Human Gait Estimation Using a Wearable Camera

Yoshihiro Watanabe, Tetsuo Hatanaka, Takashi Komuro and Masatoshi Ishikawa
Graduate School of Information Science and Technology, University of Tokyo
7-3-1 Hongo, Bunkyo-ku, Tokyo 113-8656, Japan.
Yoshihiro_Watanabe@ipc.i.u-tokyo.ac.jp

Abstract

We focus on the growing need for a technology that can achieve motion capture in outdoor environments. The conventional approaches have relied mainly on fixed installed cameras. With this approach, however, it is difficult to capture motion in everyday surroundings. This paper describes a new method for motion estimation using a single wearable camera. We focused on walking motion. The key point is how the system can estimate the original walking state using limited information from a wearable sensor. This paper describes three aspects: the configuration of the sensing system, gait representation, and the gait estimation method.

1. Introduction

There is a growing need for ubiquitous sensing systems that are capable of estimating the position and pose of humans. This type of sensing system is expected to contribute greatly to the realization of applications in which human body motion is captured and predicted in order to provide cooperative assistance, for example, using image projection systems, robots, and mobility systems, as illustrated in Figure 1. The key point in such ubiquitous motion capture is to realize that the sensing should not be restricted by location.

Conventional motion capture systems have been designed mainly to generate computer graphic animations. In typical systems, multiple cameras installed at various positions around a subject (fixed cameras) observe optical markers mounted on the subject's joints [10, 3, 2]. Recently, a marker-less technique has been developed [9]. However, multiple fixed cameras are required to observe a single person at the same time without occlusions from the viewpoint of each camera. This approach is considered to hinder achievement of the goal described above.

A system configured of wearable sensors is a highly promising approach. For example, Vlasic et al. developed a novel motion capture system using accelerometers, gyroscopes, ultrasonic sources, and microphones [12]. However, one disadvantage is that the user is required to attach many sensors at the desired parts of the body. Also, because the positions of the body parts are measured using ultrasound, simultaneous usage of many such systems in crowded areas is difficult.

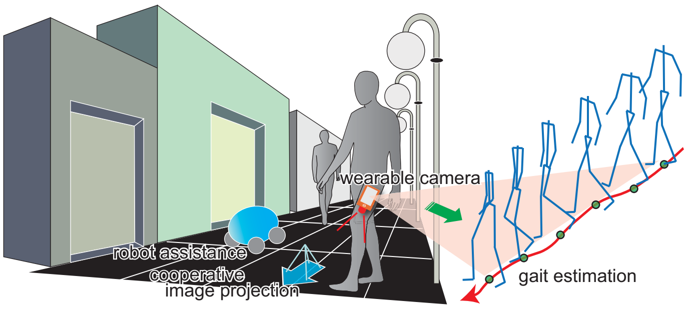

Figure 1. Human gait estimation using a wearable camera.

On the other hand, Hamaguchi et al. developed a wearable system for acquiring the walking state of a user [7]. In this system, the relative position from the user's waist to heel is measured by using an electromagnetic tracker. The position of the user is updated by using this data when a button placed on the sole is pushed. However, the issue of the complex configuration of multiple wearable sensors still remains. Also there is a system which recognizes activity by using a wearable accelerometers [5]. This system cannot utilize landmarks in the environment, so that it is difficult to avoid accumulated errors in the user's position.

On the other hand, a wearable-sensor technique using a single camera has been reported [4]. This is based on a method called Simultaneous Localisation and Mapping (SLAM), which estimates the camera motion and constructs a map of the surrounding environment. The report shows that the system can estimate its own location in an unknown environment using a single camera. Although this is not a technique for estimating the body pose of a human subject with a wearable camera, the framework is nevertheless promising for our purpose.

The key point of our system is to estimate the local body pose and the global position by using the minimum number of wearable sensors. In this paper, we focus on the possibility that the gait can be described as cyclic foot motions. This allows us to assume that the obtained sensor data follows a behavior based on fixed models, so that it is feasible to reconstruct the original state from the limited sensor data.

2. Motion estimation using a wearable camera

In this paper, we propose a system for motion estimation of a human subject using only a single camera. Recently, the size and the cost of sensors, including cameras, have been drastically reduced. Indeed, many mobile devices now come with such sensors installed. Therefore, it is becoming commonplace to assume that such sensors are available anywhere and anytime.

Figure 1 shows the configuration of the proposed system. The user mounts a camera on his leg, with the camera pointing downward. The camera attached to the user's body captures changes in an image of the external environment caused by the motion. We consider this setup sufficient to estimate the gait. By utilizing the observed information in conjunction with models of human motion, our technique estimates both the current actual motion and the global position. Also, because the camera can observe the external environment, it should be possible to reduce the accumulated errors that occur in sensors such as accelerometers.

The direction of the camera is not an essential requirement. In the downward orientation, we intend to remove moving objects, including other people, around the user to simplify the image recognition. A setup where the camera faces forward is also feasible, and in that case the amount of information available for location detection is expected to increase.

Under this configuration, the task involves the problem of pose estimation using a single camera under a situation where the motion is based on given models. The motion model is not unique and varies with time according to specific rules. The whole body motion must also be estimated from the camera pose.

3. Gait representation

The walking motion needs to be prepared in advance in order to recognize which motion causes changes in the captured images. This section describes the gait representation. In this paper, we assume that such advance knowledge of the gait is constructed for each person.

The inverted pendulum model is a well-known simple model of bipedal walking [6]. It is mainly used to create walking patterns of robots. However, it is not effective to complement insufficient sensor data because the inverted pendulum model does not accurately describe human walking motion.

On the other hand, in the field of computer graphics, animations have been created mainly by using the motion data directly obtained by motion capture systems. For example, Unuma et al. used a Fourier expansion in order to interpolate or extrapolate human walking [11]. This kind of frequency analysis is considered to be effective because most walking motions can be characterized as cyclic motion. The reported results show that the walking can be represented using only low-frequency information with a small number of parameters for animation. We also employed a motion-data-based representation to take advantage of this feature.

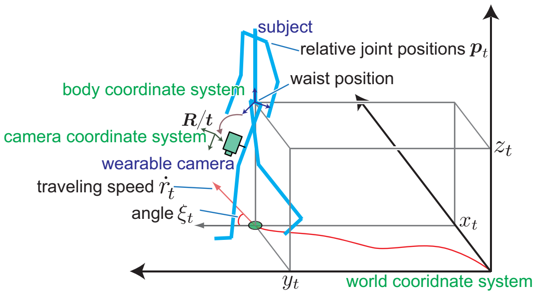

Figure 2. Parameter definitions.

The definitions of parameters are given here. In this paper, the walking state is updated with the following equation:

$$\boldsymbol{\mu}_t = \begin{bmatrix} x_t \\ y_t \\ z_t \\ \xi_t \\ \boldsymbol{p}_t \end{bmatrix} = \begin{bmatrix} x_{t-1} + \dot{r}(\eta_{t-1})\cos(\xi_{t-1}) \\ y_{t-1} + \dot{r}(\eta_{t-1})\sin(\xi_{t-1}) \\ z(\eta_{t-1}) \\ \xi_{t-1} + \dot{\xi}(\eta_{t-1}) \\ \boldsymbol{p}(\eta_{t-1}) \end{bmatrix} \quad (1)$$

Here, $[x_t, y_t, z_t]$ describes the waist position at time t, $\dot{r}_t$ is the speed in the traveling direction, and $\dot{\xi}_t$ is the angular speed. The behaviors are parameterized based on cyclic motion with phase η_t. Also, the collected data indicating the relative positions of all joints, with the waist position set as an origin, is $\boldsymbol{p}_t = \{\boldsymbol{p}_t^i | i = 1, \ldots, N_p\}$. Figure 2 illustrates the parameter definitions.

There are three coordinate systems: those attached to the world, the subject's body, and the camera. The rigid transformation between the world and the body is described by using the walking state. Also, the camera is fixed to the body, and the rigid transformation between the body and the camera is assumed to be known in advance.

Based on this definition, as the gait data, the behaviors of parameters $\boldsymbol{d}_t = [z_t, \dot{r}_t, \dot{\xi}_t, \boldsymbol{p}_t]$ are sampled in advance. Walking samples are collected by using a motion capture system. In this paper, we represent the cyclic motion by using a Fourier expansion of up to fifth order. An example in the case of the waist height is:

$$z_t = \sum_{n=0}^{5} a_n^0 \sin(n\eta_t + b_n^0). \quad (2)$$

Using this representation, the behaviors are parameterized by using (a_n^i, b_n^i). Here, the superscript i is an index identifying each parameter. The sampled gait motion is stored in the system as the feature vector $\boldsymbol{f} = \{a_n^i, b_n^i | i = 1, \ldots, N_p + 3\}$.

4. Classification of the walking models

Human walking motion has some variations. This section describes the method used to classify gaits. In this paper, we call each classified motion a walking form. Also we assume that the number of forms in steady walking is finite. The system is assumed to have a sufficient number of forms.

First, many walking samples are collected using a motion-capture system. The samples are converted to vectors $\boldsymbol{f}_k$ by applying the frequency analysis described in

Section 3. This collected data $\boldsymbol{f}_{k=1,\dots,N_s}$ is compressed to lower dimension and projected onto classification space. The projected points are classified by using a clustering method.

In this paper, principal component analysis is used to project the sampled data onto three-dimensional space. The following conversion is applied.

$$\boldsymbol{g}_k = \boldsymbol{U}(\boldsymbol{f}_k - \bar{\boldsymbol{f}}) \tag{3}$$

Here, $\boldsymbol{U}$ is the normalized orthogonal basis, $\bar{\boldsymbol{f}}$ is the mean vector of all data, and $\boldsymbol{g}_k$ is the compressed feature. Also, we used the k-means clustering method for classification [8].

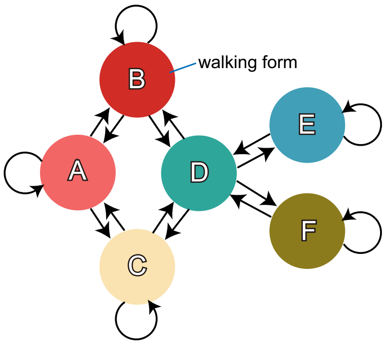

Figure 3. Transition diagram of walking forms.

Our method classifies the walking forms and generates a transition diagram describing the connection relationships between the walking forms. An example diagram is shown in Figure 3. This diagram clarifies the time-sequential changes of forms, which should make it possible to narrow the search area to achieve robust estimation. This diagram is generated based on observations made in collecting samples.

5. Human gait estimation

5.1. Overview

In the proposed method, the walking form and the present state are estimated from the changes observed in a wearable camera. The walking state for this estimation is defined as $\tilde{\boldsymbol{\mu}}_t = [x_t, y_t, \xi_t, \eta_t, \dot{\eta}_t]$. The complete state $\boldsymbol{\mu}_t$ defined in Section 3 can be reconstructed by using the estimated state $\tilde{\boldsymbol{\mu}}_t$ and the form data $\boldsymbol{f}_k$.

Our method estimates the walking state at every sampling of sensor data. An overview of the method is shown in Figure 4. In this method, multiple candidate states are generated. The total number of candidates at time t is M_t. Each candidate is tested through five steps: selecting walking forms, state prediction, likelihood estimation, resampling, and state updating. The state prediction and updating are based on the extended Kalman filter (EKF). The details of each step are described in the following subsections.

In this paper, we assume that the world coordinates of the observed points in the environment are known. There is a possibility of removing this assumption by using the SLAM framework. It should be possible to introduce this approach in our method by extending the state in the EKF.

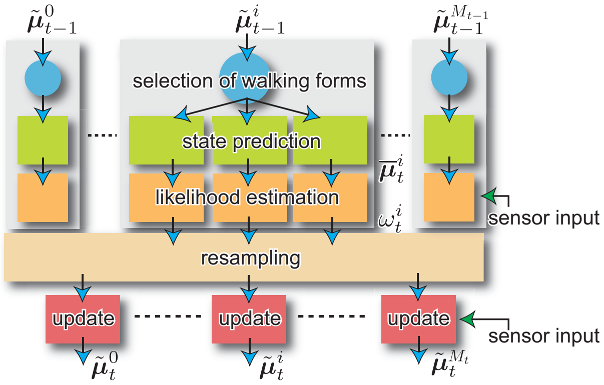
$\tilde{\boldsymbol{\mu}}^0_{t-1}$
$\tilde{\boldsymbol{\mu}}^i_{t-1}$
$\tilde{\boldsymbol{\mu}}^{M_{t-1}}_{t-1}$
selection of walking forms
state prediction
likelihood estimation
$\bar{\boldsymbol{\mu}}^i_t$
ω^i_t
sensor input
resampling
update
update
update
sensor input
$\tilde{\boldsymbol{\mu}}^0_t$
$\tilde{\boldsymbol{\mu}}^i_t$
$\tilde{\boldsymbol{\mu}}^{M_t}_t$

Figure 4. Overview of the walking estimation technique.

In addition, the intrinsic parameters of the camera used are assumed to be known in advance. Therefore, if we obtain the walking state, we can project the observed spatial points onto the image plane.

5.2. Selection of walking forms

The i-th candidate state generated at time $t-1$ is represented as $\tilde{\boldsymbol{\mu}}^i_{t-1}$. The walking forms for the tests of this state at time t are selected based on the state transition diagram. For example, the candidate state $\tilde{\boldsymbol{\mu}}^i_{t-1}$ is updated at that time by using the walking data of the A-form shown in Figure 3. In this case, this state is tested in the A-form and the connected B- and C-forms, as defined in the state transition diagram.

5.3. Prediction of walking state

In the selection step, the walking form at time t for the candidate state $\tilde{\boldsymbol{\mu}}^i_{t-1}$ is selected. State prediction is applied using the walking feature of the selected form $\boldsymbol{f}_k$. The equation of state is

$$\tilde{\boldsymbol{\mu}}^i_t = \boldsymbol{g}(\tilde{\boldsymbol{\mu}}^i_{t-1}) + \epsilon_t = \bar{\boldsymbol{\mu}}^i_t + \epsilon_t. \tag{4}$$

Here, ϵ_t is process noise, and $\boldsymbol{g}$ is a nonlinear state prediction function. In this step, we only calculate the vector $\bar{\boldsymbol{\mu}}^i_t$. In the function $\boldsymbol{g}$, the parameters $[x_t, y_t, \xi_t]$ in the state $\tilde{\boldsymbol{\mu}}^i_t$ are predicted based on equation (1). The phase parameters are predicted by the following equation:

$$\begin{bmatrix} \eta_t \\ \dot{\eta}_t \end{bmatrix} = \begin{bmatrix} \eta_{t-1} + \dot{\eta}_{t-1} \\ \dot{\eta}_{t-1} \end{bmatrix}. \tag{5}$$

5.4. Likelihood estimation and resampling

The likelihood of the i-th candidate state ω^i_t is estimated using the likelihood at the previous time $t-1$ based on the following equation:

$$\omega^i_t = C\lambda^i_t \omega^i_{t-1}. \tag{6}$$

In this update, if the candidate is predicted based on the same walking form at the previous time, a higher value is given to the likelihood as an advantage. This operation is controlled by the parameter C. If the forms between two successive estimations are the same, C is set to be larger than one; otherwise C is set to one.

The parameter λ_t^i is the reliability of the prediction tested by the sensing data. Using the predicted state, the spatial points based on world coordinates can be projected onto the image plane of the camera. Using the sum of the errors between the projected points and the observed corresponding points, we define the reliability λ_t^i as

$$\lambda_t^i = \frac{1}{\sqrt{2\pi}\sigma} \exp\{-\frac{(\mid \boldsymbol{u}_t - \boldsymbol{h}(\overline{\boldsymbol{\mu}}_t^i) \mid / N_t - \nu)^2}{2\sigma^2}\}. \tag{7}$$

Here, $\boldsymbol{u}_t$ is the N_t observed image points, and $\boldsymbol{h}(\overline{\boldsymbol{\mu}}_t)$ is the projected image points of the corresponding known spatial points. As shown in the equation, the error is represented using a Gaussian distribution. In the experiment, we set the average ν and the variance σ^2 as 0 and 10, respectively.

In the resampling step, the candidate states are filtered. Our method leaves only the candidates whose likelihoods are high. The state update described in the next section is applied to those remaining candidates. For the sake of convenience, the numbering of the candidate states is initialized after this operation.

5.5. State update

The prediction errors obtained in the previous step are caused by the sensor errors and differences between the stored walking form and the observed one. It remains possible that the observed form is slightly different from the prepared one. The error is eliminated by the update operation using the captured sensor data. The observation model is described by

$$\boldsymbol{u}_t = \boldsymbol{h}(\tilde{\boldsymbol{\mu}}_t) + \delta_t. \tag{8}$$

The update operation is based on the following equations, namely, the framework of the EKF technique:

$$\begin{aligned} \overline{\boldsymbol{\Sigma}}_t &= \boldsymbol{G}_t \boldsymbol{\Sigma}_{t-1} \boldsymbol{G}_t^T + \boldsymbol{P}_t & (9)\\ \boldsymbol{K}_t &= \overline{\boldsymbol{\Sigma}}_t \boldsymbol{H}_t^T (\boldsymbol{H}_t \overline{\boldsymbol{\Sigma}}_t \boldsymbol{H}_t^T + \boldsymbol{Q}_t)^{-1} & (10)\\ \tilde{\boldsymbol{\mu}}_t &= \overline{\boldsymbol{\mu}}_t + \boldsymbol{K}_t(\boldsymbol{u}_t - \boldsymbol{h}(\overline{\boldsymbol{\mu}}_t)). & (11) \end{aligned}$$

Here, $\boldsymbol{G}_t$ is the Jacobian of the state prediction $\boldsymbol{g}$ described in Section 5.3, $\boldsymbol{H}_t$ is the Jacobian of the observation model $\boldsymbol{h}$ described in Section 5.4, and $\boldsymbol{P}_t, \boldsymbol{Q}_t, \sum_t$ are the covariance matrices of process noise, observation noise, and the system.

In our method, the state prediction $\boldsymbol{g}$ is changed with the transition of forms. Although this could cause a problem in the EKF technique, from the experiments, we confirmed that this is not a problem in the estimation because of the robustness in EKF as well as our design, where the transitions of the forms at the possible phases are smooth.

6. Experiment

6.1. Simulation using sample walking data

In this experiment, we used the data in the motion capture library of the Carnegie Mellon University Graphics Lab [1]. The data was for five gaits, including two types of "slow walk", two types of "walk", and one type of "run". All walks were performed by the same person.

First, the five sets of data were converted to frequency features $\boldsymbol{f}$ using the method described in Section 3. Next, those features were classified using the method described in Section 4. The classification result is shown in Figure 5. The features were classified into three types. The points in the same cluster are shown in the same color. The three classified types correspond to "slow walk", "walk", and "run", respectively. This means that the classified results matched the initial classification decided visually based on the walking speed. The transition diagram is shown in Figure 5 by a black line. In this experiment, the connection was simple. Transitions are allowed only between "slow walk" and "walk", and between "walk" and "run". Also, the walking data stored in the system is selected from those data. For "walk" and "slow walk", one of two data sets was used. For "run", the single data set was used directly.

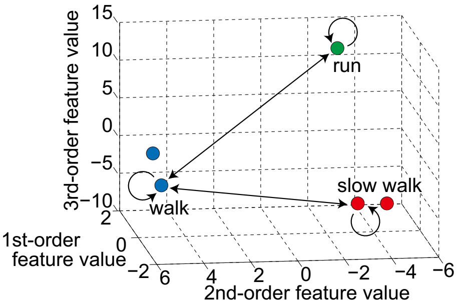

Figure 5. Classification result and transition diagram.

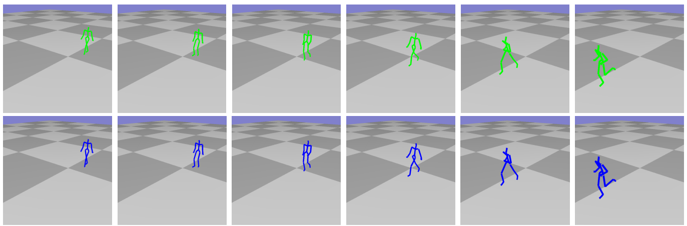

Figure 6. The input human gait and the estimated gait.

The generated walking motion was created by connecting the three types of walking forms. In this experiment, the order of form changes was "slow walk", "walk", and "run". For the "slow walk" and "walk" motions, the data that was not selected in the advance setup described above was used. The "run" motion was the same as the one that the system stores. The sensor input was created by projecting the spatial points onto the camera image plane based on the created motion. The camera was assumed to be placed on the outer side of the right thigh.

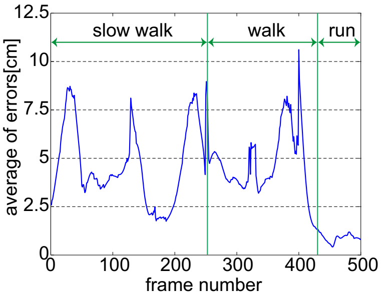

Figure 7. Average errors in joint positions.

The estimation result is shown in Figure 6. The green subject in the upper part of the figures shows the input motion. Also the blue subject in the bottom part of the figure shows the estimated motion. The pairs of figures arranged one above another show the poses at the same point in time. Figure 7 shows the time variation of the average error in the joint positions. The errors were within 10 cm, and so the achieved accuracy is considered to be high. Although in this experiment the actual motion and the stored motion were different and the timings at which the forms change were unknown, the form transition and the state update in the estimation were not influenced by such problems, and worked well to reconstruct the original walking motions.

6.2. Experiment using actual sensing data

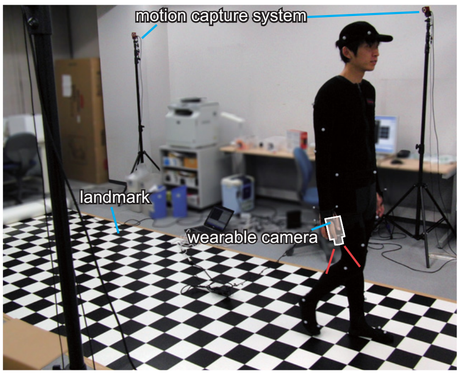

Figure 8. Photograph of the experimental environment.

This section describes an experiment using the actual sensing data. A photograph of the experiment is shown in Figure 8. In this experiment, the camera was placed on the outer side of the right thigh. The camera used was a Basler Pioneer. In this experiment, we used 400 × 400-pixel images captured at 250 fps. A chessboard pattern was set on the ground so that we assumed that the world coordinates of the observed points were known. Also, the observation model described in Section 5.5 does not involve lens distortion. Therefore, the images were used after applying distortion correction.

First, the walking motions were sampled. The motion capture system used for this sample collection was a NaturalPoint OptiTrack. The sampled motions were classified

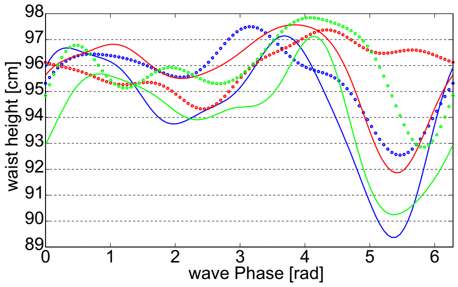

Figure 9. Sample waveform of six different motions.

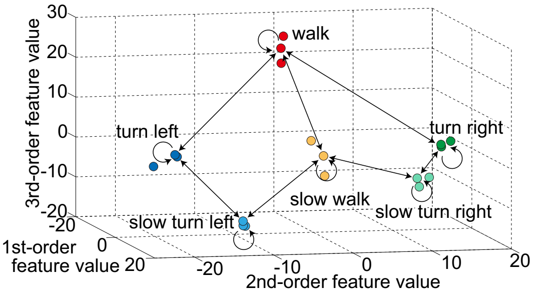

Figure 10. The classified result using the actual sensor data.

after being converted to frequency features. As an example, the waist height behaviors of six different motions are shown in different colors in Figure 9. The classification result is shown in Figure 10. The motions were classified into six forms: “walk”, “left-turning walk”, and “right-turning walk” each at two different rates. Three samples were collected for each motion. The motion of each group that was closest to the center of each cluster was stored in the system. Also, the defined transitions are shown in Figure 10 as the black lines.

Based on this setup, the estimation was tested during walking. In the estimation, the subject walked straight ahead, turned left, and walked straight ahead again. Also, he attempted to walk at slow speed. A Harris detector was used to detect the feature points in the captured image. The results of estimation are shown in Figure 11. In the figure, the photographs captured during this experiment and the corresponding estimated poses are compared.

Also, Figure 12 shows the time variation of the average errors in the image plane between the observed image points and the image points at which the corresponding spatial points were projected based on the estimation results. A sufficient number of feature points in the images could not be detected, so that the errors became high. However, most errors were under around 10 pixels. Also even when the feature points were not detected and the walking motion was different from the stored motion, the estimation was stable.

Figure 13 shows the time variation of the estimated walking forms. In this experiment, if the subject achieved the ideal motion, the observed walking could be represented using only three forms. However, in the figure, the forms varied. This was because similar poses were included in neighboring motions under a situation where the system captured

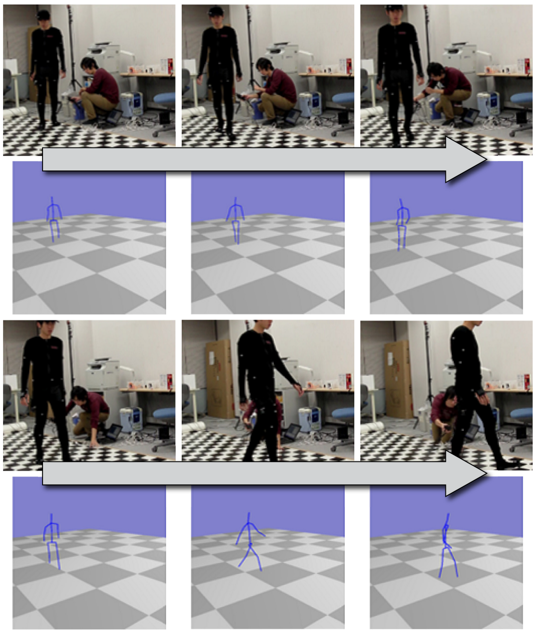

Figure 11. Estimation results of walking motion.

motions for a brief time. This is not a critical problem if the estimated pose is unnatural. One way to improve the estimation accuracy is to introduce a physical model of the form change instead of representing it using the values expressed as equation (6).

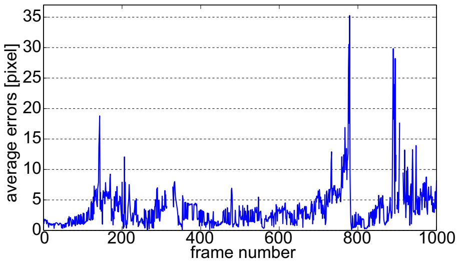

Figure 12. Time variation of errors evaluated by projecting the spatial points onto the image plane based on the estimated pose.

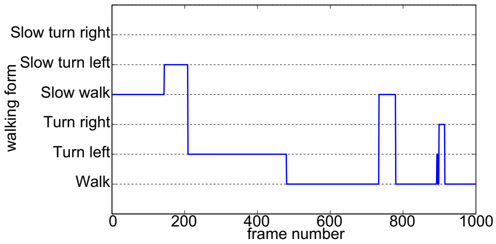

Figure 13. Time variation of the walking forms.

7. Conclusion

Ubiquitous motion capture technology will become increasingly important in many applications. This paper proposes a technique for estimating human gait with a simple sensor configuration. In our method, advance knowledge of the user's motion allows use of only a single camera as the wearable sensor.

This is a new approach to wearable sensors, with which it is possible to estimate the body motion and the global position using only the changes in images of the external environment caused by the user's motion. Another novel aspect is the proposed representation of the various walking motions in a compact dataset. We also proposed utilizing this sensing configuration and the stored knowledge to achieve effective walking estimation. This method integrates changes in walking form using a transition diagram and state estimation using an extended Kalman filter (EKF). Experiments showed that our method worked well when applied to actual human motion.

As a next step, we plan to add typical non-cyclic motions, such as stumbling. Also the system needs to focus on not only motions on flat ground but also on stairs. In addition, the system must be able to handle unknown environments.

References

[1] CMU graphics lab motion capture database. http://mocap.cs.cmu.edu/.

[2] OptiTrack. http://www.naturalpoint.com/optitrack/.

[3] VICON. http://www.vicon.com/.

[4] A. J. Davison, I. D. Reid, N. D. Molton, and O. Stasse. MonoSLAM: Real-time single camera SLAM. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 29(6):1052–1067, 2007.

[5] J. Frank, S. Mannor, and D. Precup. Activity and gait recognition with time-delay embeddings. In *Proc. 24th AAAI Conference on Artificial Intelligence*, 2010.

[6] M. Garcia, A. Chatterjee, A. Ruina, and M. Coleman. The simplest walking model: Stability, complexity, and scaling. *ASME Journal of Biomechanical Engineering*, 120:281–288, 1998.

[7] A. Hamaguchi, M. Kanbara, and N. Yokoya. User localization using wearable electromagnetic tracker and orientation sensor. In *Proc. IEEE International Symposium on Wearable Computers*, pages 55–58, 2006.

[8] J. A. Hartigan and M. A. Wong. A k-means clustering algorithm. *Journal of the Royal Statistical Society*, 28:100–108, 1979.

[9] T. B. Moeslund and E. Granum. A survey of computer vision-based human motion capture. *Computer Vision and Image Understanding*, 81:231–268, 2001.

[10] S. Rosenthal and J. Pella. The process of motion capture: Dealing with the data. *Computer Animation and Simulation*, 97:3–18, 1997.

[11] M. Unuma, K. Anjyo, and R. Takeuchi. Fourier principles for emotion-based human figure animation. In *Proc. ACM SIGGRAPH*, pages 91–96, 1995.

[12] D. Vlasic, R. Adelsberger, G. Vannucci, J. Barnwell, M. Gross, W. Matusik, and J. Popovic. Practical motion capture in everyday surroundings. *ACM Transaction on Graphics*, 26:35, 2007.