Multisensor Human Tracker based on the Markov Chain Monte Carlo Method

Takuya Murakita, Tetsushi Ikeda, and Hiroshi Ishiguro
Department of Adaptive Machine Systems, Osaka University
E-mail: murakita@ed.ams.eng.osaka-u.ac.jp

Abstract

We describe a human tracking system that is resistant to environmental changes and covers a wide area. Low-cost floor sensors that have a simple structure can track people in a wide area. However, sensor readings are discrete and sometimes missing. A Markov Chain Monte Carlo method (MCMC) is a promising tracking algorithm for these kinds of signals. We applied two prediction models to the MCMC: a linear Gaussian model and a highly nonlinear bipedal model. The Gaussian model was efficient in terms of computational cost while the bipedal model discriminated people more accurately than the Gaussian model. The Gaussian model can be used to track a number of people, and the Bipedal model can be used in situations where more accurate tracking is required. In the last part of the paper, we suggest a framework for an integrated tracking system based on multisensor data fusion.

1. Introduction

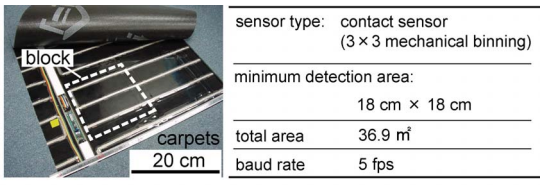
We describe a human tracking system that is fast, accurate, and resistant to environmental changes, which is required in fields such as security, traffic, surveillance, and so forth. Tracking systems employing ultrasonic sensors, infrared sensors, vision sensors, or other commonly used conventional sensors have been applied to these kinds of applications; however, they have several disadvantages such as a large detection area, costly architecture, and susceptibility to disturbance.

Floor pressure sensors avoid these disadvantages. However, little research using the sensors has been reported. For example, The ORL Active Floor [1] developed by Addlesee et al. and The Smart Floor [2] created by Orr and Abowd are used for human identification based on footstep features. Yet their research interest is human identification rather than human tracking. Morishita et al. [3] take a research approach that more closely resembles ours in developing the High Resolution Pressure Sensor. Yet the area (2.0 m by 2.0 m) is too small to perform human tracking.

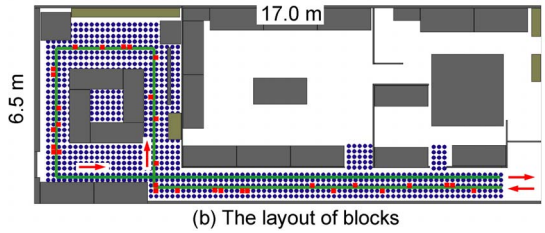
In contrast to these floor sensors, ours have a simpler structure and cover enough area (37 square meters) to track people. Lossy signals of the sensors were processed by the Markov Chain Monte Carlo method, which implemented two prediction models: a generic linear model, called the Gaussian model, and a highly nonlinear model referred to as the bipedal model. We carried out four comparative experiments to examine the tracking performance of these two models.

The experiments revealed that the two models can accurate track a person within 58 cm of the true position. The Gaussian model required much less computation, while the bipedal model discriminated people more finely. We can apply the Gaussian model to track a number of people and the bipedal model in situations where more accurate tracking is required.

Although the human tracker works well in almost all scenes of daily life, there is still room for improvement in its discrimination performance. We will tackle this issue with a multisensor data fusion technique. In the last part of the paper, the concept of integration of floor sensors, infrared sensors, and vision sensors is described.



(a) Specification of the Info Floor System



(b) The layout of blocks

Figure 1. The system architecture

2. The tracking system

We installed the InfoFloor system VS-SS-F (Vstone Corporation, Osaka, Japan) that tracks people with 1140 detection units called "blocks" (Figure 1a). A block is an 18 cm by 18 cm binary pressure sensor. An example of signals is shown as red squares in Figure 1b. The 20 seconds history of the walking signals of a person is illustrated. Ideally, there must be alternating footsteps corresponding to each foot of a person; however, there are several missing footsteps because the sensors are insulated by thick carpets.

One possible tracking algorithm for discrete and missing signals is the Kalman filter; however, the advantage of the filter is limited to linear or nearly linear systems. The filter is not effective for extremely nonlinear movements such as the rapid turn of an object, because predictions by the filter strongly depend on past observations.

It is sensible to treat human walking as a Markov process, if a position at a time step strongly depends on the position just before that time step rather than on many past positions. In a system that does not incorporate knowledge of people's habits, this may be a reasonable assumption. A Markov Chain Monte Carlo method (MCMC) is a highly suitable tracking algorithm for these kinds of signals. The advantage of the MCMC also appears when some observations have been lost. If an observation at a time step is lost and no particles hit their prediction, the particles are re-sampled assuming the observation of a uniform distribution. Then generally, the predicted distribution becomes so complicated that it cannot readily be expressed mathematically. The distribution approximated by the particles is the most probable estimation of human position. Consequently, the MCMC is very resistant to signal loss; therefore, it is suitable for tracking with simple floor sensors.

3. Implementation of the MCMC

3.1 The Gaussian model

There are various ways of formulating MCMC [4]. In the field of visual processing, Isard and Blake formulated the CONDENSATION algorithm [5]. We followed the algorithm because floor sensor signals can be considered to be binary images.

The prediction model was formulated in two ways. One of them was a generic linear model, which we call the Gaussian model:

$$\mathbf{x_t} = \mathbf{x_{t-1}} + \mathbf{v_t} + \mathbf{w_t} \qquad (1)$$

where $\mathbf{x_t}$ denotes a position of a particle at time step t, $\mathbf{v_t}$ represents a velocitie that were calculated employing the past six footsteps:

$$\mathbf{v_t} = \frac{1}{6}\sum_{i=0}^{5}\left(\mathbf{x_{t-i}} - \mathbf{x_{t-i-1}}\right) = \frac{1}{6}\left(\mathbf{x}_t - \mathbf{x}_{t-6}\right) \qquad (2)$$

$\mathbf{w_t}$ is Gaussian noise that has a variance σ^2.

The observation model for the Gaussian model was

$$p_t(\Theta_t \mid \mathbf{x}_t) = \sum_{j=1}^{M} N(\boldsymbol{\theta}_j, \sigma_{obs}) \qquad (3)$$

where the set $\Theta_t = \{\boldsymbol{\theta}_1,...,\boldsymbol{\theta}_M\}$ denotes observed M blocks and $N(\boldsymbol{\theta}_j, \sigma_{obs})$ is a normal distribution whose mean is the j[th] block $\boldsymbol{\theta}_j$ and variance σ_{obs}.

3.2 The bipedal model

The other prediction model was a bipedal model which is highly specialized for human tracking. As shown in Figure 4c, the bipedal model has four kinds of internal prediction models based on the fact that the patterns of

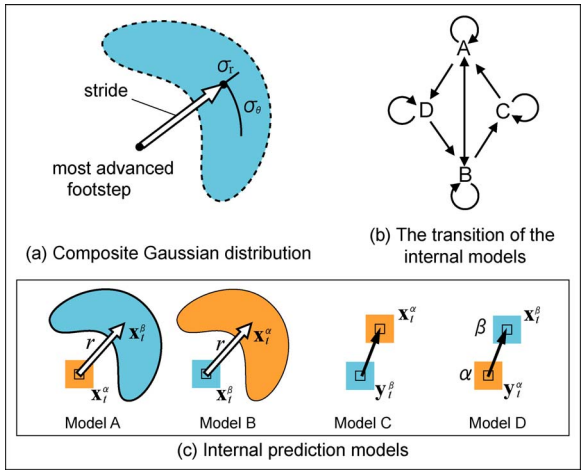


Figure 2. The bipedal model

Table 1. The transition condition

Internal model at time t-1	Observation in α domain	Observation in β domain	Weight of the particles	Internal model at time t
A	*	*	$\rho^{\alpha}\times\pi^{\beta}$	D
	*		ρ^{α}	A
		*	π^{β}	B
			1	A [1)]
B	*	*	$\pi^{\alpha}\times\rho^{\beta}$	C
	*		π^{α}	A
		*	ρ^{β}	B
			1	B [1)]
C	*	*	$\rho^{\alpha}\times\rho^{\beta}$	C
	*		ρ^{α}	A
		*		seize [2)]
			1	C [1)]
D	*	*	$\rho^{\alpha}\times\rho^{\beta}$	D
	*			seize [2)]
		*	ρ^{β}	B
			1	D [1)]

1) Tracking will be stopped after 12 times iteration. 2) Impossible transition

sensor activation by a person fall into four categories:

A. Activation by the right foot only
B. Activation by the left foot only
C. Activation by both feet with the right foot forward
D. Activation by both feet with the left foot forward

The four internal models consist of two distributions: square distribution and composite Gaussian distribution. Model A and B have a square and a composite Gaussian. Model C and D are composed of two squares.

A square distribution was formulated as follows on the lattice coordinates with a unit being a block:

$$A(\mathbf{x}) = 0.2\left(|\mathbf{x}| = 0\right),\ 0.1\left(|\mathbf{x}| = 1, \sqrt{2}\right),\ 0\left(otherwise\right) \quad (4)$$

Composite Gaussian distribution was represented in cylindrical coordinates as follows:

$$B(\mathbf{x}) = B_c(r,\theta) = \frac{1}{2\pi\sigma_r\sigma_\theta}\exp\left[-\frac{1}{2}\left(\frac{(r_0 - r)^2}{\sigma_r^2} + \frac{(\theta_0 - \theta)^2}{\sigma_\theta^2}\right)\right] \quad (5)$$
$$r_0 = |\mathbf{x}|, \quad \theta_0 = \arg(\mathbf{x})$$

where $\mathbf{x}$ is a vector in cartesian coordinates. σ_r and σ_θ are standard deviations of Gaussian noise.

The bipedal model was represented by 13 dimensional particles:

$$\mathbf{s}_t = \left[\left(\mathbf{x}_t^\alpha\right)^T \quad \left(\mathbf{x}_t^\beta\right)^T \quad \left(\mathbf{y}_t^\alpha\right)^T \quad \left(\mathbf{y}_t^\beta\right)^T \quad r_t\right]^T \quad (5)$$

where $\mathbf{x}_t^\alpha$ or $\mathbf{x}_t^\beta$ represent the mean of blocks currently activated by the most advanced footstep. $\mathbf{y}_t^\alpha$ and $\mathbf{y}_t^\beta$ denote the mean of predecessor footsteps when sensors are activated by both feet. $\mathbf{y}_t^\alpha$ and $\mathbf{y}_t^\beta$ are ignored in model A and B. r indicates the length of the stride inherited from the previous particle. The correspondence of notation α and β to the left or right foot is dependent on initialization.

A combination of square distributions or square and composite Gaussian distribution is applied to the elements of the particle. For example, in the case of model A:

$$\mathbf{x}_{t+1}^\alpha = \mathbf{x}_t^\alpha + \mathbf{w}_a$$
$$\mathbf{x}_{t+1}^\beta = \mathbf{x}_t^\alpha + \mathbf{w}_b \quad (6)$$

where $\mathbf{w}_a$ and $\mathbf{w}_b$ are noise that are generated from square and composite Gaussian distribution respectively. $\mathbf{x}_t^\alpha, \mathbf{x}_t^\beta, \mathbf{y}_t^\alpha, \mathbf{y}_t^\beta$ have internal weights $\pi^\alpha, \pi^\beta, \rho^\alpha, \rho^\beta$ respectively. The weights determine a total weight π according to Table 1. As shown in Figure 4b, internal models are changed at each time step based on the observations.

4. Experimental results

4.1. Tracking performance and the required number of particles for the MCMC

We show the effectiveness of floor sensors and the MCMC by four comparative experiments to be shown in section 4.1 through 4.3. Firstly, tracking accuracy vs. walking speeds was examined.

10 people walked a 30 meter test course as shown in Figure 1b. They did three times at arbitrary speeds, and we obtained 30 test samples. Tracking accuracy was regarded as the number of tracking errors. The parameter of the Gaussian model is the standard deviation (S.D.) σ of the Gaussian noise. The bipedal model has two parameters: σ_r , σ_θ of the composite Gaussian noise as illustrated in Figure 2a. In the experiment, these two models were examined for three values of parameters that were chosen according to the result of preliminary experiments. In experiment 4.1 and 4.2, the proportion of the two parameters of the bipedal model was fixed to $\sigma_\theta = 3\sigma_r$. The result is shown in Figure 3.

Second, tracking accuracy and the number of particles were examined employing the same data set. The number of tracking errors was summed to absorb the variances in walking speeds. The result is shown in Figure 4.

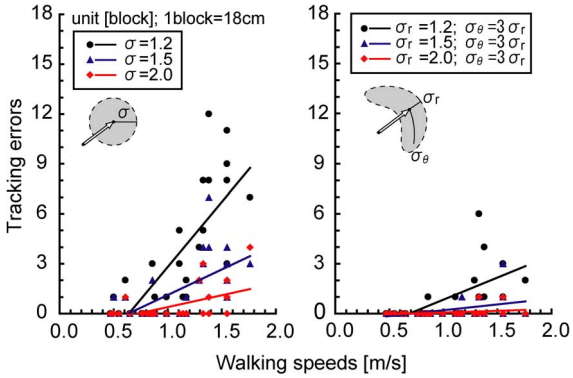


Figure 3. Tracking errors vs. walking speeds

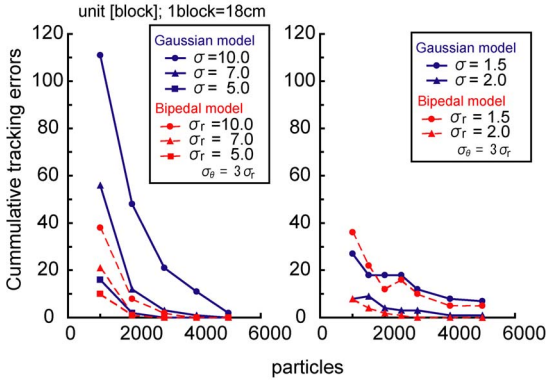
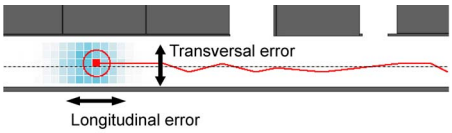


Figure 4. Tracking errors vs. number of particles

4.2. Estimation errors of human positions

It is very difficult to measure positions of people correctly; therefore, we assumed that test subjects walk along the test course precisely with uniform walking speed, and that an error of position can be decomposed into two directions: a transversal error and a longitudinal error. As shown in Figure 5 (top), the transversal error is a lateral deviation from the test course, and the longitudinal error is a deviation from uniform walking in the direction of travel.

We examined the errors of 10 test subjects and averaged them (Figure 5). The errors at the corners of the test course were excluded because of difficulty in defining the errors.



Gaussian model

	Long.	Trans.
max	**0.41** [m]	0.09 [m]
min	0 [m]	0 [m]
mean	0.19 [m]	0.08 [m]

Bipedal model

	Long.	Trans.
max	0.59 [m]	0.09 [m]
min	0 [m]	0 [m]
mean	0.21 [m]	**0.01** [m]

Figure 5. Estimation errors of human position

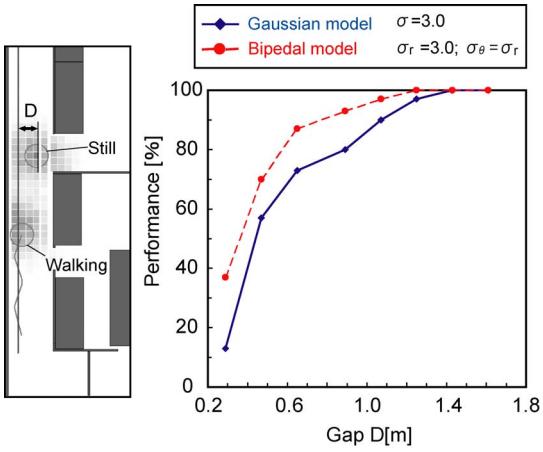


Figure 6. Discriminating two persons

4.3. Discriminating people in a crowded area

The tracking system often fails to track people in a crowded area. Assuming that the system would be applied to everyday use, these kinds of situations are inevitable; therefore, it is necessary to examine to what extent the system can discriminate. We settle the problem as shown in Figure 6 (left).

There are two persons. One of them stands still during the experiment. The other person walks close to the still person. If the system could discriminate them, the two persons are continuously tracked. However, if not, two trackers may track only one of the two people. We examined 30 test samples with a variety of gaps D. The results are shown in Figure 6 (right).

5. Discussion and Conclusion

Figure 3 shows that the tracking errors became zero if $\sigma \geq 3.0$ for the Gaussian model and $\sigma_r \geq 2.0$ for the bipedal model. However, larger variances require more particles to properly approximate a prediction distribution. If the number of particles is not enough, the distribution becomes sparse and tracking errors occur. Figure 4 demonstrates that.

A notable point is that the tracking errors become larger despite less variance (fig. 4 right). This is not owing to a sparse distribution but to an inaccurate mean of the distribution. Therefore, the number of tracking errors is not improved even when the number of particles becomes greater. This reversal shows that the tracking system requires a minimum number of particles at around $\sigma = 3.0$ and $\sigma_r = 2.0$. These values coincide with the result of the former experiment as mentioned above. This is reasonable because the minimal number of particles is required to have an accurate prediction distribution.

Figure 5 shows that the system can track people with a mean error of about 20 cm. An error of a tracking system using vision sensors examined by Sogo et al. [6] was 0.17 m at its maximum. The S.D. of the errors was 0.0393 m. In terms of computational efficiency, the estimation precision of the floor sensors is not worse than that of the distributed omnidirectional vision system (DOVS), although the evaluation methods differ.

Figure 6 shows that the floor sensors discriminated two people perfectly, if the gap D is larger than 1.42 m. Even if the interval is only 0.8 m, the bipedal model discriminated two people at nearly 90% accuracy. For every interval smaller than 1.4 m, the bipedal model discriminated more effectively than the Gaussian model.

Based on these experimental results, we could make the following conclusions:

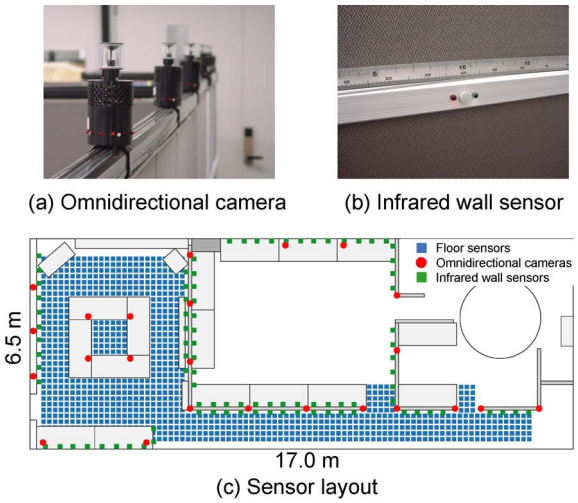
(1) The system can perfectly track a person who walks alone.
(2) The system can track a person within 59 cm error.
(3) The system can independently track people at 90% accuracy, if they maintain a gap of more than 80 cm.

These facts indicate that the system will work in everyday use.

6. Further discussion

6.1 A concept of integrated tracking system using floor sensors, infrared sensors, and vision sensors

This section describes a concept of data fusion among heterogeneous sensors. As shown in Figure 7, we have 24 omnidirectional vision sensors and 100 infrared wall sensors in addition to 1140 floor sensors. They work to complement each other rather than compete (Table 2). For example, vision sensors, using the background subtraction method, hardly fail to detect a human silhouette even when floor sensors fail because of light



(a) Omnidirectional camera (b) Infrared wall sensor
(c) Sensor layout

Figure 7. Sensors to be added

body weight, and so on. On the other hand, vision sensors are very sensitive to change of brightness. Yet infrared sensors and/or floor sensors will be able to detect people even if the vision sensors cannot. Thus, we can expect that a multisensor data fusion technique will significantly decrease tracking and estimation errors, while increasing the discrimination performance of the integrated tracking system.

As illustrated in the Joint Directors of Laboratories (JDL) data fusion model [7], there are several stages at which data from multiple sensors are fused. Data from our integrated tracking system can be fused at two levels: the physical level and the feature level. At the physical level, data from heterogeneous and homogeneous sensors are fused at the same time. We consider the MCMC is most effective for the physical level data fusion. Meanwhile at the feature level, human positions are calculated separately by the three kinds of sensors, and they are fused based on classical stochastic method.

6.2 Physical level data fusion based on the Markov chain Monte Carlo method

At the physical level, all data from individual sensors are fused at the same time. Then human position detection such as triangulation will be very complicated because there are too many sensors and geometrical constraints. The MCMC offers a lucid theoretical framework, which enables the tracking system to be used in a wider area with minimal computation (Figure 8a).

We expect that the following observation models are statistically obtained. Firstly, the observation model of the floor sensor will be Gaussian as mentioned in section three. However, the effect of time decay should be added in order for the observation model to be consistent with the faster frame rate of the other sensors. Given an activated floor sensor $\mathbf{x}_{t=T}$ at time T, the observation model will be described as:

Table 2. Sensor characteristics

Sensor	Primary observable data	Primary disturbance	Preprocessing	Principle of human position detection	Estimation error of human position
Omnidirectional camera	Visible light	Change of lightness	Background subtraction	Triangulation	0.17 m
Floor pressure sensor	Mechanical pressure	Variable pressurization	-----	Directly detect human weight	0.41 m
Infrared wall sensor	Infrared ray	Heat source	Frame difference	Triangulation	Much larger than the two sensors

$$p^{(f)}(\mathbf{z}_t \mid \mathbf{x}_{t=T}) = N(\mathbf{x}_{t=T}, \sigma_t)$$
$$\sigma_t = c \cdot V \cdot (t - T) \qquad (7)$$

where V denotes walking speed. c indicates a constant which is obtained empirically. If the sensor obtains no observation, the observation model becomes a uniform distribution.

Observation models of the vision sensors and infrared wall sensors are expressed in cylindrical coordinates as follows:

$$p^{(v\ or\ w)}(\mathbf{z}_t \mid \mathbf{x}_t) = \frac{1}{2\pi\sigma_r\sigma_\theta}\exp\left[-\frac{1}{2}\left(\frac{(r_0 - r)^2}{\sigma_r^2} + \frac{(\theta_0 - \theta)^2}{\sigma_\theta^2}\right)\right] \qquad (8)$$

where (r_0, θ_0) corresponds to $\mathbf{x}_t$.

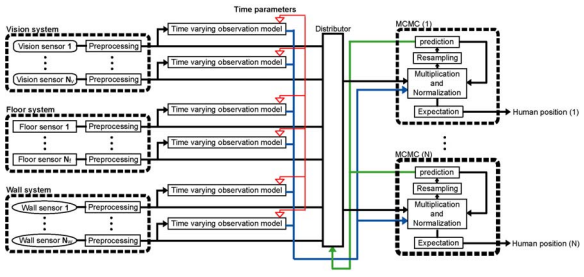
If observation models of the three sensors are given, a prediction distribution of a person is represented as:

$$p(\mathbf{x}_t \mid \mathbf{z}_t) = k \cdot \prod_{i=1}^{24} p_i^{(v)}(\mathbf{z}_t \mid \mathbf{x}_t) \times \prod_{j=1}^{1140} p_j^{(f)}(\mathbf{z}_t \mid \mathbf{x}_t) \times \prod_{k=1}^{100} p_k^{(w)}(\mathbf{z}_t \mid \mathbf{x}_t) \times p(\mathbf{x}_{t-1}) \qquad (9)$$

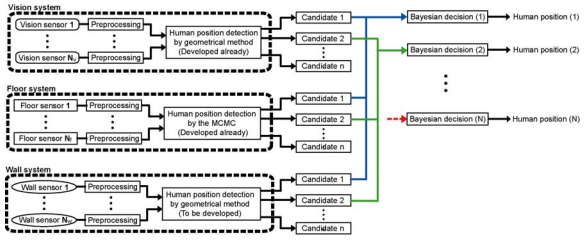
where k is a normalization constant.

Then a possible prediction model may be an isotropic model:

$$p(\mathbf{x}_t \mid \mathbf{x}_{t-1}) = \mathbf{x}_{t-1} + N(\sigma_p) \qquad (10)$$



(a) Physical level data fusion



(b) Feature level data fusion

Figure 8. Two levels of data fusion

where $N(\sigma_p)$ is a Gaussian whose standard deviation is σ_p. σ_p is proportional to the walking speed. It seems that consideration of velocity for a prediction model, such as in equation 4, is no longer effective, because the frame rate of the integrated tracking system, which is synchronized with that of the vision sensors, is relatively high.

6.3 Feature level data fusion based on a classical stochastic framework

We have already developed human trackers using floor sensors and omnidirectional vision sensors [6]. They work well if the environment is stable. One of the most straightforward and traditional method for their data fusion is to simply choose the best estimation of the person's position based on the Bayesian decision rule (Figure 8b).

The classical stochastic method may not be efficient if the scale of the tracking system become greater (larger area and/or higher resolution) because an exponential amount of computational resources is required for human position detection. However, if the scale is reasonable, it provides a robust approach to human position estimation, which would be more accurate than the MCMC approach.

We expect that physical level data fusion using the MCMC is most appropriate for tracking in a wide area and that the feature level data fusion based on a classical stochastic framework is best suited for precise tracking in a confined area.

7. References

[1] M. D. Addlesee, A. Jones, F. Livesey and F. Samaria. "The ORL Active Floor", *IEEE Personal Communications*, **5**(4), 1997, 35-41.

[2] *Proceedings of the 2000 Conference on Human Factors in Computing Systems,* The Hague, Netherlands, April 2000.

[3] *Proceedings of the 2002 IEEE/RSJ International Conference on Intelligent Robots and Systems EPFL*, Lausanne, Switzerland, October 2002.

[4] R. M. Neal. "Probabilistic inference using Markov chain Monte Carlo methods". *Technical Report CRG-TR-93-1*, *Department of Computer Science*, University of Toronto, 1993.

[5] M. Isard and A. Blake, "Contour tracking by stochastic propagation of conditional density", *Proceedings of the European Conference on Computer Vision*, Cambridge, UK, 1996.

[6] T. Sogo, H. Ishiguro, and M.M. Trivedi, "N-ocular stereo for real-time human tracking", *In Panoramic Vision: Sensors, Theory and Applications (R. Benosman and S.B. Kang, Eds.)*, Springer Verlag, 2001.

[7] D. L. Hall and J. Llinas, "An Introduction to Multisensor Data fusion", *Proceedings of the IEEE*, **85**(1), Jan 1997.