EFFICIENT INITIALIZATION OF MIXTURES OF EXPERTS FOR HUMAN POSE ESTIMATION

Huazhong Ning, Yuxiao Hu, Thomas Huang

ECE Department, U. of Illinois at Urbana-Champaign, Urbana, IL 61801. {hning2,hu3, huang}@ifp.uiuc.edu

ABSTRACT

This paper addresses the problem of recovering 3D human pose from a single monocular image. In the literature, Bayesian Mixtures of Experts (BME) was successfully used to represent the multimodal image-to-pose distributions. However, the EM algorithm that learns the BME model may converge to a suboptimal local maximum. And the quality of the final solution depends largely on the initial values. In this paper, we propose an efficient initialization method for BME learning. We first partition the training set so that each subset can be well modeled by a single expert and the total regression error is minimized. Then each expert and gate of BME model is initialized on a partition subset. Our initialization method is tested on both a quasi-synthetic dataset and a real dataset (HumanEva). Results show that it greatly reduces the computational cost in training while improves testing accuracy.

Index Terms**—** Initialization, Bayesian Mixtures of Experts, Human Pose Estimation

1. INTRODUCTION

Robust recovery of 3D human pose in monocular images or videos is an actively growing field. Effective solutions would lead to breakthroughs in a wide range of applications spanning visual surveillance, video indexing and retrieval, HCI, and so on. Unfortunately, this problem is extremely challenging due to both the internal complexity of the articulated human body and the external variations of the scene [1]. Among the approaches for human pose estimation, the *discriminative methods*, that learns direct image-to-pose mappings by training on labeled human poses, gained interests in the literature due to its fast test speed.

These discriminative methods differ in the organization of training set and in the runtime hypothesis selection [2], varying from Bayesian mixtures of experts (BME) [2, 3, 4], linear/kernel regression [5], manifold embedding [6], nearest-neighbor retrieval from typical examples [7], mixture of probabilistic PCA [8], to mixture of multi-layer perceptrons [9]. We choose the BME model, because the multi-modalities in the image-to-pose distributions can be well modeled by the mixtures of experts. It has produced superior results on human pose estimation in the literature [2, 3, 4].

For the BME model [10, 11], EM algorithm is used to estimate the parameters of both the gate network and expert network. However, there is no grurantee that the EM iteration converges to a global maximum likelihood estimator, although it does not decrease the observed data likelihood function. This means that, for multimodal distributions, an EM algorithm may converge to a local maximum or saddle point and the quality of the final solution depends largely on initial values. This happens frequently for problems with large parameter space like human pose estimation. Hence, choosing initial values is an crucial step in learning BME models, while it did not gain much attention in previous work [2, 3, 4].

In this paper, we propose an efficient initialization method by exploiting a special property of the BME model: makes an assumption that the data can be faithfully modeled by a combination of expert linear/kernel regressors. By this assumption, the data roughly consist of piecewise subsets and each subset can be well modeled by a single regressor. We first design an iterative algorithm to find such a partition. Then each expert and gate is initialized on a partition subset. Since the partition is a good (discrete) approximation of the gating network of the BME model, this partition initialization makes the EM estimation converge much faster than starting from a random initialization. On the other hand, our partition algorithm itself converges very fast, so the extra cost of initialization is negligible compared to the reduced cost of EM estimation. In our experiments on large scale data, our partition initialization also improves the testing accuracy.

2. BAYESIAN MIXTURES OF EXPERTS

The image-to-pose distribution is highly non-linear and multimodal. This leads us to use the BME model [10, 11] to model it, since BME was specifically designed to model multimodality. The BME model has produced superior results on human pose estimation [2, 3, 4]. Here we give a brief introduction. Suppose $\mathbf{x}$ is vector representation of human image and $\mathbf{y}$ is human pose, the model with M experts is:

$$p(\mathbf{y}|\mathbf{x},\Theta) = \sum_{i=1}^{M} g(\mathbf{x},\nu_i)p(\mathbf{y}|\mathbf{x},T_i,\Lambda_i) \quad (1)$$

where

$$g(\mathbf{x}, \nu_i) = \frac{e^{\nu_i^T \mathbf{x}}}{\sum_j e^{\nu_j^T \mathbf{x}}} \quad (2)$$

$$p(\mathbf{y}|\mathbf{x}, T_i, \Lambda_i) \sim \mathcal{N}(f_i(T_i, \mathbf{x}), \Lambda_i) \quad (3)$$

Here $\Theta = \{\nu_i, T_i, \Lambda_i | i = 1, 2, \cdots, M\}$ consists of the parameters of the BME model. $p(\mathbf{y}|\mathbf{x}, T_i, \Lambda_i)$ is a Gaussian distribution with mean $f_i(T_i, \mathbf{x})$ and covariance matrix Λ_i, and it is an *expert* that transforms the input into output prediction. f(.) is a linear or non-linear function, *e.g.*, a linear case $f_i(T_i, \mathbf{x}) = T_i\mathbf{x}$. Then the predictions from different experts are combined in a probabilistic mixture model. Note that the mixing proportions of the experts, $g(\mathbf{x}, \nu_i)$, are *input dependent* and normalized to 1 by the softmax construction. They reflect the distributions of the outputs in the training set. They work like gates that can competitively switch-on multiple experts for some input domains, allowing multi-modal conditionals. They can also pick a single expert for unambiguous inputs by switching-off other experts.

The parameter Θ can be estimated by maximizing $L = \sum_k \ln p(\mathbf{y}^{(t)}|\mathbf{x}^{(t)}, \Theta)$ where $(\mathbf{x}^{(t)}, \mathbf{y}^{(t)})$ are image-pose pairs. This can be achieved through EM algorithm.

3. INITIALIZATION OF BME LEARNING

As mentioned in Section 2, the parameters of the BME model are estimated through EM algorithm. However, EM algorithm has an intrinsic limitation that there is no guarantee of convergence to global optimum. For multimodal distributions like the image-to-pose data, it often converges to a local maximum or saddle point. And the quality of the final solution depends heavily on starting values. This means that choosing initial values is a crucial step of BME learning. But it did not gain much attention in previous work [2, 3, 4].

3.1. The Initialization Algorithm

A common approach for escaping a local maximum is to run the EM algorithm several times with random initialization and return the best result. However, this is extremely time consuming for problems with large parameter space. In this paper, we propose an efficient initialization method for BME learning by exploiting a property of BME model: it assumes that each data point is generated from a single or a combination of expert(s). We simplify this assumption by assuming that each data point is generated by a single expert. Then if the data subset generated by a certain expert is known, this expert and its associated gate can be efficiently and roughly estimated from this subset. This estimation might not be accurate enough but it is expected to be a good initialization. This motivates such an initialization method: first partition the training dataset into non-overlapping subsets such that each subset can be well modeled by a single expert; then estimate each expert and its gate from a partition subset, and use this estimation as starting values of EM iteration.

Denote the training dataset as $\mathcal{V} = \{(\mathbf{x}^{(t)}, \mathbf{y}^{(t)}) | t = 1, 2, \cdots, N\}$ where $(\mathbf{x}^{(t)}, \mathbf{y}^{(t)})$ is an image-pose pair. Subsets $\mathcal{V}_1, \mathcal{V}_2, \cdots, \mathcal{V}_M$ is a partition of $\mathcal{V}$ if and only if $\mathcal{V} = \bigcup_i \mathcal{V}_i$ and $\mathcal{V}_i \bigcap \mathcal{V}_j = \varnothing$ for $i \neq j$. The objective of partitioning the training dataset is to minimize the total intra-class regression error, *i.e.*, to minimize the error function

$$\sum_{i=1}^{M} \min_{f_i, T_i} \sum_{(\mathbf{x}, \mathbf{y}) \in \mathcal{V}_i} \|f_i(T_i, \mathbf{x}) - \mathbf{y}\|^2 \quad (4)$$

where $\{\mathcal{V}_i\}_{i=1}^M$ is a partition and $f_i(T_i, .)$ is a regressor on $\mathcal{V}_i$.

We propose an iterative algorithm to find the partition that minimizes the error function in Eqn. 4. It is summarized in Algorithm 1. In Step 2, estimation of regressors depends on the regression functions $f_i(T_i, .)$. For the linear case $f_i(T_i, \mathbf{x}) = T_i\mathbf{x}$ (this is also what we used), $T_i = (Y_i X_i^T)(X_i X_i^T)^{-1}$, where (X_i, Y_i) stacks all pairs $(\mathbf{x}, \mathbf{y}) \in \mathcal{V}_i$ with each column corresponding to a sample.

Algorithm 1 Iterative Algorithm for Searching Partition

```
1: Start from a random partition $\mathcal{V}_1, \mathcal{V}_2, \cdots, \mathcal{V}_M$.
2: Estimate the expert regressor $f_i(T_i, .)$ on subset $\{\mathcal{V}_i\}_{i=0}^M$.
3: For each image-to-pose pair $(\mathbf{x}, \mathbf{y}) \in \mathcal{V}$, compute its regression residual $\|f_i(T_i, \mathbf{x}) - \mathbf{y}\|^2$ on each expert regressor. Let $k = \arg\min_i \|f_i(T_i, \mathbf{x}) - \mathbf{y}\|^2$, $(\mathbf{x}, \mathbf{y})$ is reassigned to subset $\mathcal{V}_k$
4: Repeat Step 2 until convergence is achieved
```

After the partition is found by Algorithm 1, the estimated $f_i(T_i, .)$'s are used as starting values of the expert network. The starting values of gating network are chosen based on the fact that the partition itself can be viewed as a discrete approximation of the gates. Denote $\Phi = [\nu_1, \cdots, \nu_M]$ containing all the gate parameters and $X = [\mathbf{x}^{(1)}, \mathbf{x}^{(2)}, \cdots, \mathbf{x}^{(N)}]$ all the training images. For each $\mathbf{x}^{(t)}$, if it comes from the subset $\mathcal{V}_i$, we set $g(\mathbf{x}^{(t)}, \nu_i) \approx 1$ and $g(\mathbf{x}^{(t)}, \nu_j) \approx 0, j \neq i$, or equivalently $\nu_i^T \mathbf{x}^{(t)} = \tau$ and $\nu_j^T \mathbf{x}^{(t)} = \epsilon, j \neq i$ with $\tau \gg \epsilon$. Let $\Phi \mathbf{x}^{(t)} = \mathbf{u}^{(t)}$ and $U = [\mathbf{u}^{(1)}, \mathbf{u}^{(2)}, \cdots, \mathbf{u}^{(N)}]$, *i.e.*, $\Phi X = U$. We use $\Phi = (UX^T)(XX^T)^{-1}$ as the starting values of gating network.

Fig. 1 (a) shows a toy example that illustrates the iterations of finding the partition. Fig. 1 (b) plots the gates of 3 experts estimated by EM algorithm with starting values initiated by our method. The EM algorithm converges in 17 iterations, while it needs more than 100 iterations with a random initialization. Enlarge for better visualization.

3.2. Discussions on the Algorithm

Firstly, convergence of Algorithm 1 is guaranteed. In Step 2, for a given partition, the error function Eqn. 4 is minimized with respect to $\{f_i(T_i, .)\}_{i=1}^M$ yielding regressors on the currently assigned subsets. In Step 3, given a current set of regressors, Eqn. 4 is minimized by assigning each training

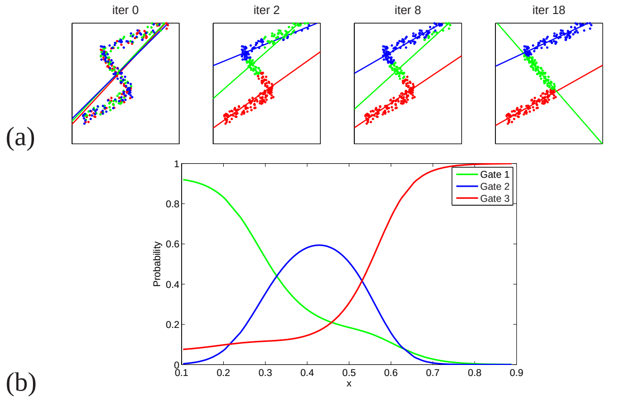

Fig. 1. An illustrative dataset consists of 300 samples generated from 3 linear models: $y = 0.5x + 0.1 + \varepsilon$, $y = -1.1x+1+\varepsilon$, and $y = 0.4x+0.6+\varepsilon$, with $\varepsilon \sim \mathcal{N}(0, 0.05^2)$. (a) Iterations for searching the partition. In each iteration, the data subsets and the estimated regressors (lines) are marked in different colors. (b) Gates of 3 experts estimated by EM algorithm that is initialized by our method.

sample to the subset on which the regression residual is the smallest. In other words, each of Steps 2 and 3 decreases the value of the error function in Eqn. 4. Since there are at most M^N possible partitions, the process will always terminate.

Secondly, like EM algorithm, Algorithm 1 has the same limitation that the result may represent a suboptimal local minimum. On the face of it, our approach does not solve the initialization problem, but transfers it from BME learning to Algorithm 1. But here the key point is that we transfer the initialization problem from a difficult task to a simpler one. Actually, Algorithm 1 converges very fast, and its iterations require much less computational cost than that in BME learning. So, it is feasible to repeat Algorithm 1 with many different random initial partitions, and choose the best partition to initialize the EM algorithm, while it is too time-consuming to repeat the BME learning for so many times. In addition, the number of required iterations in BME learning is largely reduced if it starts from our initialization.

Thirdly, our approach also improves the testing accuracy in our experiments. One possible reason is that Algorithm 1 solves a simpler problem and hence converges to global minimum with higher probability.

Finally, note that our initialization algorithm can be applied to other problems where BME is used. And the objective function Eqn. 4 tries to discover local structures existing in the dataset. This is a common problem in data analysis. Algorithm 1 gives an iterative solution.

4. IMAGE REPRESENTATION

We choose bag of local descriptors [12] to represent the human images. Our local descriptor, called *Appearance and Position Context (APC) descriptor*, is specifically designed to not only capture the co-occurrence and context information within the local informative structures but also encode their relative spatial positions. It is extracted in the following steps. Centered at each point that has large gradient, the local region is partitioned into log-polar sectors [13]. Suppose from inner to outer, the sectors are numbered $1, 2, ..., B$, and θ_i, m_i is the orientation and magnitude of the dominant gradient in sector i. Then the local descriptor is represented as $(x, y, \theta_1, r_1, ..., \theta_B, r_B)$ where x, y are the center coordinate and, $r_i = m_i/m_1$ is the normalized magnitude that basically removes the contrast of the image. The size of the local region is chosen to cover the average length of human limbs.

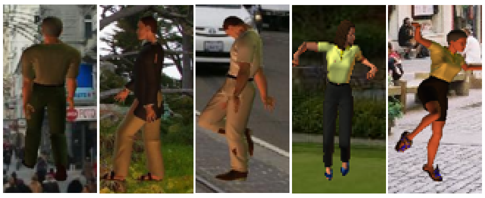

Fig. 2. Some sample synthetic human images in our quasi-synthetic dataset.

5. EXPERIMENTS

5.1. On Quasi-synthetic Dataset

We constructed a quasi-synthetic human database with large variations, by animating 3D human avatars using real motion data and placing the synthetic images on real backgrounds. Fig. 2 gives some sample synthetic images. The 3D human pose has 52 degrees of freedom (DOF), 1 for global orientation and 51 for 17 joints (each upper limb has 4 joints, lower limb has 3, and chest, neck, and head has one, respectively). Our dataset consists of about 131,468 labeled samples, much larger and more complex than the existing quasi-synthetic datasets, like 8,262 samples in [2], 2,500 in [5], 1,200 in [3], and 9,741 in [6].

The experiment setup is as follows. We use 8 experts for the BME model. 60% sequences of the dataset are randomly selected for training and 40% are left for testing. We report mean (over all 55 angles or an individual angle) RMS absolute difference errors between the true and estimated joint angle (vectors), in degrees as in [5]: $D(\mathbf{y}, \mathbf{y}') = \frac{1}{m}\sum_{i=1}^{m} |(y_i - y_i')mod \pm 180^o|$. We compare the performances on two settings: (1) *rand*: BME learning with random initialization is repeated for 10 times and the one with largest likelihood is chosen; (2) *ours*: Algorithm 1 is repeated for 10 times and choose the partition with smallest regressor error to initialize BME learning that does not repeat. Table 1 lists the training time and the average RMS errors over all angles for the two settings. Our approach reduces the training time to $1/18$ of traditional BME learning while improves the testing accuracy to some extent.

method	initialization		BME		time(h)	RMS
	rep#	ave iter#	rep#	ave iter#		
rand	-	-	10	93	7.50	6.05^{o}
ours	10	21.3	1	22	0.42	5.89^{o}

Table 1. Total training time and the average RMS error. *rep#*: repetition number; *ave iter#*: average iteration number.

algorithm	mean	std	time(s)
Zhou [14]	0.303	0.075	40.55
Bissacco [15]	0.274	0.116	3.28
Ours	**0.241**	**0.158**	**0.21**

Table 2. Mean and standard deviation (*std*) of the pose estimation errors (L_2 error norm) and inference time.

5.2. On HumanEva Dataset

To compare with other state of the art methods [14, 15], we also test our approach on a real human motion dataset–HumanEva–made publicly available by the Brown Group [16]. The experiment uses only the walking sequences having a total of 2950 frames (first trial of subject S1, S2, and S3), as [15] did. All images are taken from a single camera (C1) because our approach recovers human pose from a single view. We use walking sequences in the original training subset for training and those in the original validation subset for testing. The original testing subset is not used because motion data were not provided for it.

Here the joint angle trajectories are normalized as [14, 15] did, so that $\mathbf{y}$ is a zero-mean unit variance process. In this way, each angle in $\mathbf{y}$ contributes equally to the error function. We use the relative L_2 error norm [15]: $\|\hat{\mathbf{y}} - \mathbf{y}\|/\|\mathbf{y}\|$ where $\mathbf{y}$ is the ground truth and $\hat{\mathbf{y}}$ is the estimation. Table 2 shows the mean and standard deviation of the relative L_2 pose error norms on the walking sequences. Our approach outperforms the other state of the art algorithms [14, 15] in estimation accuracy. And as to the computational speed of inference, our approach is 15 times faster than [14, 15]. Fig. 3 shows some sample frames together with the estimated pose represented as the outline of a cylinder-based human model superimposed onto the original images. We visualize the estimated pose on cameras: C1, C2, and C3, and the ground truth on camera C1 only. Note that our estimations are obtained only from images captured by camera C1.

6. CONCLUSION

This paper uses Bayesian mixtures of experts (BME) to represent the multimodal image-to-pose distributions. However, the quality of the final BME model depends largely on the initial values of the learning process. We propose an efficient initialization method for BME learning. We partition the training set so that each subset can be well modeled by a single expert and the total regression error is minimized. Then each expert and gate of BME model is initialized on a partition subset. We test our approach on both a quasi-synthetic dataset and a real dataset to verify its effectiveness.

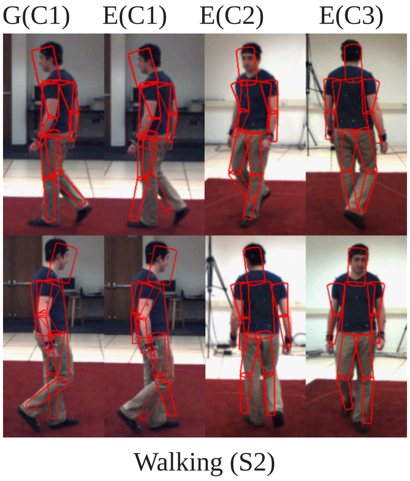

Fig. 3. Sample estimation results. Each column shows ground truth (G) or estimation (E) projected to cameras: C1, C2, C3.

7. REFERENCES

[1] H. Ning, T. Tan, L. Wang, , and W. Hu, “People tracking based on motion model and motion constraints with automatic initialization,” *Pattern Recognition*, 2004.

[2] C. Sminchisescu, A. Kanaujia, and D. Metaxas, “Bm^{3}e: Discriminative density propagation for visual tracking,” *PAMI*, 2007.

[3] F. Guo and G. Qian, “Learning and inference of 3d human poses from gaussian mixture modeled silhouettes,” *ICPR*, 2006.

[4] L. Sigal and M. J. Black, “Predicting 3d people from 2d pictures,” *AMDO*, 2006.

[5] A. Agarwal and B. Triggs, “Recovering 3d human pose from monocular images,” *PAMI*, 2006.

[6] A. Kanaujia, C. Sminchisescu, and D. Metaxas, “Semi-supervised hierarchical models for 3d human pose reconstruction,” *CVPR*, 2007.

[7] A. Fathi and G. Mori, “Human pose estimation using motion exemplars,” *ICCV*, 2007.

[8] K. Grauman, G. Shakhnarovich, and T. Darell, “Inferring 3d structure with a statistical image-based shape model,” *ICCV*, 2003.

[9] R. Rosales and S. Sclaroff, “Learning body pose via specialized maps,” *NIPS*, 2002.

[10] C. Bishop and M. Svensen, “Bayesian mixtures of experts,” *Uncertainty in Artificial Intelligence*, 2003.

[11] M. Jordan and R. Jacobs, “Hierarchical mixtures of experts and the em algorithm,” *Neural Computation*, vol. 6, 1994.

[12] L. Fei-Fei and P. Perona, “A bayesian heirarchical model for learning natural scene categories,” *CVPR*, 2005.

[13] G. Mori and J. Malik, “Estimating human body configurations using shape context matching,” *ECCV*, 2002.

[14] S.K. Zhou, B. Georgescu, X.S. Zhou, and D. Comaniciu, “Image based regression using boosting method,” in *ICCV*, 2005.

[15] A. Bissacco, M.-H. Yang, and S. Soatto, “Fast human pose estimation using appearance and motion via multi-dimensional boosting regression,” *CVPR*, 2007.

[16] L. Sigal and M. J. Black, “Humaneva: Synchronized video and motion capture dataset for evaluation of articulated human motion,” *Technical Report CS-06-08, Brown University*, 2006.