Abstract—Multiple moving object tracking is a challenging problem in Robotic Space. To handle this, particle filter and SOM are used to track the system state based on information from multiple cameras. The random moving object is tracked based on the integrated information from distributed network cameras. In order to achieve these goals, we present a method for representing, tracking and human following by fusing distributed multiple vision systems in Robotic Space, with application to pedestrian tracking in a crowd. And the article presents the integration of color distributions into SOM(Self Organizing Map) based particle filtering. Particle filters provide a robust tracking framework under ambiguity conditions. We propose to track the moving objects by generating hypotheses not in the image plane but on the top-view reconstruction of the scene. Comparative results on real video sequences show the advantage of our method for multi-motion tracking. Simulations are carried out to evaluate the proposed performance. Also, the method is applied to the intelligent environment and its performance is verified by the experiments.

I. INTRODUCTIONS

Detection of moving objects like as pedestrian, vehicles, etc, has been utilized in industrial robotic systems, for example, in the recognition and monitoring of unmanned systems that also require compression of moving images [1],[2],[3],[4]. Trajectory prediction of moving objects is required for a mobile manipulator that aims at the control and observation of motion information such as object position, velocity, and acceleration. Prediction and estimation algorithms have generally been required for industrial robots. For a simple example, in a pick-and-place operation with a manipulator, the precise motion estimation of the object on the conveyor belt is a critical factor in stable grasping. A well-structured environment, such as the moving-jig that carries the object on the conveyor belt and stops when the manipulator grasps the object, might obviate the motion estimation requirement. However, a well-structured environment limits the flexibility of the production system, requires skillful designers for the jig, and incurs a high maintenance expense; eventually it will disappear from automated production lines.

To overcome these problems, to tracking a moving object stably without stopping the motion, the trajectory prediction of the moving object on the conveyor belt is necessary [5]. The manipulator control system needs to estimate the most accurate position, velocity, and acceleration at any instance to capture the moving object safely without collision and to pick up the object stably without slippage. When the motion trajectory is not highly random and continuous, it can be modeled analytically to predict the near-future values based on previously measured data [6]. However, this kind of approach requires significant computational time for high-degrees-of-freedom motion, and its computational complexity increases rapidly when there are many modeling errors. In addition, performance is highly sensitive to the change of the environment. Those state-of-the-art techniques perform efficiently to trace the movement of one or two moving objects but the operational efficiency decreases dramatically when tracking the movement of many moving objects because systems implementing multiple hypotheses and multiple targets suffer from a combinatorial explosion, rendering those approaches computationally very expensive for real-time object tracking [7].

![Fig. 1. Robotic Space structured as intelligent environment by distributed cameras](image)

It is necessary for the intelligent environment to acquire various information about humans and robots in the environment. When the environment does not know where humans and robots are respectively, the environment cannot give the enough service to the appropriate user as for the physical service especially. Therefore, it is considered that how to get the location information is the most necessary of all. The system with multiple color CCD cameras is utilized as one of the means to acquire the location information in an intelligent environment. It can achieve the human centered system because the environment acquires the location of human noncontactly and the equipment of the special devices isn’t required for human. Moreover, camera has the advantage in wide monitoring area. It also leads to acquisition of details about objects and the behavior recognition according to image

---

Taeseok Jin¹, and Hyun-Sik Kim²

¹Dept. of Mechatronics Engineering, DongSeo University, Busan 617-716, KOREA
²Dept. of Robot System Engineering, TongMyong University, Busan, 608-711, KOREA

Email: jints@dongseo.ac.kr

---

482
processing. Our intelligent environment is achieved by distributing small intelligent devices which don’t affect the present living environment greatly.

II. VISION SYSTEMS IN ROBOTIC SPACE

A. Previous Research for Tracking

Neural networks can be classified into two categories: supervised learning- and unsupervised learning methods. In most of the previous research, the supervised learning method was adopted to overcome the nonlinear properties [10],[12]. Since the supervised learning method requires the relation between input and output [9] at all times, it is not suitable for real-time trajectory estimation for which the input-output relation cannot be obtained instantaneously in the unstructured environment. Therefore, in this study, SOM (Self Organizing Map), that is, a type of unsupervised learning method, was selected to estimate the highly nonlinear trajectory that cannot be properly predicted by the Particle filter. Also, SOM is a data-sorting algorithm, which is necessary for real-time image processing since there is so much data to be processed. Among the most popular data-sorting algorithms, VQ (Vector Quantization), SOM, and LVQ (Learning Vector Quantization), SOM is selected to sort the data in this approach since it is capable of unsupervised learning. Since VQ is limited to the very special case of a zero neighborhood and LVQ requires preliminary information for classes, neither of them is suitable for the unsupervised learning of the moving trajectory. Fig. 2 shows the estimation and tracking system for this research. The input for the dynamic model comes from either the Particle filter or SOM according to the following decision equation:

\[
P_{\text{predicted}} = k \cdot \text{Particle Filter}_{\text{out}} + (1-k) \cdot \text{SOM}_{\text{out}}
\]  

where \( k = 1 \) for \( \text{error} \leq \text{threshold} \) and \( k = 0 \) for \( \text{error} > \text{threshold} \).

The threshold value is empirically determined based on the size of the estimated position error.

III. IMAGE DATA PROCESSING

A. Extraction of Walking Humans

Classifying the walking-human pattern in the dynamically changing unstructured environment has not yet been tackled successfully [13]. Therefore, in this research, the camera was fixed on a stable platform in order to capture static environment images. To estimate the states of the human motion characteristics, the trajectory of the walking-human was pre-recorded and analyzed. Fig. 3(a) and Fig. 3(b) represent the human images at (t-1) instance and (t) instance, respectively.

As recognized in the images, most parts of the CCD image correspond to the background. After eliminating the background, the difference between the two consecutive image frames can be obtained to estimate the walking-human motion. To compute the difference, either the absolute values of the two image frames or the assigned values can be used. The difference method is popular in image pre-processing for extracting desired information from the whole image frame, which can be expressed as

\[
\text{Output}(x, y) = I_{\text{mage1}}(x, y) - I_{\text{mage2}}(x, y)
\]

The difference image between Fig. 3(a) and Fig. 3(b) is represented in Fig. 4. When the difference image for the whole time interval can be obtained, the trajectory of the moving object can be calculated precisely.

B. Target Regions Encode

Particle filtering provides a robust tracking framework, as it models uncertainty. Particle filters are very flexible in that they not require any assumptions about the probability distributions of data. In order to track moving objects (e.g. pedestrians) in
video sequences, a classical particle filter continuously looks throughout the 2D-image space to determine which image regions belong to which moving objects (target regions). For that a moving region can be encoded in a state vector. In the tracking problem the object identity must be maintained throughout the video sequences. The image features used therefore can involve low-level or high-level approaches (such as the colored-based image features, a subspace image decomposition or appearance models) to build a state vector. A target region over the 2D-image space can be represented for instance as follows:
\[
r = \{l, s, m, \gamma\}
\]
(3)
where \(l\) is the location of the region, \(s\) is the region size, \(m\) is its motion and \(\gamma\) is its direction. In the standard formulation of the particle filter algorithm, the location \(l\), of the hypothesis, is fixed in the prediction stage using only the previous approximation of the state density. Moreover, the importance of using an adaptive-target model to tackle the problems such as the occlusions and large-scale changes has been largely recognized. For example, the update of the target model can be implemented by the equation
\[
\tilde{r} = (1 - \lambda)\tilde{r}_{-1} + \lambda E[r]
\]
(4)
where \(\lambda\) weights the contribution of the mean state to the target region. So, we update the target model during slowly changing image observations.

IV. TRACKING WALKING HUMANS

A. Top-view Plan

In a practical particle filter [5],[6] implementation, the prediction density is obtained by applying a dynamic model to the output of the previous time-step. This is appropriate when the hypothesis set approximation of the state density is accurate. But the random nature of the motion model induces some non-zero probability everywhere in state-space that the object is present at that point. The tracking error can be reduced by increasing the number of hypotheses (particles) with considerable influence on the computational complexity of the algorithm. However in the case of tracking pedestrians we propose to use the top-view information to refine the predictions and reduce the state-space, which permits an efficient discrete representation. In this top-view plan the displacements become Euclidean distances. The prediction can be defined according to the physical limitations of the pedestrians and their kinematics. In this paper we use a simpler dynamic model, where the actions of the pedestrians are modeled by incorporating internal (or personal) factors only. The displacements \(M'_{\text{topview}}\) follows the expression
\[
M'_{\text{topview}} = A(\gamma'_{\text{topview}})M_{\text{topview}}^{-1} + N
\]
(5)
where \(A(.)\) is the rotation matrix, \(\gamma'_{\text{topview}}\) is the rotation angle defined over top-view plan and follows a Gaussian function \(g(\gamma'_{\text{topview}}; \sigma_{\gamma})\), and \(N\) is a stochastic component. This model proposes an anisotropic propagation of \(M\) : the highest probability is obtained by preserving the same direction. The evolution of a sample set is calculated by propagating each sample according to the dynamic model. So, that procedure generates the hypotheses.

B. Estimation of Region Size

The size of the search region represents a critical point. In our case, we use the a-priori information about the target object (the pedestrian) to solve this tedious problem. We assume an averaged height of people equal to 160 cm, ignoring the error introduced by this approximation. That means, we can estimate the region size \(s\) of the hypothetical bounding box containing the region of interest \(r = \{l, s, m, \gamma\}\) by projecting the hypothetical positions from top-view plan in Fig. 5.

A camera calibration step is necessary to verify the hypotheses by projecting the bounding boxes. So this automatic scale selection is a useful tool to distinguish regions. In this way for each visual tracker we can perform a realistic partitioning (bounding boxes) with consequent reduction in the computational cost. The distortion model of the camera's lenses has not been incorporated in this article. Under this approach, the processing time is dependent on the region size.

C. Object Model Update

In multi-motion tracking, the hypotheses are verified at each time step by incorporating the new observations (images). A well known measure of association (strength) of the relationship between two images is the normalized correlation.
\[
d_{i,j} = corr_{\text{norm}}(t \text{ arg } e_{i,j} ; \text{ hypothesis}_{i,j})
\]
(6)
where \( i \) : target region, and \( j \) : an hypothesis of the target region \( i \). The observation of each hypothesis is weighted by a Gaussian function with variance \( \sigma \).

\[
\hat{h}^{(i,j)} = \frac{1}{\sqrt{2\pi \sigma_{\text{dc}}}^j} e^{-\frac{(1-d_{\text{dc}})^2}{2\sigma_{\text{dc}}}^j} \tag{7}
\]

where \( \hat{h}^{(i,j)} \) is the observation probability of the hypothesis \( j \) tracking the target \( i \). The obvious drawback of this technique is the choice of the region size (defined in previous section) that will have a great impact on the results. Larger region sizes are less plagued by noise effects.

V. CONCLUSIONS

In this paper, the proposed tracking method adds an adaptive appearance model based on color distributions to particle filtering. The color-based tracker can efficiently and successfully handle non-rigid and fast pedestrian under different appearance changes. Moreover, as multiple hypotheses are processed, objects can be tracked well in cases of occlusion or clutter. This research proposes estimation and tracking scheme for a moving object using images captured by multi cameras. In the scheme, the state estimator has two algorithms: the particle filter that estimates the states for the linear approximated region, and SOM for the nonlinear region. The decision for the switchover is made based on the size of the position estimation error that becomes low enough for the linear region and becomes large enough for the nonlinear region. The effectiveness and superiority of the proposed algorithm was verified through experimental data and comparison. The adaptability of the algorithm was also observed during the experiments. For the sake of simplicity, this research was limited to the environment of a fixed-camera view. However, this can be expanded to the moving camera environment, where the input data might suffer from higher noises and uncertainties. As future research, selection of a precise learning pattern for SOM in order to improve the estimation accuracy and the recognition ratio, and development of an illumination robust image processing algorithm, remain.

ACKNOWLEDGMENT

This Research was supported Basic Science Research Program through the National Research Foundation of Korea(NRF) funded by the Ministry of Education, Science and Technology (No. 2010-0021054).

REFERENCES