.

Given the set of state sequences, we must sample the HDP-HMM parameters. We use the standard procedure as outlined in [1] to sample the state transition probabilities from the number of transitions and hyper parameters α, γ. We now sample from (3), the observation density parameters $\psi_{1:K} = (\mu_{1:K}, \Sigma_{1:K})$ given the set of observations $\mathcal{X}_{1:K}$ belonging to states. We do this one state at a time making use of slice sampling techniques. Specifically, using (3), we formulate a likelihood function

$$L\left(\mu_k^c, \Sigma_k^c \middle| \mathcal{X}_k^c, \theta^{\setminus c}, \theta'\right) = \mathcal{N}(\mathcal{X}_k^c; \mu_k^c, \Sigma_k^c) * \prod_{c' \in \setminus c} \prod_{k'} D\left(\mathcal{N}(\mu_k^c, \Sigma_k^c), \mathcal{N}\left(\mu_{k'}^{c'}, \Sigma_{k'}^{c'}\right)\right) * \; e^{-\xi\,|\theta^c - \theta'^c|} \tag{5}$$

In slice sampling, an auxiliary variable $u \sim \mathbb{U}[0, L(\phi^t)]$ is drawn from a likelihood function using the current state ϕ^t. From the current state, a new state is proposed and is accepted if $L(\phi^{t+1}) > u$. Elliptical slice sampling [7] provides a much more efficient way of proposing new states even for high dimensional variables if the variables have Gaussian priors. It defines a full ellipse around the current state and provides richer updates by adapting step-sizes effectively. We use Elliptical slice sampling to sample observation density parameters using the likelihood function in (5). We set a Gaussian prior for the mean $\mu \sim N(\mu_0, \Sigma_0)$. We assume diagonal co-variance and set independent log-normal priors for the standard-deviations i.e. $\log(\sqrt{\Sigma}^d) \sim N(\zeta, \Delta)$.

5. EXPERIMENTS

We illustrate the results of our approach using the MSR-Action3D data set [15]. This dataset has annotated 3D joint positions extracted from depth image sequences captured by an infrared light camera. The actions are performed in the context of interacting with a games console. Each action contains 21 instances and we use 60% of the instances for training and the rest for testing. We apply our algorithm for classifying 13 actions. Figure 2 shows the actions from this dataset we use in our experiments.

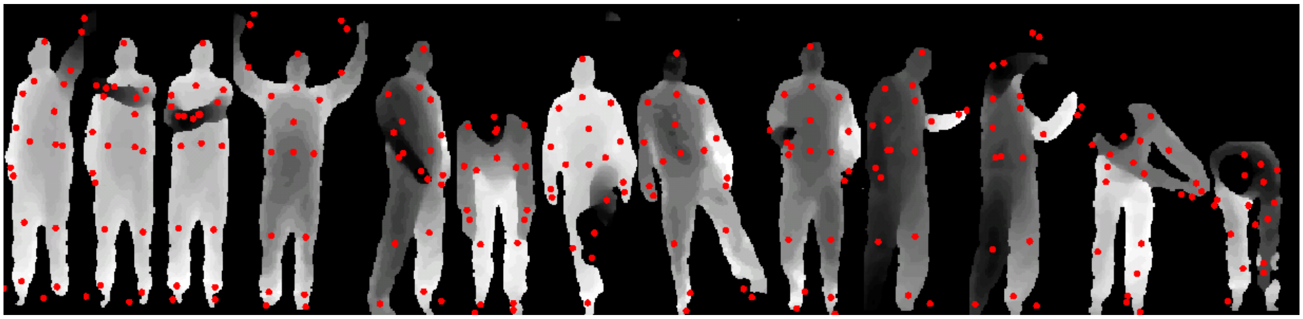

Figure 2. Example actions annotated with 3D joint positions from the MSR-Action3D [15] dataset.

From the 20 joint positions, we compute 19 relative joint positions based on a pre-defined skeleton hierarchy. By using relative positions as features we ensure invariance to uniform translation of the body. We use the Bhattacharya distance to compute the distance D between the normal distributions in (5). We use a gamma prior for the Dirichlet hyper parameters α, γ and set ξ used in (5) to a small value allowing θ^c and θ'^c to vary freely. For numerical convenience, we use log-likelihood.

We report an overall classification accuracy (proportion of correct classifications to total number of classifications) of 89.7%. The confusion matrix is shown in Figure 3.

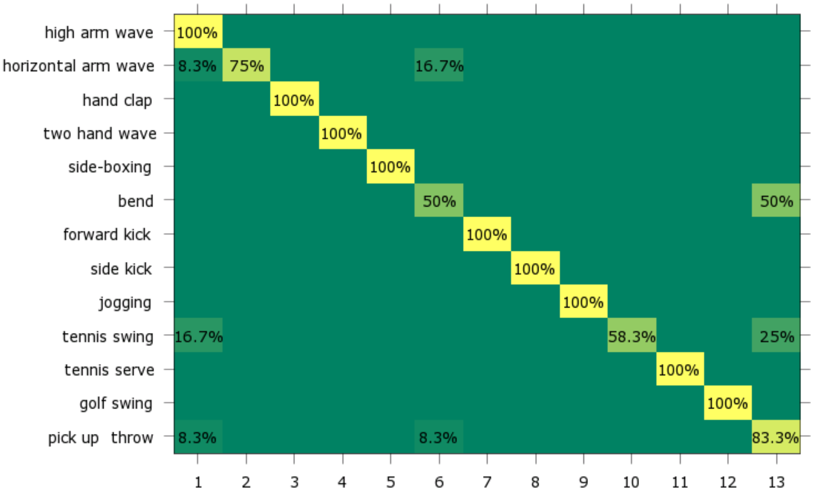

Figure 3. Confusion matrix for the classification results.

6. CONCLUSION

We have proposed here an action classification technique based on non-parametric HMMs with the parameters trained in a discriminative way. Specifically, we have formulated a flexible yet powerful discriminative likelihood function and applied elliptical slice sampling to efficiently sample posterior parameters. Our experiments have demonstrated the utility of this approach. We intend to introduce different discriminative criterions and apply this technique in the future to classify complex activities involving humans and objects.

REFERENCES

[1] Teh, Yee Whye, Michael I. Jordan, Matthew J. Beal, and David M. Blei. "Hierarchical dirichlet processes." Journal of the American Statistical Association 101.476 (2006).

[2] Fox, Emily B., Erik B. Sudderth, Michael I. Jordan, and Alan S. Willsky. "An HDP-HMM for systems with state persistence."Proceedings of the 25th international conference on Machine learning. ACM, (2008).

[3] Lin, Tien-ho, Naftali Kaminski, and Ziv Bar-Joseph. "Alignment and classification of time series gene expression in clinical studies." Bioinformatics 24.13 : i147-i155 (2008).

[4] He, Xiaodong, Li Deng, and Wu Chou. "Discriminative learning in sequential pattern recognition." Signal Processing Magazine, IEEE 25.5 : 14-36 (2008).

[5] Lasserre, Julia A., Christopher M. Bishop, and Thomas P. Minka. "Principled hybrids of generative and discriminative models." Computer Vision and Pattern Recognition, (2006).

[6] Neal, Radford M. "Slice sampling." Annals of Statistics 705-741 (2003).

[7] Murray, Iain, Ryan Prescott Adams, and David JC MacKay. "Elliptical slice sampling." arXiv preprint arXiv:1001.0175 (2009).

[8] Aggarwal, J. K., and Michael S. Ryoo. "Human activity analysis: A review."ACM Computing Surveys (CSUR) 43.3: 16 (2011).

[9] Han, Jungong, Ling Shao, Dong Xu, and Jamie Shotton. "Enhanced Computer Vision with Microsoft Kinect Sensor: A Review." IEEE Transactions on Cybernetics (2013).

[10] Wang, Jiang, Zicheng Liu, Ying Wu, and Junsong Yuan. "Mining actionlet ensemble for action recognition with depth cameras." Computer Vision and Pattern Recognition (CVPR), (2012).

[11] Hughes, Michael C., and Erik B. Sudderth. "Nonparametric discovery of activity patterns from video collections." Computer Vision and Pattern Recognition Workshops, (2012).

[12] Kooij, Julian FP, Gwenn Englebienne, and Dariu M. Gavrila. "A non-parametric hierarchical model to discover behavior dynamics from tracks." Computer Vision–ECCV (2012).

[13] Bargi, Ava, R. Y. D. Xu, and Massimo Piccardi. "An online HDP-HMM for joint action segmentation and classification in motion capture data." CVPRW, (2012).

[14] Jiang, Hui. "Discriminative training of HMMs for automatic speech recognition: A survey." Computer Speech & Language 24.4 : 589-608 (2010).

[15] Li, Wanqing, Zhengyou Zhang, and Zicheng Liu. "Action recognition based on a bag of 3d points." Computer Vision and Pattern Recognition Workshops (CVPRW), (2010).