INFERRING 3D BODY POSE USING VARIATIONAL SEMI-PARAMETRIC REGRESSION

Yan Tian[‡,†], *Yonghua Jia*[‡], *Yuan Shi*[§], *Yong Liu*[†], *Hao Ji*[§], *Leonid Sigal*[¶]

‡Hikvision Digital Technology Co. Ltd, Hangzhou, P.R.China
†Beijing University of Posts and Telecommunications,Beijing, P.R.China
§Carnegie Mellon University, Pittsburgh, US
¶Disney Research, Pittsburgh, US

ABSTRACT

To deal with multi-modality in human pose estimation, mixture models or local models are introduced. However, problems with over-fitting and generalization are caused by our necessarily limited data, and the regression parameters need to be determined without resorting to slow and processor-hungry techniques, such as cross validation. To compensate these problems, we have developed a semi-parametric regression model in latent space with variational inference. Our method performed competitively in comparison to other current methods.

Index Terms— Image motion analysis, unsupervised learning, regression model, latent variable model

1. INTRODUCTION

Though abundant applications for marker-less motion capture in activity recognition and human computer interaction, monocular pose estimation remains a difficult task; Complexity of the pose, angle of the camera, the shadow or noise in the image make even an easy pose difficult to identify.

Most prior research in this area can be classified into two types of approach: *generative* and *discriminative*. *Generative* approaches [1] define an image formation model by predicting the appearance of a body given a hypothesized state of the body (pose); an inference framework is then used to infer the posterior. Since the inference often takes the form of non-convex search in a high-dimensional space of body articulations, these methods are computationally expensive, and can suffer from local convergence. *Discriminative* approaches [2] avoid building an explicit imaging model, and instead opt to learn regression function or conditional distribution directly. The difficulties with this class of methods are: (1) the conditional probability of pose given image features is typically multi-modal; and (2) learning high dimensional using limited training data often results in over-fitting.

To deal with multi-modality, on the parametric side, mixture models were introduced, *e.g.*, Mixture of Regressors [3] or Mixture of Experts [4] (see Fig. 1). On the non-parametric side, local models are proposed (*e.g.*, Local Gaussian Process Latent Variable Models (Local GPLVM) [5]). In both the parametric and non-parametric cases, over-fitting and generalization remain a problem due to insufficient training data.

One class of alternatives is variational methods [6]. The variational parameters gives an approximation to the marginal or conditional probabilities. Variational Inferences for mixture models have been been used in Gaussian distribution [7], Factor Analyzer [8], Independent Component Analyzer [9], Dirichlet Process [10], among others.

However, these methods do not explore correlations between local and global structure. Neglecting such valuable information can lead to inconsistent, suboptimal estimations. We present a variational semi-parametric regression model in latent space (see Fig. 1). A multi-modal joint density model can be learned in the form of Variational Bayesian framework. Local structure in latent space is determined by the components of a global Gaussian Mixture Model. Our method can deal with multi-modality in the data, and derive explicit conditional distributions for inference without over-fitting problem.

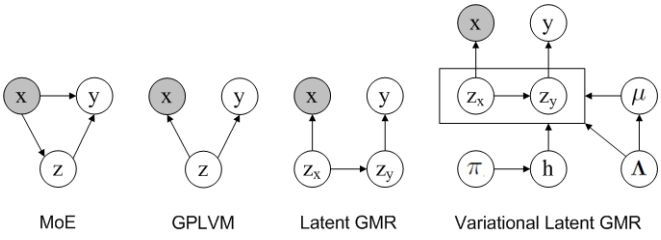

Fig. 1. Graphical models for mixture models. Gray nodes x depict observed variables, and nodes y represent target variables. z, z_x, and z_y are latent variable, and mixing coefficients, means and precision matrix in Gaussian Mixture Model are π, μ and $\mathbf{\Lambda}$ individually.

2. VARIATIONAL LATENT GAUSSIAN MIXTURE REGRESSION

2.1. Variational Bayesian Learning Model

Supposing mixing coefficient, mean and precision matrix in Gaussian Mixture Model are $\pi \in \mathbb{R}^K$, $\mu \in \mathbb{R}^{d\times K}$ and $\mathbf{\Lambda} \in \mathbb{R}^{d\times d\times K}$, respectively, where K is the number of components, and d is the dimension of data. Observed data are denoted by $\mathbf{Z} \in \mathbb{R}^{d\times N}$, and latent variables are represented by $\mathbf{H} \in \mathbb{R}^{d\times K}$, where N is the number of training data. As described in the Section 1, if the density function is achieved with a Gaussian Mixture Model using Expectation Maximization (EM) procedure with K-means initialization, then the optimal number of components in the mixture needs to be determined.

Variational inference involves two probability distributions— posterior distribution p and variational distribution q. If q can adjust free variational parameters to approximate p, and it is restricted to the factorized form

$$q(\mathbf{H}, \pi, \mu, \mathbf{\Lambda}) = q(\mathbf{H})q(\pi, \mu, \mathbf{\Lambda}). \tag{1}$$

Then $q(\mathbf{H})$ can be gained by maximizing the marginal log-likelihood

$$\max \log p(\mathbf{Z}) = \int q(\mathbf{H}) \frac{p(\mathbf{Z}, \mathbf{H}, \pi, \mu, \mathbf{\Lambda})}{q(\mathbf{H})} d_{\mathbf{H}} + KL(q \parallel p), \tag{2}$$

where $KL(q \parallel p)$ is a penalty to reduce the distance between the posterior and the variational distribution.

The optimal solution can be obtained by calculating the derivative with respect to $q(\mathbf{H})$. This can be equivalent to minimizing KL distance between $q(\mathbf{H})$ and $p(\mathbf{Z}, \mathbf{H}, \pi, \mu, \mathbf{\Lambda})$, and the solution is

$$\log q^*(\mathbf{H}) = \mathbb{E}_\pi[\log p(\mathbf{H}|\pi)] + \mathbb{E}_{\mu,\mathbf{\Lambda}}[\log p(\mathbf{Z}|\mathbf{H}, \mu, \mathbf{\Lambda})] + const. \tag{3}$$

The factor $q(\pi, \mu, \mathbf{\Lambda})$ can further factorize as

$$\log q(\pi, \mu, \mathbf{\Lambda}) = q(\pi) \prod_k q(\mu_k, \mathbf{\Lambda}_k), \tag{4}$$

and the optimal solution $q^*(\pi)$,$q^*(\mu)$,$q^*(\mathbf{\Lambda})$ can be gained with the same method as $q^*(\mathbf{H})$

$$\log q^*(\pi) = \log p(\pi) + \mathbb{E}[\log p(\mathbf{H}|\pi)] + const, \tag{5}$$

$$\log q^*(\mu, \mathbf{\Lambda}) = \log p(\mu, \mathbf{\Lambda}) + \mathbb{E}[\log p(\mathbf{Z}|\mathbf{H}, \mu, \mathbf{\Lambda})] + const. \tag{6}$$

Supposing the prior distributions needed are

$$p(\pi) = Dir(\pi|\alpha), \tag{7}$$

$$p(\mu, \mathbf{\Lambda}) = \prod_{k=1}^K \mathcal{N}(\mu_k|\mathbf{m}, (\beta\mathbf{\Lambda})^{-1}) \mathcal{W}(\mathbf{\Lambda}|\mathbf{W}, v), \tag{8}$$

$$p(\mathbf{H}|\pi) = \prod_{n=1}^N \prod_{k=1}^K \pi_k^{H_{nk}}, \tag{9}$$

$$p(\mathbf{Z}|\mathbf{H}, \mu, \mathbf{\Lambda}) = \prod_{n=1}^N \prod_{k=1}^K \mathcal{N}(\mathbf{z}_n|\mu_k, \mathbf{\Lambda}_k^{-1})^{H_{nk}}, \tag{10}$$

where α is a parameter of the Dirichlet Distribution, and $\mathbf{m}, \beta, \mathbf{W}, v$ are parameters of the Gaussian-Wishart Distribution.

The optimization of variational posterior distribution can be gained by EM algorithm. In the E step, the current variational distributions over the parameters are employed to evaluate the moments, then in the M step, these moments are fixed and used to re-compute the variational distribution over the parameters. The algorithm iterates until a final result is obtained.

2.2. Learning the Structure of a Probability Distribution

Given observations, $\mathbf{z}_x \in \mathbb{R}^{d_1}$, and targets, $\mathbf{z}_y \in \mathbb{R}^{d_2}$, where d_1 is dimension of the observation, and d_2 is dimension of the target space, $d = d_1 + d_2$, we learn the correlation between local structure and global structure by using the Latent Gaussian Mixture Regression. Assuming the joint data samples, $(\mathbf{z}_x, \mathbf{z}_y)$, follow the Gaussian mixture distribution with K mixture components,

$$P(\mathbf{z}_x, \mathbf{z}_y) = \sum_{k=1}^K \pi_k P(\mathbf{z}_x, \mathbf{z}_y; \mu_k, \mathbf{C}_k), \tag{11}$$

where $P(\mathbf{z}_x, \mathbf{z}_y; \mu_k, \mathbf{C}_k)$ is the multivariate Gaussian density function. The parameters of model include prior weights, π_k, means, $\mu_k = [\mu_{k,\mathbf{z}_x}\ \mu_{k,\mathbf{z}_y}]^T$, and covariances matrices, $\mathbf{C}_k = [\mathbf{C}_{k,\mathbf{z}_x}\ \mathbf{C}_{k,\mathbf{z}_x\mathbf{z}_y}; \mathbf{C}_{k,\mathbf{z}_y\mathbf{z}_x}\ \mathbf{C}_{k,\mathbf{z}_y}] = {\mathbf{\Lambda_k}}^{-1}$, of each Gaussian component.

The global regression function can be expressed as a mixture of conditional distributions,

$$P(\mathbf{z}_y|\mathbf{z}_x) = \sum_{k=1}^K \omega_k P(\mathbf{z}_y|\mathbf{z}_x; m_k, \sigma_k^2), \tag{12}$$

where the mixing weights ω_k are defined as:

$$\omega_k = \frac{\pi_k P(\mathbf{z}_x; \mu_{k,\mathbf{z}_x}, \mathbf{C}_{k,\mathbf{z}_x})}{\sum_{j=1}^K \pi_j P(\mathbf{z}_x; \mu_{j,\mathbf{z}_x}, \mathbf{C}_{j,\mathbf{z}_x})}. \tag{13}$$

The mean and the variance of the conditional distribution $P(\mathbf{y}|\mathbf{x})$ can be acquired in closed form by

$$m_k = \mu_{k,\mathbf{z}_x} + \mathbf{C}_{k,\mathbf{z}_y\mathbf{z}_x} \mathbf{C}_{k,\mathbf{z}_x}^{-1} (\mathbf{z}_x - \mu_{k,\mathbf{z}_x}), \tag{14}$$

$$\sigma_k^2 = \mathbf{C}_{k,\mathbf{z}_y} - \mathbf{C}_{k,\mathbf{z}_y\mathbf{z}_x} \mathbf{C}_{k,\mathbf{z}_x}^{-1} \mathbf{C}_{k,\mathbf{z}_x\mathbf{z}_y}. \tag{15}$$

Given a new input, a prediction can be obtained by computing expectation over $P(\mathbf{z}_y|\mathbf{z}_x)$

$$E[P(\mathbf{z_y}|\mathbf{z_x})] = \sum_{k=1}^K \omega_k m_k. \tag{16}$$

Because m_k depends on current input $\mathbf{z}_x$, our model is a semi-parametric regression model, and it determines the local structure of the data but according to the components of a global Gaussian mixture model.

2.3. Initialization

Representations for the observation and full body pose configuration are high dimensional and complex by nature. Nonlinear dimensionality reduction techniques like manifold learning identify the latent space from observation and body pose individually.

Besides that, Locality Preserving Projections (LPP) [11], can be simply applied to any new data point to locate it in the reduced representation space by finding the optimal linear approximations to the eigenfunctions of the Laplace Beltrami operator on the manifold.

For example, given a training data set of N poses, $\mathbf{Y} = \{\mathbf{y}^{(1)}, \mathbf{y}^{(2)}, ..., \mathbf{y}^{(N)}\} \in \mathbb{R}^{d_\mathbf{y} \times N}$, we want to find a transformation matrix $\mathbf{A} = [\mathbf{a}_1, ..., \mathbf{a}_{d_\mathbf{z}}]^T$ of basis vectors, $\mathbf{a}_i$, that maps these points to a set of latent points $\mathbf{Z} = \{\mathbf{z}^{(1)}, \mathbf{z}^{(2)}, ..., \mathbf{z}^{(N)}\} \in \mathbb{R}^{d_\mathbf{z} \times N}$ ($d_\mathbf{z} \ll d_\mathbf{y}$), such that $\mathbf{z}^{(i)}$ is a low dimensional manifold embedding representation of a high dimensional space pose $\mathbf{y}^{(i)}$. Following [11], this can be expressed as:

$$\begin{aligned} \min_{\mathbf{A}} \quad & tr(\mathbf{A^T Y L Y^T A}) \\ \text{subject to} \quad & \mathbf{A^T Y D Y^T A} = \mathbf{I}, \end{aligned} \tag{17}$$

Where $\mathbf{D}$ is a diagonal matrix whose entries are column sums of weight matrix $\mathbf{W}$, and $\mathbf{W}$ incurs a heavy penalty if neighboring training points are mapped far apart; $\mathbf{L} = \mathbf{D} - \mathbf{W}$ is Laplacian matrix.

3. EXPERIMENTS

The performance of our method is given in this section.

Data set. (1) The Poser Data Set – synthetic sequences of body postures produced by Poser software [12]. The motion sequences come from six categories: walk, run, dance, fall, prone, and misc (see Fig. 2). A total of five sequences within each category are broken into: three training and two testing sequences, with each sequence containing approximately 500 frames. The size of each synthetic image is 500×490 pixels. We represent body pose in terms of 3D positions of 23 joints, resulting in $d_\mathbf{y} = 69$. All poses are represented in relative terms by subtracting the skeleton root (pelvis) from all other joint centers in every frame.

(2) Carnegie Mellon University Graphics Lab Motion Capture Database – real image/mocap data set publicly available from [13]. We chose sequences of Subject 2 as training data, and use sequences in Subject 1, 8, 15 and 17 as test data, as their foregrounds are easy to be calculated. The size of each image is 240×352 pixels. We represent body pose in terms of 3D joint positions, resulting in $d_\mathbf{y} = 93$. Again, all poses are represented relative to the skeleton root (pelvis).

Image Features. A number of representations for image features have been introduced over the years, *e.g.*, Scale invariant feature transform (SIFT) [14] or histogram of shape context [15], to name a few. Similar to prior work, we rely on silhouette features and encode them using 70D histogram of shape context.

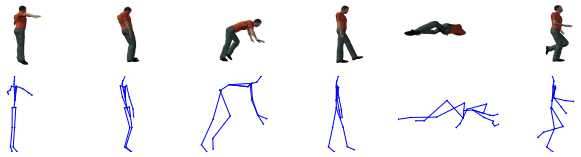

Fig. 2. Synthesized data generated by Poser 7 software.

Error (cm)		KR	LR	MoE	LGMR	VLGMR
dance	S1	10.85	5.83	5.72	5.60	**5.41**
	S2	10.37	5.23	5.04	4.91	**4.68**
prone	S1	11.36	6.55	6.40	5.88	**5.73**
	S2	12.46	6.36	6.28	6.19	**6.00**
falls	S1	15.32	10.40	10.25	10.05	**9.77**
	S2	16.31	11.50	11.26	10.92	**10.68**
walk	S1	11.65	6.36	6.06	5.93	**5.80**
	S2	9.15	3.55	3.34	3.15	**3.12**
miscs	S1	8.32	3.59	3.42	3.28	**3.24**
	S2	19.27	12.19	12.10	11.80	**11.64**
run	S1	8.94	4.70	4.64	4.31	**4.27**
	S2	11.65	5.96	5.79	5.62	**5.44**
Average		12.13	6.85	6.69	6.47	**6.31**

Table 1. Evaluation of different algorithms on the Poser data set (for details see text).

Comparison. We compare our Latent Gaussian Mixture Regression (**LGMR**) model, Variational Latent Gaussian Mixture Regression (**VLGMR**) model with kernel regression (**KR**), linear regression (**LR**), and mixture of experts (**MoE**). The results are shown in Table 1 and Table 2. In our method, the Locality Preserving Projections (LPP) is trained to keep 95% of the original energy.

We can see that since our features and data are sparse, kernel regression approaches tend to work poorly. Mixture models, *e.g.*mixture of experts and Latent Gaussian Mixture Regression, learn through unsupervised methods, and suffer from over-fitting, as a result, their performance degrades in test sequences. By minimizing KL distance between the posterior and the variational distribution, Variational Latent Gaussian Mixture Regression solves the problem, and obtains the best performance at the cost of extra training time. Though Variational Latent Gaussian Mixture Regression is as complex as the Mixture of Experts algorithm, its Mean Square Error is much less than the Mixture of Experts, and it does not need to employ cross-validation to choose parameters, so it actually spends less time in training than Latent Gaussian Mixture Regression.

We also show subjective results of different approaches in Carnegie Mellon University Graphics Lab Motion Capture Database in Fig 3. In real video sequence, it is difficult to predict a 3D human gesture owing to the cluttered background, shadow, and lighting. As we use the latent variable method, our result is less sensitive to noise in the real video, which makes it more accurate than other methods.

Error (cm)		KR	LR	MoE	LGMR	VLGMR
Subject 1	01-01	18.27	16.18	15.79	14.49	**14.33**
	01-02	22.27	23.01	21.77	18.49	**18.32**
	01-03	32.88	34.74	34.12	32.76	**32.45**
Subject 8	08-01	19.00	13.93	13.14	11.78	**11.61**
	08-02	17.59	19.95	19.17	18.66	**18.55**
	08-03	16.69	22.22	18.33	15.45	**15.29**
Subject 15	15-01	20.24	13.64	13.25	13.14	**13.01**
	15-02	15.90	14.26	13.59	13.42	**13.25**
	15-03	28.82	23.18	23.01	22.95	**22.78**
Subject 17	17-01	29.49	25.49	23.68	23.23	**23.08**
	17-02	21.43	13.93	13.42	12.01	**11.86**
	17-03	21.43	15.84	14.77	14.55	**14.40**
Average		21.99	19.68	18.61	17.54	**17.41**
Train time		0	0.06	72.18	19.38	78.52
Test time		10.28	0.02	0.17	1.14	1.15

Table 2. Evaluation of different algorithms in Carnegie Mellon University motion capture database; the learning and inference time is also given in (*seconds*).

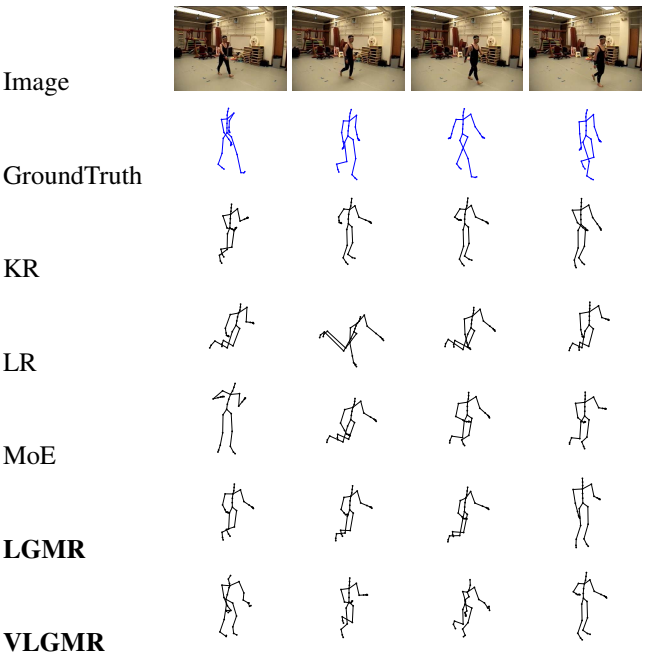

Fig. 3. Evaluation on frames 58, 68, 78, and 88 of sequence 04 in subject 08 from the Carnegie Mellon University motion capture database.

4. CONCLUSION

We have identified 3D poses in video data using a novel method of semi-parametric regression models based on variational inference. Our algorithm creates a stick figure that mirrors the person's pose in the video to a fair degree of accuracy. We arrive at a few close predictions to the ground truth pose in the video by using a Latent Gaussian Mixture Regression Model, and then enhancing the model's performance using a Variational Bayesian framework. Advantages to our method include that it can handle multi-modality in the data and derive explicit conditional distributions for inference that result in greater accuracy. We show that our performance compares well to related parametric and non-parametric models in the original high-dimensional space.

5. REFERENCES

[1] C. Sminchisescu and B. Triggs, "Covariance scaled sampling for monocular 3d body tracking," in *CVPR*, 2001.

[2] T. Jaeggli, E. Koller-Meier, and L. Van Gool, "Learning Generative Models for Multi-Activity Body Pose Estimation," *International Journal of Computer Vision*, vol. 83, no. 2, pp. 121–134, 2009.

[3] A. Agarwal and B. Triggs, "Monocular human motion capture with a mixture of regressors," in *CVPR*, 2005.

[4] C. Sminchisescu, A. Kanaujia, Z. Li, and D. Metaxas, "Discriminative density propagation for 3d human motion estimation," in *CVPR*, 2005.

[5] R. Urtasun and T. Darrell, "Sparse probabilistic regression for activity-independent human pose inference," in *CVPR*, 2008.

[6] C.M. Bishop, *Pattern recognition and machine learning*, Springer-Verlag New York, Inc. Secaucus, NJ, USA, 2006.

[7] H. Attias, "A variational Bayesian framework for graphical models," in *NIPS*, 2000.

[8] Z. Ghahramani and M.J. Beal, "Variational inference for Bayesian mixtures of factor analysers," in *NIPS*, 2000.

[9] RA Choudrey and SJ Roberts, "Variational mixture of Bayesian independent component analyzers," *Neural Computation*, vol. 15, no. 1, pp. 213–252, 2003.

[10] D.M. Blei and M.I. Jordan, "Variational inference for Dirichlet process mixtures," *Bayesian Analysis*, vol. 1, no. 1, pp. 121–144, 2006.

[11] X. He and P. Niyogi, "Locality preserving projections," in *NIPS*, 2003.

[12] E frontier. Curious Labs Poser. Computer Software., ," .

[13] CMU Motion Capture Database. http://mocap.cs.cmu.edu/, ," .

[14] L. Bo and C. Sminchisescu, "Structured output-associative regression," in *CVPR*, 2009.

[15] L. Sigal, A. Balan, and M.J. Black, "Combined discriminative and generative articulated pose and non-rigid shape estimation," in *NIPS*, 2007.