# Enhancing Probabilistic Appearance-Based Object Tracking with Depth Information: Object Tracking under Occlusion

Kourosh MESHGI[†] Yu-zhe LI[†] Shigeyuki OBA[†] Shin-ichi MAEDA[†] and Shin ISHII[†]

[†] Graduate School of Informatics, Kyoto University, Gokasho, Uji-shi, Kyoto, 611-0011 Japan

E-mail: † {meshgi-k,li-yuzhe,oba,ichi,ishii}@sys.i.kyoto-u.ac.jp

**Abstract** Object tracking has attracted recent attention because of high demands for its everyday-life applications. Handling occlusions especially in cluttered environments introduced new challenges to the tracking problem; identity loss, splitting/merging, shape changes, shadows and other appearance artifacts trouble appearance-based tracking techniques. Depth-maps provide necessary clues to retrieve occluded objects after they reappear, recombine split group of objects, compensate drastic appearance changes, and reduce the effect of appearance artifacts. In this study, we not only proposed a consistent way of integrating color and depth information in a particle filter framework to efficiently perform the tracking task, but also enhanced the previous color-based particle filtering to achieve trajectory independence and consistency with respect to the target scale. We also exploited local characteristics to represent the target objects and proposed a novel confidence measure for them. Applying to simple tracking problems, the performance of our method is discussed thoroughly.

**Keyword** RGB-D particle filter, depth map, explicit occlusion handling, enhanced bounding box

## 1 Introduction

Object tracking is of emerging demands for everyday life of people with various applications ranging from human-computer interfaces, to human behavior analysis, video communication/compression, virtual/augmented reality and surveillance. To approach the object tracking problem, there are two major strategies which are known as bottom-up and top-down methods [1]. In a bottom-up approach, objects are segmented out from a frame-wise image which is used for tracking. In contrary, a top-down method generates hypotheses and verifies them using the image; model-based [10, 11] and template matching approaches [12] are typical examples. Particle filter, which was used in this study, is one of top-down approaches, since it generates multiple hypotheses so that image features are evaluated based on those hypotheses.

Employing probabilistic framework, in this paper a particle filter is detailed which handles object tracking even under persistent occlusions, is highly adaptive to object scale and trajectory, and perform a color and depth information fusion. The algorithm use rectangular bounding boxes as hypotheses of target presence. The target, described by a color histogram and median of depth, was compared to each bounding box in terms of the Bhattacharyya distance. In order to increase the accuracy of target localization, and granting scale adaptation to system, each bounding box was divided into regular grids, for each confidence measure was calculated as a ratio of foreground pixels to all pixels in the grid. A novelty of the proposed method is to handle long-lasting occlusions explicitly using an occlusion flag attached to each particle which signals if the bounding box is occluded, and serves to allow a search for the target in an increasing area in effect and to suppress the update of the target template.

Object tracking algorithms can be categorized according to the type and configuration of cameras used [7], and it can be 2D based on monocular cameras, 3D in the case of stereo cameras or multiple cameras, and 2.5D on Microsoft Kinect (which combines 2D image with depth map). Trackers using 2D views often rely on appearance models, where employed models have one-to-one correspondence to objects in the image [6]. These trackers suffer from occlusions and fail to handle object interactions since modeling all possible object interactions is intractable. On the other hand, 2.5D or 3D approaches are more robust to occlusions but are prone to major tracking issues.

There are a plenty of literature and a wealth of tools for enabling object tracking from 2D images taken by a single camera (a famous survey could be found in [13]). They include model-based, appearance-based, and feature-based methods [7]. Although tracking separated targets has become a popular competition scenario, multiple object tracking still remains a challenging task due to dynamic change of object attributes, such as color distribution, shape and visibility [7].

New generation of trackers, the ones that track multiple objects could be categorized into generative and discriminative methods. Generative models keep the status of each object represented by a probability distribution function. Existing studies in this line used particle filtering [15], Monte Carlo-based methods [9] and Bayesian networks with HMM [17]. If real-time computation is mandatory, generative methods trades the number of evaluated solutions and the granularity of each solution. The general trend toward compensating the computational charge of having multiple hypotheses is to use compact appearance models e.g. color histograms or color distribution [15, 16]. Color distributions, as a well-defined feature, play a crucial role in many of the successful generative tracking algorithms. Mean shift tracker [12] uses color distributions and employs multiple hypotheses and a model of the system dynamics. Another trend to use color histograms is to model them as Gaussian mixture models in combination with multi-hypothesis adaptive models [15].

Being widely used in generative schemes, particle filtering is a sophisticate technique that applies a recursive Bayesian filter based on sample set [10, 11]. The prominent advantage of this scheme is its applicability to nonlinear and non-Gaussian systems [1]. Among various versions of particle filtering, this study is based on Condensation algorithm [10, 11] which was developed initially to track objects in cluttered environments. Particle filtering is robust as it benefits from multiple hypotheses. In the case of short-time occlusion, the probability of object states decreases but particles still remain in the tracking process. Even if the occlusion continues, we suppress track loss by allowing particles to scatter gradually to look for the target object all over the frame.

Typically, generative models do not address occlusion explicitly, but since they maintain a large set of hypotheses some of them survive and are recovered after the occlusion. In contrary discriminative models generally deal directly with problem of occlusion detection [6]. Although these methods are robust against partial and temporary occlusions, long-lasting occlusions hinder their tracking heavily.

Yet other approaches are employing stereo and/or multiple cameras or data fusion with range sensors. Some tracking methods focus on usage of depth information only [20], while others use depth information for better foreground segmentation [21], or employ depth information to statistically estimate 3D object positions [22]. As depth information provides valuable information about z-order of the object, it can be used to facilitate tracking and handle occlusions.

Occlusions can be classified into three classes: (1) dynamic occlusion, in which there are pixels close to camera; (2) scene occlusion, in which still objects are closer to camera than the target object; and (3) apparent occlusion, a result of shape change, silhouette motion, shadows, or self-occlusions [8]. When devising an update model for target, it is important to consider which type of occlusion we deal with. If occlusions are of (1) or (2), the update of object model should be performed slowly to keep memory, but in the case of (3), a fast update is preferred to keep focus on the target.

Tracking in hybrid domain of color and depth needs to maintain a balance between information obtained from each. Difficulties a tracker faces include appearance changes (illumination changes, shadows, affine transformations, non-rigid body deformations, and occlusion), parameters of the sensors and their compatibility (field of view, position, resolution, and signal-to-noise ratio) and segmentation inherent problems (partial segmentation split and merge). The design choices we made in this system, tackles most of these challenges as will be explained later in this manuscript. Following this introduction, next we describe our proposed method, and compare it with adaptive color based particle filter tracker [1] in single and multiple target cases along with short and long occlusions. The manuscript then concludes with discussion and future works.

## 2 Proposed model

### 2.1 Overview of the proposed model

Proposed model, a version of particle filter, is one of state-space models which consist of the observation model and state-transition model. The state specifies a bounding box for each particle along with the correspondent occlusion flags. Then the observation model describes how the bounding box is observed by RGB-camera and depth sensor while the state-transition model describes how the bounding box with occlusion flag transits along with time.

### 2.2 Target Representation

A rectangular bounding box is used as hypothesis of target presence. The bounding box at time $t$ is specified by four parameters $B_t = \{x_t, y_t, w_t, h_t\}$ where $(x_t, y_t)$ denotes

2D-coordinate of the top-left corner of the bounding box, $w_t$ and $h_t$ denote the width and height of the bounding box, respectively. Furthermore, the bounding box is divided into several grid cells (in the experiment below, the bounding box is partitioned into four equal parts) to fit the local statistics and represent the partial occlusion. Then a binary occlusion flag $z_{i,t} \in \{0,1\}$ is introduced for each cell; $z_{i,t} = 0$ represents *i*-th cell is not occluded while $z_{i,t} = 1$ represents *i*-th cell is occluded. $Z_t = \{z_{i,t} \mid i = 1, \cdots, C\}$ denotes the set of all the occlusion indicators in the bounding box. As mentioned above, we set $C = 4$ in our experiment. Since both $B_t$ and $Z_t$ is not observed directly, they are treated as the hidden state and estimated according to the state-space model described below.

### 2.3 Observation

We employed a fixed camera in this paper and the stationary background could be assumed. Having this assumption, the background could be extracted using several frames of the video, so that the median value of each pixel in several frames is considered to be the background, both in RGB and depth domains (Fig 1). This method is a modified version of algorithm proposed in [3] which first uses a variance filter to balance the brightness variation, and then a temporal median background update technique is used for obtaining accurate reference background.

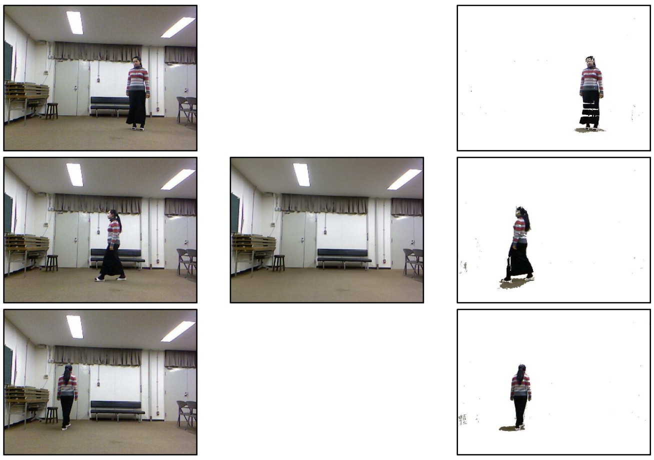

**Fig 1. Left:** Several frames of the video **Center:** Their extracted background. **Right:** Background subtracted image and d-map (with same threshold).

The depth information is assumed to be collected by cheap range sensors like Kinect. The relation between raw depth values and metric depth has been experimentally determined to be a hyperbolic relationship, described in [2]. Another artifact of such sensors is their sensitivity to IR-absorbing material, especially in long distances. Figure 2 illustrates these two effects. In this step, the out-of-range values are clipped to valid range, and hyperbolic correction was applied to remove sensor systematic error, and the result is up-sampled to match the RGB resolution using linear interpolation. Since the way we use the depth data is robust to noise, outliers and noise due to the high absorbing material, is not treated in our approach.

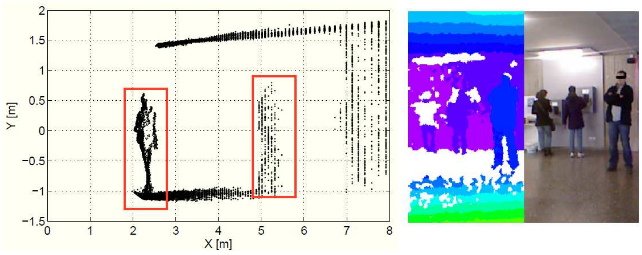

**Fig. 2. Left:** Effects of hyperbolic resolution loss. Side view of two example persons at different ranges from the sensor. Close subjects are accurately described in high detail. Farther away, quantization is becoming increasingly dominant and heavily compromises the shape information on people. Geometric approaches to people detection will perform poorly in such data. **Right:** Example frame to illustrate that IR-absorbing surfaces at large distances lead to blobs of missing depth data (upper body of leftmost subject, white means missing data). [5]

### 2.4 Observation model

The preprocessed foreground RGB image at time $t$ is denoted as $Y_{rgb,t}$ and depth image at time $t$ is stated as $Y_{depth,t}$ . Assumed to be independent in this paper, these two components from the observation $Y_t = \{Y_{rgb,t}, Y_{depth,t}\}$ .The observation model for $Y_t$ given by $B_t$ and occlusion flags $Z_t$ is elaborated as follows.

$$\begin{aligned} & p(Y_t \mid B_t, Z_t, \boldsymbol{\theta}_t) = \prod\nolimits_i p(Y_t \mid B_t, i, z_{i,t}, \theta_{i,t}) \\ & p(Y_t \mid B_t, i, z_{i,t} = 1, \theta_{i,t}) = const \\ & p(Y_t \mid B_t, i, z_{i,t} = 0, \theta_{i,t}) \\ \propto\ & p(\# Y_{i,t} \mid B_t, \theta_{i,t}) p_{rgb}(Hist(Y_{rgb,i.t}) \mid B_t, \theta_{i,t}) p_d(\overline{Y}_{depth,i,t} \mid B_t, \theta_{i,t}) \end{aligned} \tag{1}$$

where $i$ is a grid index, $\# Y_{i,t}$ , $Hist(Y_{rgb,i,t})$ and $\overline{Y}_{depth,i,t}$ denote the ratio of foreground pixels, histogram of $Y_{rgb,t}$ and the median value of $Y_{depth,t}$ in the $i$-th grid cell of the bounding box $B_t$ , respectively. Finally $\boldsymbol{\theta}_t = \{\theta_{i,t} \mid i = 1, \cdots, C\}$ is a set of adaptive parameters containing typical RGB histogram and typical depth which are explained later.

We employed beta distribution for $p(\# Y_{i,t} \mid B_t)$ which is defined over the domain of [0, 1] parameterized by two positive shape parameters, $a_i$ and $b_i$, that control the shape of the distribution.

$$p(\# Y_{i,t} \mid B_t) \propto \left(\# Y_{i,t}\right)^{a_i - 1} \left(1 - \# Y_{i,t}\right)^{b_i - 1} \tag{2}$$

We extract sufficient samples of target bounding boxes and fit the beta distribution to foreground ratio separately on each cell (fig 3). This distribution works as the confidence of each grid cell since it tends to take high value when the grid cell is dominated by the foreground pixels while it tends to take low value, or zero when the cell does not contain the significant number of the foreground pixels.

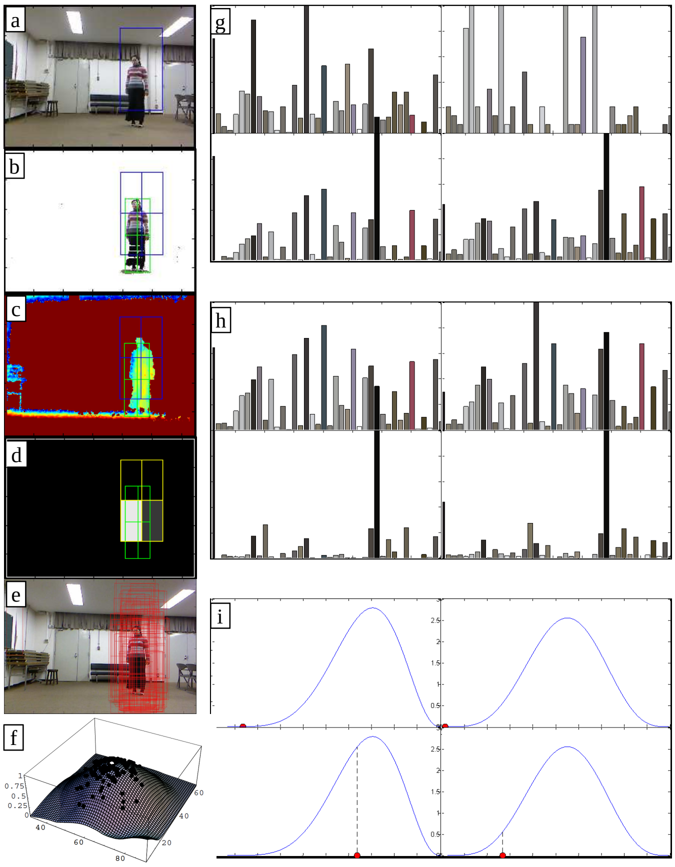

**Fig 3. (a)** A 2×2 particle bounding box vs. target bounding box (green); **(b)**The gridding of bounding box, on the background subtracted image; **(c)** Corresponding depth map; **(d)** Confidence map; brighter cells are more reliable cells to rely on. **(e)** Particles around target. **(f)** Surface plot of the Bhattacharyya coefficient of a small area around the target. The black points illustrate the centers of some of the particles while the white point represents the mean location, close to the peak of the plot [1]; **(g)** The histogram of colors of the illustrated sample bounding box and **(h)** target; **(i)** Beta distributions fit to a 2×2 grid bounding box to track the target and the confidence calculate for grid cells of left image.

The likelihood of the color histogram is defined as follows

$$p_{rgb}(Hist(Y_{rgb,i.t}) \mid B_t, \theta_{i,t}) \propto \exp\left(d_{rgb}\left(Hist(Y_{rgb,i.t}), q_{i,t}\right)\right) \qquad (4)$$

where $d_{rgb}\left(Hist(Y_{rgb,i.t}), q_{i,t}\right)$ denotes the Bhattacharyya distance. The Bhattacharyya distance of two discrete distribution $p$ and $q$ with $m$ bins (number of color bins) weighted with control parameter $\sigma_{rgb}$ is defined as follows:

$$d_{rgb}(p,q) = \frac{1}{\sigma_{rgb}}\sqrt{1-\sum_{i=1}^{m}\sqrt{p_i q_i}} \qquad (5)$$

Finally $p_d(\overline{Y}_{depth,i,t} \mid B_t, \theta_{i,t}) = N(\mu_{i,t}, \sigma_{depth})$ is assumed to be a Gaussian distribution whose mean and variance are $\mu_{i,t}$ and $\sigma_{depth}$ , respectively. Since the resolution and signal-to-noise ratio of the depth information are assumed to be low, only the median value is evaluated for the likelihood to make the likelihood robust.

Note that our observation model for each grid cell does not depend on the size, shape and rotation of the target, implying our grid cell model is scale, shape and rotation invariant.

## 2.5 State-Transition Model

Since the occluder does not affect how the target changes with time assuming there is no direct interference such as collision between targets or target and environment, independence between the state-transitions of bounding box and occlusion flag is imposed.

$$p(B_{t+1}, Z_{t+1} \mid B_t, Z_t) = p(Z_{t+1} \mid Z_t)\, p(B_{t+1} \mid B_t) \qquad (6)$$

Bounding box transition probability $p(B_{t+1} \mid B_t)$ is represented by Gaussian distribution whose center is specified by $B_t$ and the covariance matrix $\Sigma$ is diagonal matrix assuming the independence between four parameters $B_t = \{x_t, y_t, w_t, h_t\}$ . On the other hand, occlusion flag transition probability $p(Z_{t+1} \mid Z_t)$ is a discrete distribution. Since $Z_t$ has $2^C$ distinct states, the distribution is represented by $2^C \times 2^C$ matrix $T$ . Representing by full matrix, state transition of the occlusion flag $Z_t$ can take into account of the spatial and temporal continuity of the occluder.

## 3 Inference and learning of the proposed model

### 3.1 Inference by particle filter

Since our model is complex and non-Gaussian, particle filter is used for the filtering. Among various versions of particle filter, this research is based on Condensation algorithm [10, 11] which was developed initially to track objects in clutter. Modeling uncertainty, particle filter very is robust as it benefits from multiple hypotheses.

### 3.2 Learning

Parameters to be set are $\{a_{i,t}, b_{i,t}, q_{i,t}, \mu_{i,t} \mid i = 1,\cdots,C\}$ , $\sigma_{rgb}$ , $\sigma_{depth}$ , $\Sigma$ , and $T$ . Among them, $\{a_{i,t}, b_{i,t} \mid i = 1,\cdots,C\}$ are learned from several training sequences as mentioned earlier while $\sigma_{rgb}$ , $\sigma_{depth}$ , $\Sigma$ , and $T$ are settled by hand and fixed through the whole experiment. $\boldsymbol{\theta}_t = \{q_{i,t}, \mu_{i,t} \mid i = 1, \cdots, C\}$ are adaptively learned to follow the temporal and spatial change of the target. However, we have to carefully update them because the observation does not include meaningful information when there is the occlusion. We resolve this problem by utilizing the occlusion flag.

Having $N$ particles from which several of them have the occlusion flag set ( $z = 1$ ), we can vote between particle if the estimated target using expectation of all bounding boxes is expected to be occluded or not. Since the vote of these particles should have different effect on the result proportional to their probability, we simply take the expectation of all particle flags to validate the voting.

$$\theta_{i,t+1} = \begin{cases} \lambda\theta_{i,t} + (1-\lambda)\overline{\theta}_{i,t} & (z_{i,t} = 0) \\ \theta_{i,t} & (z_{i,t} = 1) \end{cases} \qquad (7)$$

in which $\overline{\theta}_{i,t}$ denotes the $Hist(Y_{rgb,i,t})$ for $q_{i,t}$ and $\overline{Y}_{depth,i,t}$ for $\mu_{i,t}$ . $\lambda$ is a forgetting factor, which balances the adaptivity and robustness.

## 4 Experiments

In order to evaluate the performance of our algorithm we prepared a toy dataset in laboratory environment containing two scenarios. In the first one a single person is walking, mostly in parallel with camera z-plane and in some parts towards the camera to test the tracking accuracy and scale adoptability of the tracker. The appearance of the subject changed drastically in several frames, and several rapid changes in direction of movement and velocity as well were observed, while the depth information of those frames remains intact to test the robustness of algorithm against changes of appearance, direction, and velocity of movement. (fig. 5). In the second scenario, the same video is used while a rectangular space of the data is occluded manually. The data is acquired with Microsoft Kinect, with image resolution of 640×480 and depth image resolution of 320×240. The dataset is also provided with target bounding box coordinates and occlusion status (states are: *none*, *partial*, *full*) as ground truth.

To evaluate the system, we applied following two criteria that are specially designed metrics to evaluate bounding boxes for multiple objects tracking, which were proposed for a classification of events, activities and relationships (CLEAR) project, and called CLEAR criteria [23]. The multiple object tracker precision(MOTP) shows the ability of the tracker to estimate precise object positions, independent of its skill at recognizing object configurations, keeping consistent trajectories, etc. Additionally the multiple object tracker accuracy (MOTA) criterion cares for such high level objectives:

$$MOTP = \frac{\sum_{i,t} d_t^i}{\sum_t c_t} \qquad (8)$$

$$MOTA = 1 - \frac{\sum_t \left( m_t + fp_t + mme_t \right)}{\sum_t g_t} \qquad (9)$$

where $c_t$ is the indicator for tracker match the correct target, $d_t^i$ is the error in estimated position for matched object-hypothesis pairs, $g_t$ is number of objects present in frame *t*, $m_t$ , $fp_t$ , and $mme_t$ are the number of misses, of false positives, and of mismatches, respectively, for time *t*.

Furthermore, we introduce a measure for scale adaptation, and since the tracker is not constrained to maintaining the aspect ratio of the object, it has two components:

$$SA = \frac{\sum_t \sum_{i \in c_t} \sqrt{(\Delta w_{i,t})^2 + (\Delta h_{i,t})^2}}{\sum_t c_t} \qquad (10)$$

in which $\Delta w_t$ and $\Delta h_t$ are the difference of the estimated width and height of estimated target with ground truth at time *t* and the measure will be sum of distance of estimated dimensions of bounding box to true value for all successfully matched objects. Lower values of SA indicates better adaptation of algorithm to scale.

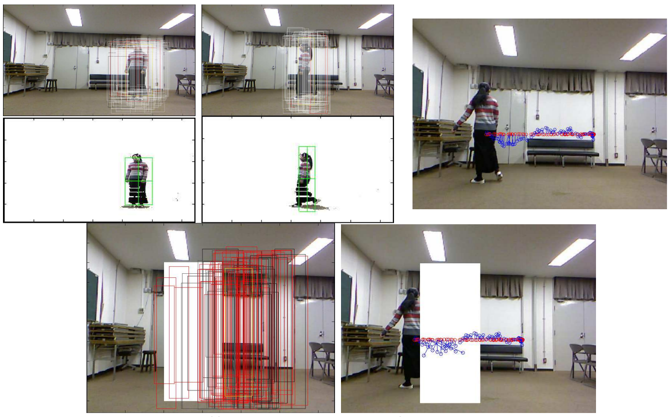

**Fig 5. Top:** Two different screenshots of the walking person tracking dataset with internal state of tracker. Each bounding box represents a particle. Red boxes indicate particles with occlusion flag of one while gray-scale boxes are related to not occluded particles. In the latter case, brighter color shows more probable bounding box. Green box indicates the estimated target. The rightmost figure is the tracking result of objects over time (blue) and ground truth (red) **Bottom:** The same video with a toy occlusion attached. The particles are scattered to search a wider area of scene, while their occlusion flag indicates that they are occluded. The

expected target box, reasonably experienced some fluctuation during this period.

In the experiments the performance of RGB particle filter tracker proposed in [1], our proposed algorithm without enabling the bounding box gridding option and with gridding option enabled (2×2 grid), and last-mentioned algorithm plus occlusion flag is compared in regard to three criteria of MOTA, MOTP, and SA defined earlier. Note that all of these algorithms are particle filters run via same platform with the same number of particles. The results of the walking person scenario and walking person scenario with occlusion are shown in tables 1 and 2 respectively.

**Table 1:** Comparison of Algorithms performance for 1-person tracking scenario. The scenario contains changes in movement velocity, trajectory, appearance and scale.

| Tracker | MOTP | MOTA | SA |
|---|---|---|---|
| RGB | 87.2 | 97.1% | 112.8 |
| RGB-D | 38.1 | 100% | 99.9 |
| RGB-D Grid 2×2 | 23.6 | 100% | 48.5 |
| RGB-D Grid 2×2+ Occlusion Flag | 24.1 | 100% | 51.2 |

As it can be inferred from second row of table 1, the precision of tracking increases with adding depth information, as this channel is invariant to appearance changes and mitigate the low recognition ability of color domain tracker. Third row illustrates the importance of constraining size of bounding box using information obtain by local regions. This local information is propagated to the decision via combining gridding strategy and confidence measure. Fourth row suggests that although there is no occlusion in this scenario, it should perform well.

**Table 2:** Comparison of Algorithms performance for 1-person tracking scenario. The scenario is similar to one in Table 1, but there is a toy occlusion in the middle of the subject trajectory.

| Tracker | MOTP | MOTA | SA |
|---|---|---|---|
| RGB | 153.1 | 57.2% | 98.8 |
| RGB-D | 93.2 | 59.1% | 91.9 |
| RGB-D Grid 2×2 | 73.1 | 46.1% | 59.2 |
| RGB-D Grid 2×2+ Occlusion Flag | 53.1 | 83.3% | 67.5 |

Table 2 supports the conclusions drawn from table 1. By the way the second row shows the role of depth information in finding the subject quickly after it reappears. The fourth row illustrates the effectiveness of proposed occlusion handling method, as it is deduced from MOTA.

## 5 Conclusion & Future Works

In this paper we described a probabilistic framework for tracking objects in hybrid space of color and depth, in which we devised the ideas of gridding bounding box and occlusion flag. The gridding bounding box was devised for better representation of local statistics of foreground image and occlusion than that of simple bounding boxes. The occlusion flag for each box was for distinguishing the occluded and un-occluded cases explicitly, which suppressed the template and extended the search space under the occlusion. In addition, we introduced a confidence measure that evaluates ratio of fore- and background pixels in a box in order to track such box that has appropriate value of the ratio. Also we utilized the depth information effectively to judge who occludes the others.

Giving flexibility to the bounding box size, our method prone to involve partial occlusion artifacts such as splitting and merging. This scheme is equipped with bounding box as representation, in which histogram of colors and median of depth is extracted as feature, to compare the similarity to target features. We devised gridding bounding box for better representation of the local statistics and local occlusion and employed Bhattacharya distance to compare distance of two color distributions in each grid. In addition we introduced a confidence measure to focus on the foreground image. Strictly speaking, it makes the tracking system to focus on the box where an appropriate ratio of foreground and background pixels is realized. Furthermore the occlusion is handled in our framework with an occlusion flag for each particle which distinguish the occluded case and un-occluded case explicitly. It enables the explicit suppression of the template and effective extension of the search space under the occlusion. Also we utilized the depth information effectively to judge who occludes the others.

Using a bounding box to represent objects, and in this case walking people, and giving flexibility to their size, could have put our method prone to partial occlusion artifacts such as splitting and merging artifacts. But since there are multiple hypotheses for target in each frame, which stochastically search around expected target location and scale, there is no need to explicitly handle them. Clearly this secondary outcome is an advantage over cases which special care for splitting and merging problem is required. For example in the multi-object tracking system [7], level set scheme is utilized to handle contour splitting and merging or in another case, authors of [6] merged the adjacent boxes having the same velocity, while the probability mask is periodically analyzed to check the presence of two or more well-separated connected components

Another design choice we made was giving flexibility to box dimension, such that they vary around the dimensions of expected target, with a freedom degree parameterized by variance of the Gaussian distribution they are sampled from. This freedom, aligned with gridding and confidence measure grants scale adaptation for the tracker. Other methods either fails to show satisfying scale adaptation such as mean-shift tracker [12], or are unable to recover after sudden changes in scale or scale change during occlusion such as [1].

To enhance this algorithm, it is possible to make use of occlusion information of each grid to handle different combinations of occlusion patterns in state transition matrix. It is also possible to let the instances of the tracker interact with each other: exchanging information about depth, more accurate state transition, etc.

Our experiments shows that exploiting information in depth channel helps resolving loss of track for abrupt appearance changes, and increase the robustness of the method. Additionally decomposing bounding box into regular grid improves scale adaptation of the algorithm, preventing the size of bounding box to bewilder around the optimal value. Finally by adding an explicit stochastic occlusion handling mechanism, the algorithm could handle longer occlusion times without losing track of object because of small search region, or corrupted template by irrelevant data. The algorithm overtook its appearance based ancestors regarding localization accuracy, scale adaptation, and occlusion handling.

## References

[1] K. Nummiaro, E. Koller-Meier, and L. Van Gool, “An Adaptive Color-based Particle Filter.” in Image and Vision Computing 21.1, 99-110, 2003.

[2] M. Quigley, K. Conley, B. P. Gerkey, J. Faust, T. Foote, J. Leibs, R. Wheeler, and A. Y. Ng, “ROS: an Open-Source Robot Operating System,” in ICRA Workshop Open Source Soft., Vol. 3, 2009.

[3] Lo, B.P.L. and S.A. Velastin, “Automatic Congestion Detection System for Underground Platforms” in proceedings of Int’l Symp. on Intell. Multimedia, Video and Speech Processing, 158-161, 2001.

[4] P. Pérez, C. Hue, J. Vermaak, and M. Gangnet. “Color-based Probabilistic Tracking,” in ECCV 2002, 661-675, 2002.

[5] L. Spinello and K.O. Arras. “People Detection in RGB-D Data.” in IROS, 2011 IEEE/RSJ Int’l Conf. on, 3838-3843. IEEE, 2011.

[6] R. Cucchiara, C. Grana, G. Tardini, and R. Vezzani. “Probabilistic People Tracking for Occlusion Handling.” in Proceedings of ICPR’04, 2004.

[7] A.T. Tran and K. Harada “Depth Assisted Tracking Multiple Moving Objects under Occlusion,” in IJCSNS 13, no. 5, 49, 2013.

[8] R. Vezzani, C. Grana, and R. Cucchiara. “Probabilistic People Tracking with Appearance Models and Occlusion Classification: The AD-HOC System,” in Pattern Recog. Letters 32, 867-877, 2011.

[9] T. Zhao and R. Nevatia, “Tracking Multiple Humans in Complex Situations,” in IEEE Trans. PAMI 26 (9), 1208–1221, 2004.

[10] M. Isard and A. Blake, “Contour Tracking by Stochastic Propagation of Conditional Density,” in ECCV, 343-356, 1996.

[11] M. Isard and A. Blake, “Condensation: Conditional Density Propagation for Visual Tracking,” in Int’l J. on Computer Vision 1(29), 5-28, 1998.

[12] D. Comaniciu, V. Ramesh and P. Meer, “Real-Time Tracking of Non-Rigid Objects using Mean Shift,” in CVPR, 142-149, 2000.

[13] A. Yilmaz, X. Li, and M. Shah, “Contour-based Object Tracking with Occlusion Handling in Video acquired using Mobile Cameras,” in IEEE Trans. PAMI 26 (11), 1531–1536, 2004.

[14] A. Bhattacharyya, “On a Measure of Divergence between Two Statistical Populations defined by their Probability Distributions,” in Bulletin of the Calcutta Mathematical Society, vol. 35, pp. 99-109, 1943.

[15] M. Isard and J. MacCormick, “Bramble: A Bayesian Multiple-Blob Tracker,” in Proc. IEEE ICCV, pp. 34–41, 2001.

[16] O. Lanz, “Approximate Bayesian Multibody Tracking,” in IEEE Trans. PAMI. 28 (9), 1436–1449, 2006.

[17] B. Wu and R. Nevatia, “Tracking of Multiple, Partially Occluded Humans based on Static Body Part Detection,” in: Proc. IEEE Int’l. Conf. on CVPR, vol. 1, 951–958, 2006.

[18] S. McKenna, S. Jabri, Z. Duric, A. Rosenfeld, and H. Wechsler, “Tracking Groups of People,” in Computer Vision Image Understanding 80 (1), 42–56, 2000.

[19] A. Jepson, D. Fleet, and T. El-Maraghi, “Robust Online Appearance Models for Visual Tracking,” in IEEE Trans. PAMI 25 (10), 1296–1311, 2003.

[20] E. Parvizi and Q. M. J. Wu, “Multiple Object Tracking Based on Adaptive Depth Segmentation,” in Canadian Conf. on Comp. and Robot Vision, 2008 (CRV '08), 273-277, 2008.

[21] S. J. Krotosky and M. M. Trivedi, “On Color-, Infrared-, and Multimodal-Stereo Approaches to Pedestrian Detection,” in IEEE Trans, on Intell. Transportation Systems, vol. 8, 619-629, 2007.

[22] R. Okada, Y. Shirai, and J. Miura, “Object Tracking Based on Optical Flow and Depth," in Int’l Conf. on Multisensor Fusion and Integration for Intelligent Systems (IEEE/SICE/RSJ), 565-571, 1996.

[23] B. Keni, and R. Stiefelhagen, “Evaluating Multiple Object Tracking Performance: the CLEAR MOT Metrics,” in EURASIP J. on Image and Video Processing, 2008.