ESANN'2009 proceedings, European Symposium on Artificial Neural Networks - Advances in Computational Intelligence and Learning. Bruges (Belgium), 22-24 April 2009, d-side publi., ISBN 2-930307-09-9.

Modeling pigeon behaviour using a Conditional Restricted Boltzmann Machine

Matthew D. Zeiler[1], Graham W. Taylor[1], Nikolaus F. Troje[2] and Geoffrey E. Hinton[1]

1- University of Toronto - Dept. of Computer Science
Toronto, Ontario M5S 2Z9 - Canada

2- Queen's University - Dept. of Psychology
Kingston, Ontario K7L 3N6 - Canada

Abstract. In an effort to better understand the complex courtship behaviour of pigeons, we have built a model learned from motion capture data. We employ a Conditional Restricted Boltzmann Machine (CRBM) with binary latent features and real-valued visible units. The units are conditioned on information from previous time steps to capture dynamics. We validate a trained model by quantifying the characteristic "head-bobbing" present in pigeons. We also show how to predict missing data by marginalizing out the hidden variables and minimizing free energy.

1 Introduction

Recent studies investigating the complex courtship behaviour of the pigeon, *Columbia livia*, demonstrate that pigeons show courtship responses not only to real partners, but also to video [1]. More recently, social behaviour in pigeons has been elicited by a virtual pigeon, driven by motion capture (mocap) data gathered from a real pigeon and rendered through a computer graphics engine [2]. Investigating avian social perception is an interesting problem, as it presents a "sandbox" in which we can experiment to better understand interactive social behaviour. A natural next step is to drive a virtual pigeon not by mocap data, but by a model learned from such data. This will give researchers both control over and insight into the complex factors underlying courtship behaviour.

The success of a temporal extension of Restricted Boltzmann Machines on modeling human motion [3] has prompted us to consider this powerful class of models for learning on mocap data captured from both single pigeons and pairs of pigeons in courtship. These models can learn complex dynamics directly from data, without imposing physics-based constraints. Their generative nature permits online synthesis of novel motion using only the learned weights and a few valid frames for initialization. Like a Hidden Markov Model (HMM), they can capture nonlinearities in the observation but their distributed binary hidden state is exponentially more powerful than the K-state multinomial used by an HMM. This allows the CRBM to capture longer-term dependencies.

In this paper, we concentrate on learning a generative model of the motion of a single pigeon. Before modeling courtship behaviour, we must first confirm that our model is capable of capturing the subtleties of pigeon motion, notably the characteristic "head-bobbing"[4]. We also focus on the more practical problem of predicting the location of the feet which are frequently occluded during capture.

ESANN'2009 proceedings, European Symposium on Artificial Neural Networks - Advances in Computational Intelligence and Learning. Bruges (Belgium), 22-24 April 2009, d-side publi., ISBN 2-930307-09-9.

2 The Conditional Restricted Boltzmann Machine

The CRBM is a non-linear generative model for time-series data that uses an undirected model with binary latent variables, $\mathbf{h}$, connected to a collection of "visible" variables, $\mathbf{v}$. The visible variables can use any distribution in the exponential family. We use real-valued Gaussian units in our experiments. At each time step, $\mathbf{v}$ and $\mathbf{h}$ receive directed connections from the visible variables at the last few time-steps. The model defines a joint probability distribution over $\mathbf{v}$ and $\mathbf{h}$, conditional on the past N observations and model parameters, θ:

$$p(\mathbf{v},\mathbf{h}|\{\mathbf{v}\}_{t-N}^{t-1},\theta) = \exp(-E\left(\mathbf{v},\mathbf{h}|\{\mathbf{v}\}_{t-N}^{t-1},\theta\right)/Z$$
$$E\left(\mathbf{v},\mathbf{h}|\{\mathbf{v}\}_{t-N}^{t-1},\theta\right)\big) = \sum_i \frac{(v_i - b_i^*)^2}{2\sigma_i^2} - \sum_j h_j b_j^* - \sum_{ij} w_{ij}\frac{v_i}{\sigma_i}h_j \qquad (1)$$

where Z is a constant called the partition function which is exponentially expensive to compute exactly. The dynamic biases, b_i^*, b_j^*, are affine functions of the past N observations. Such an architecture makes on-line inference efficient and allows us to train by minimizing contrastive divergence (for details, see [5]). Taylor et al. [3] showed that after training a CRBM on mocap data, a single model could synthesize novel motion of various styles without the need to keep a database of valid motions. The model was also used to perform on-line filling in of data lost during motion capture.

An important feature of the CRBM is that once it is trained, we can add layers like in a Deep Belief Network [6]. The previous layer CRBM is kept, and the sequence of hidden state vectors, while driven by the data, is treated as a new kind of "fully observed" data. The next level CRBM has the same architecture as the first (though it has binary visible units and we can change the number of hidden units) and is trained in the exact same way (Fig. 1). Upper levels of the network can then model higher-order structure. More layers aid in capturing multiple styles of motion, and permitting transitions between these styles [3].

3 Data gathering and preprocessing

Markers were placed at various locations on the head, torso, and both feet of a pigeon to capture the local movements of each body segment. Each pigeon was allowed to walk in an enclosed area and was recorded with an array of synchronized cameras. The collected data was cleaned to account for sensor noise and occlusion, providing (x,y,z) positions of each marker in mm with respect to a global coordinate system.

This data was converted to a hierarchy of coordinate systems relative to a body coordinate frame. The origin of the local coordinate systems for each foot and the head were defined as a translation and rotation relative to the body frame, which was in turn relative to the global coordinate system. All rotations were converted to an exponential map representation. As in [3], the root segment was expressed in a body-centred coordinate system which is invariant to ground-plane translations and rotations about the gravitational vertical. Finally all

ESANN'2009 proceedings, European Symposium on Artificial Neural Networks - Advances in Computational Intelligence and Learning. Bruges (Belgium), 22-24 April 2009, d-side publi., ISBN 2-930307-09-9.

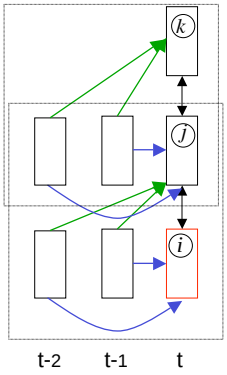

Fig. 1: Architecture of a two-layer CRBM

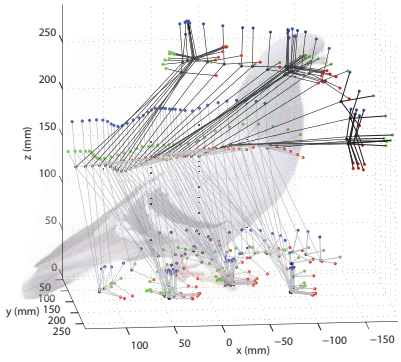

Fig. 2: The hold-phase present in generated motion. RGB points mark each frame's xyz axes.

data was scaled to have zero mean and unit variance (data is rescaled before measurement or playback). The final representation was 24 real values per frame: 6 degrees of freedom (dof) for each of 4 segments. We included all translational dof to account for the articulated, multi-segment nature of the neck and legs.

4 Experimental setup and discussion

4.1 Generation of novel motion

Following the procedure of [3], we trained a Gaussian-binary CRBM on 11583 frames of pigeon mocap data at 120 fps. A binary-binary CRBM was then trained on the real-valued probabilities of the hidden layer while driven by the training data. Each CRBM had 600 hidden units. All parameters used a learning rate of 1×10^{-3}, except for the autoregressive weights (1×10^{-5}). A momentum term was also used: 0.9 of the previous accumulated gradient was added to the current gradient. We used CD(10), i.e. 10 steps of alternating Gibbs sampling per each iteration of contrastive divergence. Both layers were conditioned on 12 previous frames. Generation of each frame involved alternating Gibbs sampling in the top two layers, followed by a single downward pass in the first CRBM. The previous frame plus a small amount of Gaussian noise was used for initialization.

If we are to carry out experiments to elicit response of real pigeons from a "virtual pigeon" driven by our model, we must ensure that our synthesized data captures the subtleties characteristic of pigeon motion. In addition to fixed foot plants, pigeons demonstrate complex "head-bobbing", defined by a distinct "thrust phase" and "hold-phase" where the head actually remains stationary [4]. We were able to generate pigeon motion that closely resembled the true motion capture data (Fig. 2). The collection of head coordinate frames shows the distinct hold-phase present in the generated motion. For videos of generated motion see: `http://www.matthewzeiler.com/videos/`.

Although the hold-phase is visually present, we have sought a quantitative

comparison to the real motion capture data. The hold-phase can be quantified by measuring rotation about the head frame's yaw axis and the path length of the head in the ground-plane with respect to the global coordinate system. A visual comparison of training data against data generated from our 2-layer CRBM model is shown in Fig. 3 along with a baseline model: a 12th order autoregressive (AR) model fit by regularized least squares.

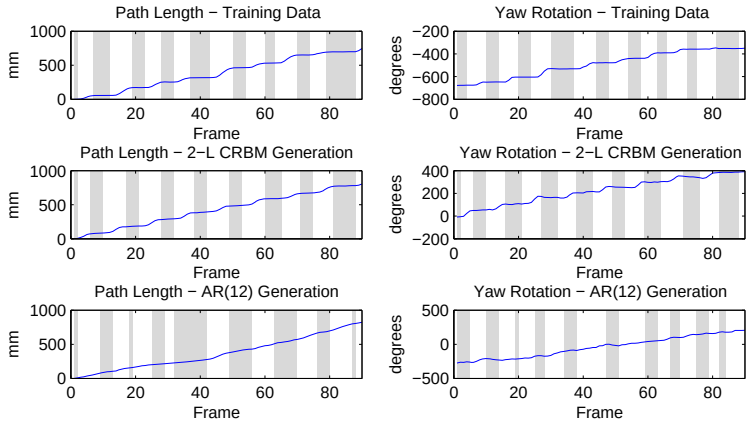

Fig. 3: Hold-Phases present in path length and yaw angles (shaded grey).

Each hold phase was detected by determining where the empirical second derivative of the respective time series was either a minimum or a maximum, which corresponds to the leveling of the hold-phase regions. In these regions the standard deviations were calculated and the mean of all these regions, along with corresponding calculation for the thrust-phases are shown in Table 1.

Model/Phase	Path Length (mm)	Yaw Angle (degrees)
Training data Hold Phase	0.58	0.96
Training data Thrust Phase	30.87	23.09
2-layer CRBM Hold Phase	3.58	2.71
2-layer CRBM Thrust Phase	31.18	19.69
AR(12) Hold Phase	11.54	6.68
AR(12) Thrust Phase	21.18	13.65

Table 1: Comparison of Mean Standard Deviation in Hold and Thrust Phases.

Data generated from the 2-layer CRBM clearly captures the two distinct phases as seen in Fig. 3 and in the differences between standard deviations. Also, since our model is stochastic, the std. dev. in the hold phase is greater than that of the mocap data. Since we are modeling ground-plane velocities instead of positions, noise introduced by sampling is integrated up and accumulates

during post-processing. The reason why our model exhibits a smaller standard deviation in the hold-phase for rotation compared to translation, may be an artifact of normalizing the data dimensions before training. While translational dimensions are expressed in mm before scaling (typically large values), rotational dimensions are expressed in exponential maps (small values) and so noise in the translational dimensions is exaggerated, relative to the rotations, when we rescale. The precision of the Gaussian units, $1/\sigma_i$, has been fixed to 1 in all of our experiments since it tends to work well in practice [3]. Either learning this parameter, or using a larger fixed value may improve these results.

4.2 Prediction of missing foot markers

Due to the relative size of torso and feet, as well as the swelling of feathers, markers on the feet of pigeons are frequently occluded. Since our model would benefit from the availability of more data, we would like to exploit all possible mocap trials. One option would be to ignore, or marginalize out any missing feet while training. Another is to use a CRBM trained on complete data to predict the missing feet. We carried out a series of experiments in which a 1-layer and 2-layer CRBM are used to predict the 6 dof corresponding to the right foot, given the other 18 dof. We compare to several baseline methods: a nearest neighbor search through the training set (based on the known data dimensions), an AR(12) model, as well as 1 and 2 layer static RBM models that do not use temporal information. 1-step prediction results are shown in Table 2.

Model	Mean Error (mm)
Nearest Neighbor	63.73
AR(12)	2.68
1-L RBM	26.43 ± 0.01
2-L RBM	31.31 ± 0.07
1-L CRBM (FE)	5.61
2-L CRBM (Gibbs)	8.27 ± 0.04

Table 2: Average 1-step prediction error for 7767 test frames. The std. dev. over 25 trials for stochastic models is also reported.

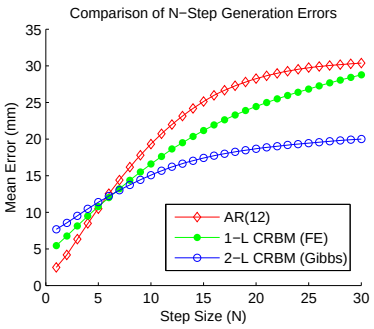

Fig. 5: Comparison of various models for N-step prediction.

For the 1-layer CRBM, we used a unique method to fill in missing markers. A method based on Gibbs sampling could be used for prediction by initializing unknown markers with their values at the previous time step and reconstructing only these markers on alternating Gibbs sampling steps as described in [3]. The downside of this approach, is that it is subject to noise. A better method may be to exploit the fact that the hidden variables are binary, and integrate them out

to arrive at the "free energy", given the model parameters and past observations:

$$F(\mathbf{v}|\{\mathbf{v}\}_{t-N}^{t-1},\theta)=\sum_i \frac{(v_i-b_i^*)^2}{2\sigma_i^2} - \sum_j \log\left(1+\exp(\sum_i w_{ij}\frac{v_i}{\sigma_i}+b_j^*)\right). \quad (2)$$

F is the negative log probability of an observation plus $\log Z$ (see Eq. 1). A setting of $\mathbf{v}$ that minimizes the free energy is found by fixing the known v_i and following the gradient of F with respect to the unknown v_i. In our tests we use a conjugate-gradient method initialized with the values at the previous frame.

One method for filling in missing data using the 2-layer CRBM is to generate as described earlier, but only replace missing dimensions during the final down-pass through the 1st CRBM. This has the disadvantage that sampling hiddens does not take into account the known visible dimensions of the current frame. Prediction can be improved by linearly blending both top-down input (from CRBM 2) and bottom-up input (from CRBM 1) to the first hidden layer. For models using temporal information, we also carried out N-step prediction (Fig. 5), showing that multi-layer models can capture longer-term dependencies.

5 Conclusion

We have proposed a model for pigeon motion based on a Conditional Restricted Boltzmann Machine. A model trained on mocap data from a single pigeon can synthesize realistic data. This is verified by videos, and through quantitative analysis of "head-bobbing". The model can be used to fill in missing foot data. Our results show that minimizing free energy with respect to the missing variables gives superior results to the Gibbs sampling method proposed in [3] for short term predictions. The benefits of adding a second layer are evident in longer-term prediction. Future work will be to drive a "virtual pigeon" using computer graphics with data generated by our model in an attempt to elicit social behaviour in pigeons. We also hope to train CRBMs on data captured from pairs of pigeons. This joint model could be used to predict the behaviour of one bird given the motion of another.

References

[1] B.J. Frost, N.F. Troje, and S. David. Pigeon courtship behaviour in response to live birds and video presentations. In *5th International Congress of Neuroethnology*, 1998.

[2] S. Watanabe and N.F. Troje. Towards a "virtual pigeon": A new technique to investigate avian social perception. *Animal Cognition*, 9:271–279, 2006.

[3] G.W. Taylor, G.E. Hinton, and S.T. Roweis. Modeling human motion using binary latent variables. In *NIPS 19*, pages 1345–1352, Cambridge, MA, 2007. MIT Press.

[4] Nikolaus F. Troje and Barrie J. Frost. Head-bobbing in pigeons: How stable is the hold phase? *The Journal of Experimental Biology*, 203(4):935–940, 2000.

[5] G.E. Hinton. Training products of experts by minimizing contrastive divergence. *Neural Comput*, 14(8):1771–1800, Aug 2002.

[6] G. E. Hinton, S. Osindero, and Y. W. Teh. A fast learning algorithm for deep belief nets. *Neural Comp.*, 18(7):1527–1554, 2006.