Silhouette-based human pose estimation using reversible jump Markov chain Monte Carlo

S.-S. Huang, L.-C. Fu and P.-Y. Hsiao

A novel approach for recovering the human body configuration based on the silhouette is presented. By considering pose inference as traversing the difference subspaces and using a data-driven mechanism, reversible jump Markov chain Monte Carlo (RJMCMC) can explore such solution space very efficiently. Experimental results are provided to demonstrate the efficiency and effectiveness of the proposed approach.

Introduction: The challenges in the problem of human pose estimation are mainly due to the appearance of variations, which include clothing, self-occlusion and articulated-deformation. According to the methodologies used, the approaches for recovering the human pose can be divided into three classes: the component-based approach [1], the template-based approach [2], and the parameterisation-based approach [3]. The limitation of the component-based approach is that detecting body parts may be difficult, especially in a cluttered image. As for the template-based approach, its performance is heavily dependent on the templates used for learning. The parameterisation-based approach is the more effective and general one, but it is time-consuming. This Letter proposes an efficient and effective parameterised-based approach.

Estimation framework: The problem of human pose estimation is formulated as computing the maximum *a posteriori* (MAP):

$$\omega^* = \arg\max_{\omega\in\Omega} \Pr(\omega|I)$$

where Ω is the solution space of the pose to be estimated and I is the currently observed image. Following the Bayes rule, the posterior probability can be decomposed into a likelihood term $\Pr(I|\omega)$ and a prior term $\Pr(\omega)$, i.e. $\omega^* = \arg\max_{\omega\in\Omega} \Pr(I|\omega)\Pr(\omega)$.

Here, a 2D pictorial human model, which consists of ten body parts as shown in Fig. 1*a* is used to present a human pose wherein each body part i is parameterised by a configuration $\Theta_i = (x_i, y_i, \theta_i, s_i, l_i)$. The definitions of these five parameters are given in Fig. 1*b*. For a human body, the adjacent body parts are connected by a flexible joint and have a geometry constraint called connectivity. Let p_{ij}^J be a joint point specified by the configuration Θ_i of the part i which is connected to the part j. Ideally, p_{ij}^J and p_{ji}^J should refer to the same point and thus the Euclidean distance between p_{ij}^J and p_{ji}^J modelled by a Gaussian distribution can be expressed as:

$$\psi_{connectivity}(\Theta_i, \Theta_j) = \exp\left\{-\frac{\|p_{ij}^J - p_{ji}^J\|^2}{2\sigma_c^2}\right\}$$

where $\|\cdot\|$ is the Euclidean distance and σ_c is a standard deviation. Besides connectivity, the widths between adjacent parts should have a consistent scaling factor. The scale consistency between the adjacent parts i and j is defined as:

$$\psi_{scale}(\Theta_i, \Theta_j) = \exp\left\{-\frac{\|f_{contrast}(S_i, S_j) - f_{contrast}(\bar{S}_i, \bar{S}_j)\|^2}{2\sigma_s^2}\right\}$$

where $f_{contrast}(a, b) = (a-b)/(a+b)$ is the contrast function, σ_s is the standard deviation, and $\bar{s}_i$ and $\bar{s}_j$ are the physical widths of parts i and j, respectively. As a result, the prior distribution of a pair of adjacent parts i and j is defined as $\Pr(\Theta_i, \Theta_j) = 1/Z_1 \psi_{connectivity}(\Theta_i, \Theta_j)\psi_{scale}(\Theta_i, \Theta_j)$, where Z_1 is a normalisation constant.

Here, we use the human silhouette to define the likelihood distribution. Let I_s be the region of the human silhouette obtained using the background subtraction technique, and I_ω be the region formed by the union of regions specified by the part configurations of the solution ω. A measure computing the region distance is introduced to define a better likelihood distribution $\Pr(I|\omega)$ of seeing I given a solution ω as:

$$\Pr(I|\omega) = \frac{1}{Z^2}\exp\{-(d_h(I_\omega, I_s) + d_h(\bar{I}_\omega, \bar{I}_s))\}$$

where $d_h(A, B) = \sum_{a\in A} \min_{b\in B}(\|a-b\|)$ measures the distance between two regions A and B, and Z_2 is a normalisation constant.

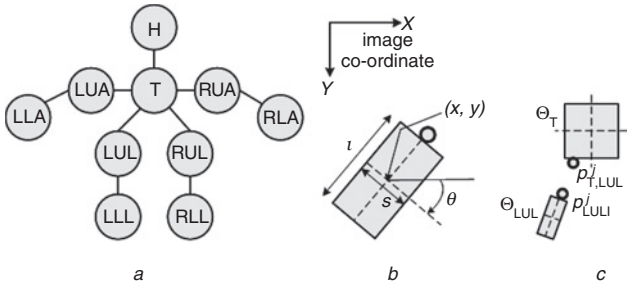

Fig. 1 *Structure of human model, configuration of body parts, and connectivity property*

a Structure of human model consists of ten parts and nine joints
b Configuration of body parts in a schematic form
c Connectivity property between parts H and LUL

Human pose inference: Inspired by the component-based approach [1], the manner of recovering the pose configuration may be considered as an assembling process. Fig. 2 shows the structure of the solution space. By constructing a solution space in this way, two merits can be obtained. First, instead of finding the solutions directly in Ω_{10} (the solution space consisting of ten body parts) as the work reported in [3] does, the solutions are explored by traversing the subspaces of different dimensionality. This drastically improves the efficiency of solution exploration because the bad solutions in the early subspace with low dimensionality will have a low probability of being visited and jump to the higher subspace. Secondly, the proposed process is a probabilistic framework but not a deterministic one. This gives the system a chance to recover the human configuration even though some body parts are difficult to detect.

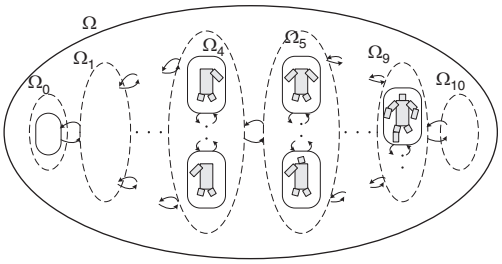

Fig. 2 *Structure of solution space Ω for problem of human pose estimation*

The solution space is high-dimensional and the posterior distribution over such a space is multimodal. It is impractical to do the exact inference. An approximation technique called RJMCMC [4] is utilised to approximate the posterior distribution by drawing samples. The three types of MCMC dynamics we utilise to explore the solution space are according to the work listed in [4] and are described as:

Drift dynamic: randomly selects one body part in the solution and then diffuses its values by following a Gaussian distribution.
Addition dynamic: adds a new body part to the assembly and jumps to the subspace with a higher dimension.
Removal dynamic: takes away one body part from the assembly and jumps to the subspace with a lower dimension.

To reduce burn-in period and mix rapidly, the key issue is how to compute an effective proposal probability for proposing the parameters in addition dynamic. Here, we achieve this by extracting all symmetry candidates using symmetry transform [5].

Results: In this Section, several images with a resolution 512×768 are considered to validate our proposed method. Our system is implemented on a personal computer with a 1.4 GHz Pentium IV processor. For each image, 100 samples are drawn from the defined posterior distribution. Fig. 3 shows the curve of posterior probability

over these 100 iterations and shows a rapid convergence rate. The average time for processing each frame is about 5.3 s and is much faster than the approaches in [3] (about 1 s). The estimated results shown in Fig. 4 are the samples with the maximum posterior. Figs. 4*a–d* are four images of a person exhibiting different postures. Figs. 4*e* and *f* demonstrate the effectiveness of our approach for the case when skin colour of body parts is invisible. The cases when some parts are occluded are shown in Figs. 4*g* and *h*. For evaluation, we compare the estimated centre position and orientation of the parts with the manually annotated ones. Table 1 shows the RMS (root mean square) errors of head, torso and limbs, respectively.

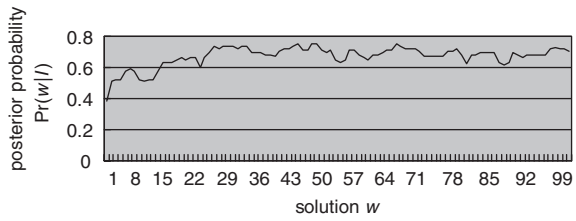

Fig. 3 *Curve of posterior probability over 100 drawn samples*

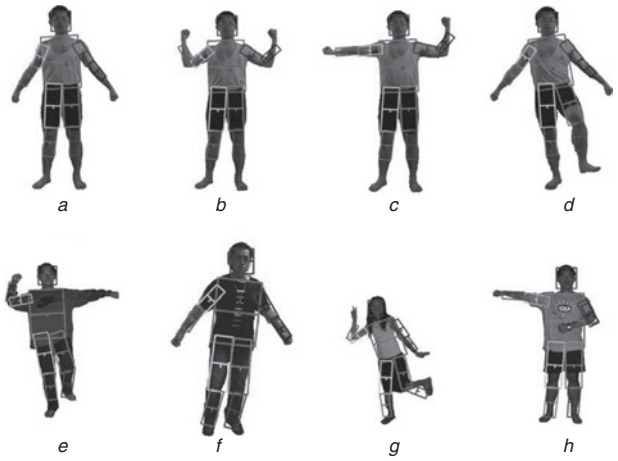

Fig. 4 *Some estimation results of proposed approach*

Table 1: Average RMS error: unit of Δd is pixel and of $\Delta\theta$ is degree

Head		Torso		Limb	
Δd	$\Delta\theta$	Δd	$\Delta\theta$	Δd	$\Delta\theta$
10.8	2.3	17.02	2.3	24.86	9.5

Conclusion: We have proposed an effective and efficient approach to recover the human pose even when the skin colour and some body parts are invisible. The efficiency is achieved by incorporating the domain knowledge and formulating the estimation problem as solution exploration over 11 subspaces with different dimensionality.

Acknowledgment: This research is sponsored by the National Science Council under the project NSC93-2752-E-002-007-PAE.

16 February 2006
Electronics Letters online no: 20060044
doi: 10.1049/el:20060044

S.-S. Huang and L.-C. Fu (*Department of Computer Science and Information Engineering, National University, No. 1, Sec. 4, Roosevelt Road, Taipei, Taiwan, Republic of China*)

E-mail: powwhuang@gmail.com

P.-Y. Hsiao (*Department of Electronic Engineering, Chang Gung University, 259, Wen-Hwa 1st Road, Kwei-Shan, Tao-Yuan, Taiwan, Republic of China*)

References

1 Ioffe, S., and Forsyth, D.A.: 'Probabilistic methods for finding people', *Int. J. Comput. Vis.*, 2001, **43**, (1), pp. 45–68
2 Shakhnarowich, G., Viola, P., and Darrell, T.: 'Fast pose estimation with parameter-sensitive hashing'. Int. Conf. on Computer Vision, 2003, Vol. 2, pp. 750–757
3 Felzenszwalb, P.F.: 'Pictorial structures for object recognition', *Int. J. Comput. Vis.*, 2005, **61**, (1), pp. 55–79
4 Green, P.J.: 'Reversible jump Markov chain Monte Carlo computation and Bayesian Model Determination', *Biometrika*, 1995, **82**, (4), pp. 711–732
5 Reisfeld, D., Wolfson, H., and Yeshurun, Y.: 'Context free attentional operations: the generalized symmetry transform', *Int. J. Comput. Vis.*, 1995, **12**, (2), pp. 119–130