International Journal of Innovative Computing, Information and Control
Volume **14**, Number **2**, April **2018**

 ISSN 1349-4198
pp. 747–755

SPATIO-TEMPORAL HUMAN MOTION ESTIMATION USING DYNAMIC CONDITIONAL RANDOM FIELD

Igi Ardiyanto

Department of Electrical Engineering and Information Technology
Universitas Gadjah Mada
Jl. Grafika No. 2, Yogyakarta 55281, Indonesia
igi@ugm.ac.id

Received June 2017; revised October 2017

Abstract. *We propose an approach for estimating human motion using spatio-temporal context. Unlike the widely used short-term prediction which employs a simple motion model, we try to establish a long-term human motion prediction by characterizing human trajectories tendency in a specific location and place. These trajectories are modeled by using a probabilistic temporal sequence model and parameterized by both person dynamics and location/topological features. Each topology is subsequently interconnected using a graph representation for acquiring a longer trajectory model. Predicted future path of the person is then generated by employing a particle filter-based predictor and integrating it with the trajectory models. Experimental results and evaluations on a real environment show benefit and feasibility of the proposed method.*
Keywords: Human motion prediction, Spatio-temporal context, Topological feature

1. **Introduction.** Understanding and perceiving human motion is really necessary for many robotic and computer vision applications. In robotics area, figuring out the human motion may assist the robot navigation in performing a better obstacle avoidance task, especially the dynamic motion of human. A correct prediction of the person motion enables the robot to produce a susceptible motion plan which handles possible future collisions. In another application, a computer vision researcher may utilize extracted information from the human motion to augment the surveillance camera capability involving the pattern of human movement.

Nevertheless, understanding the human movement is a complex problem. Many researchers try to simplify the problem by assuming that the human motion follows a simple model such as the *constant velocity model* (e.g., [1,2]). The weakness is that this simple model hardly maintains a correct prediction for a long time.

By realizing such weakness, several recent works recommend more sophisticated methods for exhibiting the human motion model. An effort to cluster and infer the human motion by using a *hidden Markov model* and *expectation-maximization* clustering is performed by [3]. Here they aimed to collect the pattern of the human trajectories. The similar method is also used in [4] by employing the hidden Markov model. In another work, [5] presented a novel probabilistic method so-called *joint probability distribution* for predicting the human motion patterns. In other research, [6] took a different perspective by engaging a class of method based on the *optimal control theory* to model a long-term destination forecasting of a moving person utilizing a semantic scene.

In more recent studies, Bera et al. [7] proposed the combination of global and local movement pattern for predicting the pedestrian paths. The drawback is that they made no assumption on the motion, of which it would affect the long term prediction. In contrast,

the work in [8] modeled the pedestrian intention using Markov decision process, but it takes a long time. This is also the problem which appears in [9,10]. Additionally, those works above do not take account of the influence of surrounding environment towards the human or pedestrian movement.

We are convinced that surrounding environment has a great influence on determining how a person will move. In particular, spatial and time context of the environment will drive and shape the human motion trajectories. As an easy instance, a person trajectory tends to follow the shape of a sidewalk pavement and walk on the side (rather than on the center) in a morning-rush time. Utilization of such information will benefit the future prediction of the human motion.

Unfortunately, most of the mentioned works do not take account of how the environment has an influence on the person movement (e.g., [1-3,11,12]). In case of [6], they take advantage of the physical attribute information of the environment (such as building, car, and pavement) for only separating the *walkable* and *non-walkable* areas.

We aim to close the above gaps by coalescing the spatio-temporal information of the environment for predicting the human motion. Here, the spatio-temporal context is described as the impact of features and attributes of the environment, including the structure and shape, towards the human trajectories which varies over the time. We subsequently propose a novel framework for tackling such problem. Human trajectory trends are initially extracted and a probabilistic sequence model is constructed, considering the person motion and spatio-temporal features of the environment. It is then integrated with a graph representation of the environment. The future path for the person is subsequently estimated using a particle filter. This paper also extends our previous work [13] which considers only the spatial relationship of the environment to the human motion. By exploiting the time into the estimation system, the spatio-temporal features can be constructed for obtaining more thorough motion prediction.

We organize the rest of this paper as follows. We first describe the human motion model as a sequence classification in Section 2. Environment representation as a graph and the usage of particle filter to help the human motion prediction are also discussed in the same section. The proposed approach is then verified on various experiments in Section 3. In the end, the conclusion and future directions of this work are provided.

2. **Modeling Human Motion Trajectory.**

Human trajectory model is established by first perceiving the human motion as a goal-oriented trajectory. In a real world, it can be explained by an easy instance as follows. An indoor environment can be semantically categorized into hallways or corridors and junctions. From a stationary perspective (e.g., the person is observed from a static observer), the human movement on the hallway can be easily predicted either it is getting close or going away. In another case, when we observe a T-junction, the other person movement will be going to left, going to right, or approaching to us. Here we intend to emulate such reasoning which infers the human movement as a goal-oriented motion.

2.1. **Notation.**

Trajectory of human motion is formally defined as $\mathcal{X} = \{x_1, x_2, \ldots, x_t\}$ which is a sequence of the human position or state until the time t. Let $\mathcal{Y} = \{Y_1, Y_2, \ldots, Y_n\}$ be n possible trajectories which leads a moving person towards each *observable goal* in $\mathcal{G}$. Let $\phi_{\mathcal{X}}$ and $\phi_{\mathcal{Y}}$ respectively represent observation of the human position and the goals from the current observer pose. We will explain *observable goal* term, later.

Let also $\mathbb{Q} \subset \mathbb{R}^2$ be two dimensional grid map obtained by a SLAM algorithm [14], the map $\mathbb{Q}$ is then simplified, such that $f_{simp} : \mathbb{Q} \mapsto \{\mathcal{P}, \mathcal{K}\}$. We borrow mapping function f_{simp} using procedure mentioned in [15] and extract a polygonal model $\mathcal{P}$, as well as the

skeleton $\mathcal{K}$ of the map. All junctions can be determined from $\mathcal{K}$ by employing a template matching over the map using junction models [15].

2.2. **Concept of observable goal for predicting human motion.** When a person/observer (or robot) observes motion of the other person, his "outlook" is basically limited by *field-of-view*. Suppose the observer is at a location near junction, his view will be restricted by walls and *line-of-sight*. Here, *line-of-sight* is defined as a non-obstacle area at which a person can pass through. Since the person cannot move through the walls, it is safe to assume that a person will make a motion following the shape of environment (i.e., corridors and junctions). Accordingly, human motion is basically coming from and to those *line-of-sight*. We then assume that *line-of-sight* becomes the possible *observable goal* for a person to move.

Field-of-view of the observer is not only spatially-limited, but also distance-limited. It then creates an observable space called *visibility polygon* which contains all of *line-of-sight*. Therefore, the observable goal can be determined by intersecting the visibility $\mathcal{V}$ and the skeleton $\mathcal{K}$. We subsequently define $\mathcal{G} = \{g_1, g_2, \ldots, g_n\}$ as the observable goal the person may lead to, as follows

$$\mathcal{G} = \{\forall q \in \mathbb{Q} | q = \mathcal{V} \cap \mathcal{K}\}, \tag{1}$$

where $\mathbb{Q}$ is the grid map obtained by SLAM algorithm, as mentioned in the notation section.

What if the environment is not in a form of narrow corridors and junctions? In this case, *field-of-view* of the observer becomes very broad and it is difficult to determine *line-of-sight*. Nevertheless, human trajectory is finite and spatially-biased. For instance, in a wide gallery, a person tends to move from an entrance, to a painting, to the other paintings, and then go to the exit. Each place subsequently becomes the possible *observable goal* for a person to move.

Utilizing the above concept of observable goal, our objective now becomes modeling the relationship between person trajectory, observable goals, the predicted motion towards the goals, and the observations as $p(\mathcal{X}, \mathcal{G}, \mathcal{Y} | \phi_{\mathcal{X}}, \phi_{\mathcal{G}})$ respectively. The model can be written under the independence assumption, as

$$p\left(\mathcal{X}, \mathcal{G}, \mathcal{Y} | \phi_{\mathcal{X}}, \phi_{\mathcal{G}}\right) = p\left(\mathcal{Y} | \mathcal{Y}, \mathcal{G}\right) p\left(\mathcal{X}, \mathcal{G} | \phi_{\mathcal{X}}, \phi_{\mathcal{G}}\right). \tag{2}$$

In principal, the first term of the right-hand side of Equation (2) is the trajectory prediction towards observable goal involving a sequence structure. It is naturally solved by a sequence classifier. The second term models the target person and observable goals. Here, we use a *Gaussian distribution*.

2.3. **Human motion as spatio-temporal sequence classification.** Human trajectory model consists of sequence of the human state or pose. Hence, it can be treated as a sequence prediction. We exploit these structures by employing dynamic conditional random field (DCRF) [16] to capture the spatial and temporal relationship between adjacent human pose constructing the trajectory. Adopting the work of [17], we model our DCRF as follows

$$p\left(\mathcal{Y} | \mathcal{X}, \mathcal{G}, \varphi\right) \propto \frac{1}{Z(\mathcal{X}, \mathcal{G}; \varphi)} \prod_t \exp\left\{ \sum_k \varphi_k f_k\left(\mathcal{Y}_t, \mathcal{X}, \mathcal{G}, t\right) \right\}, \tag{3}$$

where $f_k(\cdot)$ denotes a set of feature functions, φ_k is the parameter in the form of a set of weights to be estimated, and $Z(\cdot)$ represents the normalization factor.

Parameter φ_k is subsequently optimized using *pseudo-likelihood* (as also used by [17]), as follows

$$\mathcal{L}(\varphi) = \sum_{i=1}^{N} \log p\left(y_i \in \mathcal{Y} | x_i \in \mathcal{X}\right), \tag{4}$$

where N is the number of training data.

For the trajectory classification, the maximum score of predicted $\mathcal{Y}$ is taken, such that

$$y^* = \arg\max_{\mathcal{Y}} p(\mathcal{Y}|\mathcal{X}, \mathcal{G}; \varphi^*), \tag{5}$$

where φ^* represents the learned parameter.

2.4. **Feature function.** We exploit feature function similar to the one used in [13] for capturing the human trajectory properties. The employed features are composed by pose and topological features.

The pose features are depicted by the human coordinate at the time t, $x_t = \{x_t^1, x_t^2\} \in \mathcal{X} \subset \mathbb{R}^2$ as well as its derivation $\dot{x}$, and its motion orientation θ, as follows

$$\begin{aligned} \dot{x_t} &= \frac{(x_t - x_{t-1})}{\Delta t}, \\ \theta_t &= \tan^{-1}\left(\frac{x_t^2 - x_{t-1}^2}{x_t^1 - x_{t-1}^1}\right). \end{aligned} \tag{6}$$

Both velocity and orientation are respectively quantized into three bins and 16 bins histogram.

For the topological features, the objective of utilizing environmental topology is to comprehend the effect of environment structure to the human motion. Accordingly, a skeleton map $\mathcal{K}$ (as mentioned in the notation section above) is utilized since it has ability to capture shape and type of the environment (e.g., corridors, T-junctions, cross-junctions, and L-turn). It can be achieved by deriving the distance function towards the skeleton $\mathcal{K}$ for each element $x_i \in \mathcal{X}$, as follows

$$r(x_i) = \frac{\partial\left(e^{\|x_i - x_{\mathcal{K}}\|}\right)}{\partial x}, \tag{7}$$

where the numerator denotes the distance of x_i to the nearest point $x_{\mathcal{K}}$ in the skeleton. A high magnitude of $r(x_i)$ is expected to be obtained when a person cuts across the skeleton. This topological feature is subsequently quantized into eight bins histogram.

2.5. **Graph representation for the environment.** We have explained how to predict the human motion trajectory in a single place (e.g., junctions and corridors). When an observer moves, the place is dynamically and continuously changed. It means using only one spatial place is unfeasible. Fortunately, an indoor environment is able to be represented as a graph consisting of interconnected corridor and junction nodes, as shown in Figure 1.

One unique feature of this graph representation is that it looks like the human motion trajectory in a single place. The only difference is that it is now in a coarse and larger scale. It also means the similar technique and algorithm which handles sequence data can be utilized.

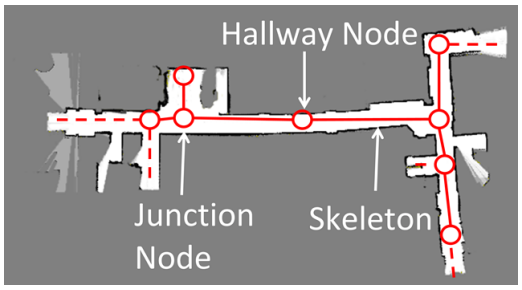

FIGURE 1. Graph representation of the map

For accommodating those structured graph representations, we employ a hierarchical trajectory classification by modifying Equation (3) as follows

$$\begin{aligned} p\left(\mathcal{Y}|\mathcal{X},\mathcal{G},\varphi\right) &\propto \frac{1}{Z(\mathcal{X},\mathcal{G};\varphi)}\prod_{t}\exp(f(\cdot)) \\ f(\cdot) &\propto \sum_{k_1}\varphi_1 f_1\left(\mathcal{Y}_t,\mathcal{X},\mathcal{G},t\right)+\sum_{k_2}\varphi_2 f_2\left(\mathcal{Y}_t,\mathcal{X},\mathcal{G},t\right). \end{aligned} \tag{8}$$

Here, $f_1(\cdot)$ represents feature functions taken from dense state of the human trajectory, while $f_2(\cdot)$ denotes coarser feature functions brought from the environment graph as shown in Figure 1 (of course, $k_2 < k_1$). Using such approaches, we expect to merge and take advantage of the information from the human motion state and the graph structure of the environment together, so that the human intention to move can be better foreseen.

2.6. **Particle filter-based estimator.** The distribution in Equation (2) is iteratively predicted using a Bayesian framework as the observation $\phi_{\mathcal{X}}$ and $\phi_{\mathcal{G}}$ are updated through the time. A particle filter framework is particularly utilized to do the job. We compose the state model by $\mathcal{S}=\{\mathcal{X},\mathcal{G},\mathcal{Y}\}$. The dynamical model is then described as

$$p\left(\mathcal{S}_t|\mathcal{S}_{t-1}\right)=p\left(\mathcal{X}_t,\dot{\mathcal{X}}_t|\mathcal{X}_{t-1},\dot{\mathcal{X}}_{t-1}\right)p\left(\mathcal{G}_t|\mathcal{G}_{t-1}\right)p\left(\mathcal{Y}_t|\mathcal{Y}_{t-1}\right). \tag{9}$$

The first term of the right-hand side of Equation (9) is modeled using a *first-order dynamical model*, and the rest is in a form of the *Gaussian distribution*.

The observation is subsequently modeled as

$$p\left(\phi_{\mathcal{X}},\phi_{\mathcal{G}}|\mathcal{X},\mathcal{G}\right)=p\left(\phi_{\mathcal{X}}|\mathcal{X}\right)p\left(\phi_{\mathcal{G}}|\mathcal{G}\right). \tag{10}$$

Same as the above, the *Gaussian distribution* model is utilized for the right-hand side of Equation (10). It is worth noting that the decision of choosing the observable goal can be determined when the confidence is above a threshold.

3. **Experiments.** All implementations of the described algorithm were done on a Windows PC (i5 2.4 GHz, 4 GB RAM) using C++ programming language.

3.1. **Dataset preparation.** We employ the same data used in [13]. A set of person trajectories is initially collected using a laser-based person tracker [18] on five different locations at our campus. These procedures capture 983 trajectory sequences in total and produce three to six trajectory classes per location. We subsequently divide the data into two different sets randomly for each location, i.e., for training and testing purposes.

We also append additional data taken from the different environment, depicting an outdoor scene at Universitas Gadjah Mada (hereby, *UGM Outdoor*). A set of person inside the scene is detected and tracked using image-based person tracker [19,20]. We manage to capture 125 trajectory sequences in total, with three classes.

3.2. **Evaluation.** The proposed method exploiting the dynamic conditional random field (DCRF) [16] is evaluated to discriminate person trajectory on each designated location. Since the problem has nature of sequence classification, it can be compared with other sequence classifiers as baseline, such as conditional random field (CRF) [21], hidden CRF (HCRF) [22], and hidden Markov model (HMM) [23].

CRF, HCRF, and DCRF are accordingly trained as a multi-class classifier for each location. There is difference between CRF-DCRF and HCRF on labeling the trajectory which is based on the observable goal. For CRF and DCRF, each state in the trajectory sequence needs to be labeled separately, while HCRF needs only one label for the whole states in the trajectory sequence. The HMM is subsequently treated as a generative model. Two different types of feature usage are engaged on each method: positional information only, and combined positional-topological features.

Accuracy of the trajectory classification is shown by Table 1. It is clear that considering the spatial context, as used in HCRF and DCRF, will increase the trajectory class recognition rate. Additionally, spatio-temporal features boost up the result over the other methods. One reasonable explanation is that the HCRF has ability to model the hidden structures of the trajectory sequence and its relationship toward one single trajectory

TABLE 1. Comparison of trajectory classification of dataset [13]

Method	Accuracy (%) on Location				
	1	2	3	4	5
CRF (pose)	52.66	73.45	46.45	43.83	42.64
HMM (pose)	58.90	77.25	54.67	50.24	51.30
HCRF (pose)	55.83	77.76	51.23	46.87	46.28
DCRF (pose)	57.45	77.34	51.02	48.68	49.52
CRF (pose + topology)	52.90	74.23	49.00	46.47	44.64
HMM (pose + topology)	60.67	75.23	58.96	52.45	48.20
HCRF (pose + topology)	63.34	80.45	61.62	57.90	53.44
DCRF (pose + topology)	65.76	79.88	63.71	59.15	53.06

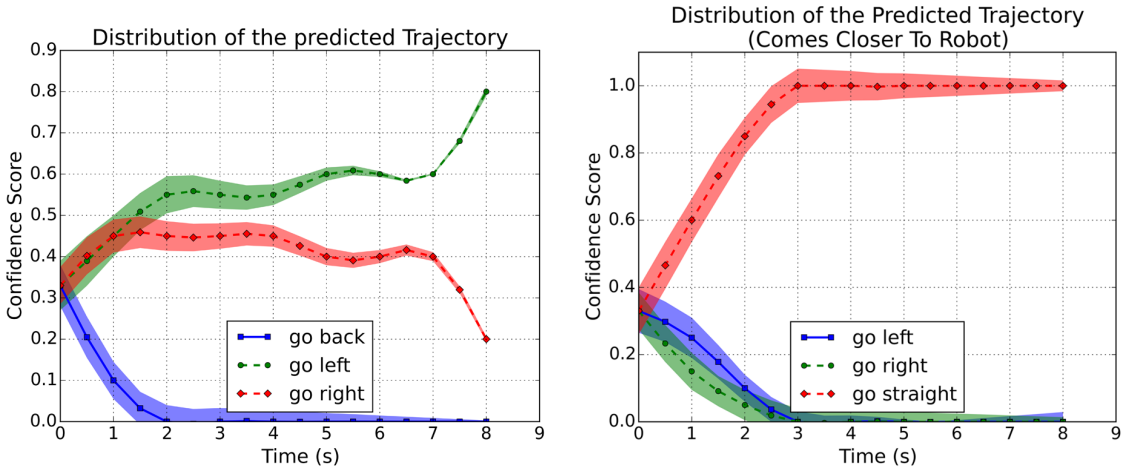

FIGURE 2. Distribution of the predicted trajectory over the time for location 1 of the dataset [13]

label, which is lack in the CRF and HMM, while the DCRF even further models the relationship over temporal states to achieve a better result.

Human motion prediction performance is qualitatively satisfying as it has been endorsed by tendency graph on Figure 2, which shows the distribution of each predicted trajectory over time. Please note that the trajectory is converged to the left side as the time grows.

3.3. **Predicting the human motion on outdoor environment.** We also conduct experiments on outdoor scene, i.e., "UGM Outdoor" which represents a 3-crossing paving road (see Figure 3). It means we have three classes of the trajectory. From geometrical point of view, this 3-crossing paving road has similar spatial form as the three-junction of the indoor setting used in the previous experiment. From Table 2, the HMM represents the generative sequential model, while the CRF's family denotes the discriminative models. In HMM, CRF and HCRF, we only care about the position and spatial information of the environment. Contrarily, DCRF considers both spatial and temporal properties of the environment. Once again, from Table 2, the usage of spatio-temporal context gives benefit for correctly predicting the human motion, with a higher accuracy.

FIGURE 3. Human motion prediction on an outdoor scheme

TABLE 2. Comparison of trajectory classification of dataset *UGM Outdoor*

Method	Accuracy (%)
CRF (pose)	55.7
HMM (pose)	55.9
HCRF (pose)	58.2
DCRF (pose)	60.5
CRF (pose + topology)	60.7
HMM (pose + topology)	62.9
HCRF (pose + topology)	66.3
DCRF (pose + topology)	66.8

4. **Conclusions.** We have presented a novel approach to predict the human motion by considering spatio-temporal context of the environment. Human trajectory tendencies are extracted by using a probabilistic sequence model which considers spatio-temporal context on each environment structure. Afterwards, each prediction on each environment structure is integrated with a graph representation. Lastly, a particle filter-based predictor is incorporated with the model to predict the human motion intention. Experimental results support the advantage of our method over other state-of-the-art approaches.

Some possible future direction of this research will be to eliminate all limitations we have mentioned in the previous section. Basically, our approach is also applicable on any structured environment which can be represented as graph, yet it needs further verification. A richer feature choice to extract the environment context also seems interesting to be evaluated later.

REFERENCES

[1] J. Cui, H. Zha, H. Zhao and R. Shibasaki, Tracking multiple people using laser and vision, *Proc. of the IEEE/RSJ Int. Conf. on Intelligent Robots and Systems (IROS)*, pp.1301-1306, 2005.

[2] A. Fod, A. Howard and M. Mataric, Laser-based people tracking, *Proc. of IEEE International Conference on Robotics and Automation (ICRA)*, pp.3024-3029, 2002.

[3] M. Bennewitz, W. Burgard, G. Cielniak and S. Thrun, Learning motion patterns of people for compliant robot motion, *Int. Journal of Robotics Research*, vol.24, no.1, pp.31-48, 2005.

[4] J. Z. BolaBola, Y. Wang, S. Wu, H. Qin and J. Niu, Application of hidden Markov model in human motion recognition by using motion capture data, *Advances in Physical Ergonomics and Human Factors*, pp.21-28, 2016.

[5] D. Vasquez, Novel planning-based algorithms for human motion prediction, *Proc. of IEEE International Conference on Robotics and Automation (ICRA)*, pp.3317-3322, 2016.

[6] K. Kitani, B. Ziebart, J. Bagnell and M. Hebert, Activity forecasting, *Proc. of European Conference on Computer Vision – ECCV*, pp.201-214, 2012.

[7] A. Bera, S. Kim, T. Randhavane, S. Pratapa and D. Manocha, GLMP – Realtime pedestrian path prediction using global and local movement patterns, *Proc. of IEEE International Conference on Robotics and Automation (ICRA)*, pp.5528-5535, 2016.

[8] V. Karasev, A. Ayvaci, B. Heisele and S. Soatto, Intent-aware long-term prediction of pedestrian motion, *Proc. of IEEE International Conference on Robotics and Automation (ICRA)*, pp.2543-2549, 2016.

[9] N. Brouwer, H. Kloeden and C. Stiller, Comparison and evaluation of pedestrian motion models for vehicle safety systems, *IEEE the 19th International Conference on Intelligent Transportation Systems (ITSC)*, pp.2207-2212, 2016.

[10] J.-Y. Kwak, B. C. Ko and J.-Y. Nam, Pedestrian intention prediction based on dynamic fuzzy automata for vehicle driving at nighttime, *Infrared Physics and Technology*, vol.81, pp.41-51, 2017.

[11] D. Vasquez, T. Fraichard and C. Laugier, Incremental learning of statistical motion patterns with growing hidden Markov models, *IEEE Trans. Intelligent Transportation Systems*, vol.10, no.3, pp.403-416, 2009.
[12] M. Luber, J. Stork, G. Tipaldi and K. Arras, People tracking with human motion predictions from social forces, *Proc. of IEEE International Conference on Robotics and Automation (ICRA)*, Anchorage, AK, USA, pp.464-469, 2010.
[13] I. Ardiyanto and J. Miura, Human motion prediction considering environmental context, *Proc. of the 14th IAPR Conf. on Machine Vision and Application*, Tokyo, Japan, pp.390-393, 2015.
[14] S. Thrun, W. Burgard and D. Fox, *Probabilistic Robotics (Intelligent Robotics and Autonomous Agents)*, The MIT Press, 2005.
[15] I. Ardiyanto and J. Miura, Visibility-based viewpoint planning for guard robot using skeletonization and geodesic motion model, *Proc. of IEEE International Conference on Robotics and Automation (ICRA)*, Karlsruhe, Germany, pp.652-658, 2013.
[16] C. A. Sutton, A. McCallum and K. Rohanimanesh, Dynamic conditional random fields: Factorized probabilistic models for labeling and segmenting sequence data, *Journal of Machine Learning Research*, vol.8, pp.693-723, 2007.
[17] J. Yin, D. H. Hu and Q. Yang, Spatio-temporal event detection using dynamic conditional random fields, *Proc. of the 21st International Joint Conference on Artificial Intelligence (IJCAI)*, pp.1321-1326, 2009.
[18] K. Koide, I. Ardiyanto and J. Miura, Person detection and tracking using camera and laser range finder for attendant robots, *Proc. of System Integration (SI)*, 2013.
[19] I. Ardiyanto and J. Miura, Partial least squares-based human upper body orientation estimation with combined detection and tracking, *Image and Vision Computing*, vol.32, no.11, pp.904-915, 2014.
[20] D. E. Pratiwi and A. Harjoko, Face recognition using principal component analysis, *Indonesian Journal of Electronics and Instrumentation Systems*, vol.3, no.2, pp.175-184, 2013.
[21] J. Lafferty, A. McCallum and F. Pereira, Conditional random fields: Probabilistic models for segmenting and labeling sequence data, *Proc. of the 18th Int. Conf. on Machine Learning*, pp.282-289, 2001.
[22] A. Quattoni, S. Wang, L. Morency, M. Collins and T. Darrell, Hidden conditional random fields, *IEEE Trans. Pattern Analysis and Machine Intelligence*, vol.29, no.10, pp.1848-1853, 2007.
[23] A. Sand, C. Pedersen, T. Mailund and A. Brask, HMMlib: A C++ library for general hidden Markov models exploiting modern CPUs, *Proc. of the 2nd Int. Workshop on High Performance Computational Systems Biology*, pp.126-134, 2010.