

Article

Content-Aware Focal Plane Selection and Proposals for Object Tracking on Plenoptic Image Sequences

Dae Hyun Bae ¹, Jae Woo Kim ² and Jae-Pil Heo ^{3,*}

¹ Department of Electrical and Computer Engineering, Sungkyunkwan University, 2066 Seobu-ro, Jangan-gu, Suwon-si, Gyeonggi-do 16419, Korea; noahark0628@gmail.com

² Electronics and Telecommunications Research Institute, 218 Gajeong-ro, Yuseong-gu, Daejeon 34129, Korea; jae_kim@etri.re.kr

³ Department of Software, Sungkyunkwan University, 2066 Seobu-ro, Jangan-gu, Suwon-si, Gyeonggi-do 16419, Korea

* Correspondence: jaepilheo@skku.edu; Tel.: +82-31-290-7965

Received: 1 November 2018; Accepted: 20 December 2018; Published: 22 December 2018



Abstract: Object tracking is a fundamental problem in computer vision since it is required in many practical applications including video-based surveillance and autonomous vehicles. One of the most challenging scenarios in the problem is when the target object is partially or even fully occluded by other objects. In such cases, most of existing trackers can fail in their task while the object is invisible. Recently, a few techniques have been proposed to tackle the occlusion problem by performing the tracking on plenoptic image sequences. Although they have shown promising results based on the refocusing capability of plenoptic images, there is still room for improvement. In this paper, we propose a novel focus index selection algorithm to identify an optimal focal plane where the tracking should be performed. To determine an optimal focus index, we use a focus measure to find maximally focused plane and a visual similarity to capture the plane where the target object is visible, and its appearance is distinguishably clear. We further use the selected focus index to generate proposals. Since the optimal focus index allows us to estimate the distance between the camera and the target object, we can more accurately guess the scale changes of the object in the image plane. Our proposal algorithm also takes the trajectory of the target object into account. We extensively evaluate our proposed techniques on three plenoptic image sequences by comparing them against the prior tracking methods specialized to the plenoptic image sequences. In experiments, our method provides higher accuracy and robustness over the prior art, and those results confirm that the merits of our proposed algorithms.

Keywords: plenoptic imaging technique; object tracking; bounding box proposal; content-based image matching

1. Introduction

Object tracking is one of the most active research topics in computer vision, since it is a core technique in various industrial applications such as video-based surveillance systems and autonomous vehicles. The goal of the object-tracking problem is to estimate the trajectory of target object in a sequence of images. Given an initial bounding box containing the target object at the beginning of the tracking task (i.e., frame index $t = 1$), the object tracker is asked to compute bounding boxes of the target object in subsequent frames (i.e., frame index $t > 1$). Although there are numerous trackers for multiple-object tracking, we restrict our discussion in this paper on the problem of single object tracking since our proposed techniques are orthogonal to such problem configurations.

There are various factors that make object tracking difficult such as the change of appearance of the object due to the rotation, translation, or illumination. Among those factors, one of the most

difficult cases is when the target object becomes partially or even fully occluded. The traditional tracking algorithm designed for 2D image sequences are highly likely to be missing the target object while it is being invisible. A few tracking techniques [1–3] have used the refocusing capability of plenoptic images to overcome the occlusion problem. Even though the object is occluded, changing the focus to the plane where the target object is maximally focused makes the target object visible while the occluders are being blurred out. Cho and Javidi [1] have reduced the occlusion problem using integral imaging technique to discover the occluded target object based on plane-by-plane reconstruction. Kim et al. [2] have presented a focus index selection method based on a focus measure [4] that identifies the maximally focused focal plane and perform the object tracking on the selected images. A follow-up work [3] has validated that applying image sharpening filters and a simple proposal technique can enhance the accuracy and robustness of [2]. Aforementioned methods have proposed solutions for the occlusion problem by using plenoptic imaging techniques and shown noticeable tracking results; however, none of them completely solved the problem. Specifically, focal plane selection algorithm in [2] can miss the optimal focal plane since their method is only relied on the focus measure, and the proposal technique of [3] does not fully use the advantage of plenoptic images.

In this paper, we propose a novel focal plane selection algorithm that takes both focus measure and visual similarity into account for the object tracking on plenoptic image sequences and design a bounding box proposal method tailored to our selected focus index. The main contributions of this paper are summarized as follows:

- We propose a novel focus index selection algorithm, *content-aware focal plane selection*, that identifies the maximally focused and visually consistent focal plane. While the prior techniques [2,3] have only used the focus measure, our selection method additionally consider whether the target object is clearly appeared in the image plane (Section 4).
- We propose a novel bounding box proposal scheme based on the optimal focus index computed by our content-aware focal plane selection algorithm. Since the focus index is directly related to the distance between the camera and object, we use the change of the focus index to accurately estimate the scale of the object in the image plane. We also design an algorithm to propose the location of candidate bounding boxes by modeling of the trajectory of the object. We generate several candidate bounding boxes based on the both scale and motion factors and select the best one among them according to the visual similarity to the target object (Section 5).

Once we identify a focus index and a bounding box proposal based on our proposed techniques, the remaining task is reduced to an ordinary object-tracking problem in 2D images. Since our method does not have any restriction on the choice of trackers, we can use any existing tracking techniques to estimate the location of object. In other words, we can take advantage of advanced tracking techniques. Figure 1 illustrates the process of the object tracking task in plenoptic image sequences. Please note that the plenoptic images can be captured by not only specialized equipment but also micro lens arrays which can be installed to portable commodity devices such as smartphones. Once the plenoptic images are acquired, they can be used in any computing devices including PCs, smartphones, and tablet.

We have extensively evaluated our proposed techniques on three different plenoptic image sequences. To show the flexibility of our method in terms of choice of trackers, we have applied three different tracking engines, Boosting [5], minimum output sum of squared error (MOSSE) [6], and MOSSE with convolutional neural networks (CNN) features [7] trackers to our proposed scheme. Experimental results show that our techniques provide both higher accuracy and robustness over the recent methods [2,3] (Section 6).

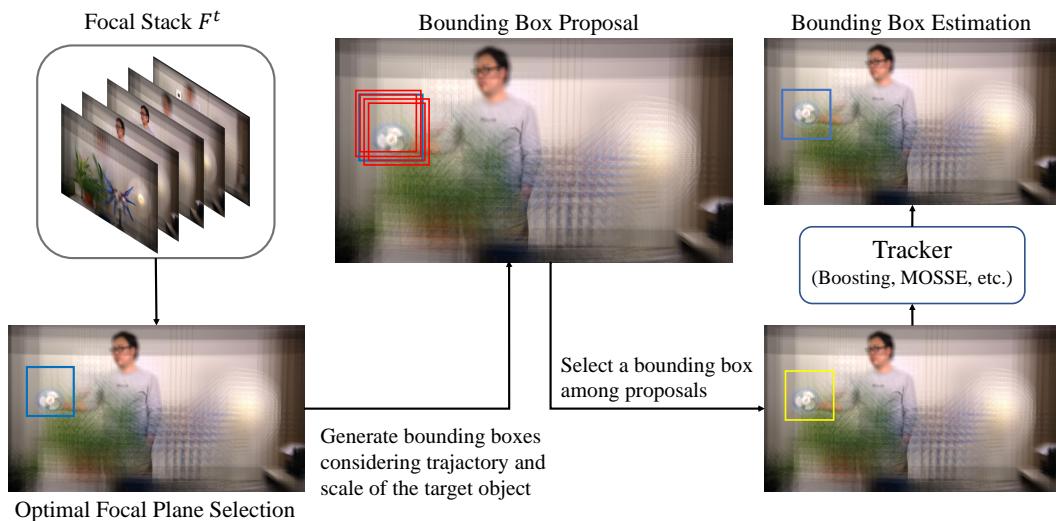


Figure 1. Procedure of our object-tracking techniques for plenoptic image sequences.

2. Related Work

Object tracking is one of the fundamental problems in computer vision. Due to its importance, there has been numerous techniques to solve the problem. Excellent surveys for those methods are available [8–10].

Among several challenges in the object tracking, one of the most difficult cases is dealing with the target object partially or even fully occluded. There have been many studies that have attempted to resolve such occlusion problem. Adam et al. [11] have proposed a representation of the target object based on multiple image fragments which is robust to the partial occlusions or appearance changes. A tracking method based on a sparse dictionary for modeling the appearance of the object is proposed in [12]. They constructed a confidence map from the sparse representation, and then located the object by voting and mean-shift. Zhang et al. [13] also reconstructed an appearance of object by sparse measurement matrix exploiting a non-adaptive random projection. Grabner et al. [14] employed the visual context into the object tracking task. Even if the target object becomes invisible, they can infer its location based on the motion of correlated objects in the scene. Babenko et al. [15] proposed multiple instance learning to train a classifier to learn an appearance of the object in robustly so that the tracker alleviates the problem of partial occlusion. A dynamic graph-based tracker has been proposed to address both deformation and occlusion problem in [16].

Tracking algorithms based on the correlation filters show remarkable results in recently object tracking challenge [17]. Such trackers learn the appearance of the target object. Bolme et al. [6] have proposed an object tracker based on a MOSSE filter. The MOSSE-based tracking algorithms compute a response map with the correlation score based on a correlation filter and an image patch. The scores can be computed by a simple convolution operation in Fourier Transform domain to achieve high efficiency in terms of computational costs. A new location of the target object is then identified by centering the peak among the correlation scores of the response map. Once the position of the target object is estimated, the correlation filter is incrementally updated with an image patch cropped from the location of the object.

There has been various work proposing improved algorithms based on the correlation filters by exploiting its simplicity and efficiency. Notable examples include a kernelized correlation filters [18], a filter trained from multiple dimensional features instead of a single channel of images [19], spatial regularization to resolve periodic boundary effects [20], and using CNN features to training correlation filters [7]. Lukezic et al. [21] uses spatial constraints to overcome the limitations regarding to the assumption of shape of the target object. Shi et al. [22] have proposed a sampling technique by considering the spatiotemporal consistency for training correlation filters. An adaptive method to

adaptively update the kernelized correlation filters according to the scene classification is proposed in [23]. Also, Xue et al. [24] attempted to improve tracking performance by reducing noises from background. They adopt the concept of cellular automata to update an appearance template.

In this paper, our contribution is more focused on handling plenoptic image sequences not concentrating on development of a new tracking algorithm working in 2D images. Our technique is designed to accept any of existing tracking methods. We, however, mostly tested our method by combining with MOSSE tracker and its variations to take advantage of their simplicity and efficiency.

On the other hand, the refocusing capability of plenoptic imaging technology inheritably enables to reduce the occlusion problem in the object tracking task. Cho et al. [1] have proposed a tracking method in the integrated image sequences. A plenoptic image sequence is consisting of focal stacks, and each focal stack is corresponding to each frame index and containing a set of images reconstructed in a plane-by-plane manner. Thus, it is crucial to select an appropriate focal plane where the target object is maximally focused. By selecting a proper focus index, the object can become visible even it is blocked by occluders. Figure 2 shows examples of plenoptic images.

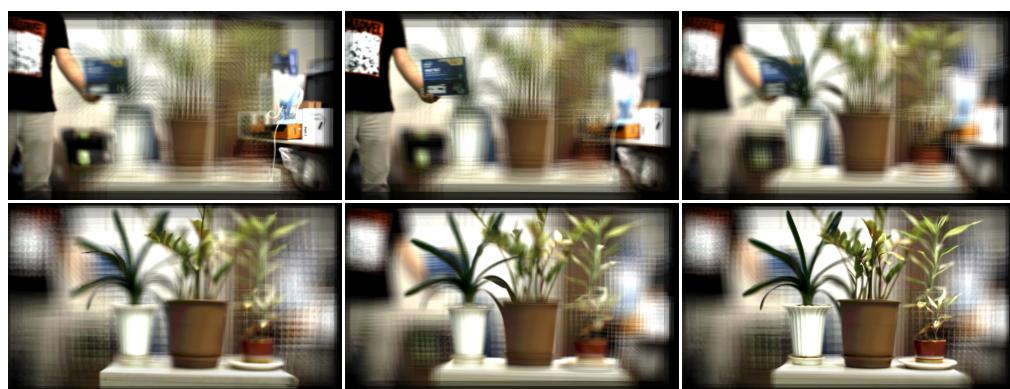


Figure 2. Example of plenoptic images.

Kim et al. [2] shows promising results by using the focus measure [4,25] to select focal planes and performing tracking on those image planes. The technique [2] is extended in Bae et al. [3] that proposed two additional steps. They performed an image sharpening process prior to estimate the location of the object, since the selected images determined by the focus measure still contain blurry reasons due to the discreteness of focal planes. They also presented a heuristic to generate proposals based on a fixed ratio from the previous frame. We found that, both techniques [2,3] can identify an undesired focus index since their focal plane selection is strongly relied on the focus measure. If there is an occluder that can have higher focus measure score than the target object, their algorithms may select the focus index where the occluder is maximally focused instead of the target object. To overcome such problem, we propose a novel focus index selection technique that considers both the focus measure and visual consistency.

Recently deep CNNs have shown state-of-the-art performance in several computer vision tasks including object classification [26–30] and detection [31–33]. The network trained for solving image classification problem can produce generic features of images, the pre-trained CNN models can be employed as image feature extractor. For instance, in the object detection problem region proposal-based architectures [31,32] use VGG [27] to extract spatial and semantic information from input images. Inspired from the success, deep features extracted from CNNs are actively used in the object tracking. Wang et al. [34] have proposed an online tracker by learning generic features of natural images based on an autoencoder trained offline. In [35], a CNN that learns relationships between the appearances and motions of objects is proposed to predict the position of a particular object in the next frame. Danelljan et al. [7] have used deep CNN features to train correlation filters instead of using single gray scale image. They specifically used activations of a convolutional layer which are multiple channels data. They have chosen the first activations of convolutional layer as CNN feature

to preserve spatial information. The proposed method achieved remarkable results on three public benchmark datasets [8,17,36]. On the other hand, Babenko et al. [37,38] have shown that the features extracted from CNNs are powerful image descriptors in image matching or retrieval tasks. In this paper, we validate our proposed technique that can take advantage of the success of deep learning in various computer vision tasks. We specifically combine our method with the MOSSE tracker trained by CNN features and use the features to select an appropriate focal plane based on the visual consistency. More details are discussed in Sections 4 and 5.

3. Problem Formulation and Overview

3.1. Problem Formulation

Let us first define notations that we will use throughout this paper. A plenoptic image sequence \mathbb{I} consists of N frames of focal stacks, and a focal stack at the frame index t is then denoted by F^t . The focal stack F^t of the frame t is defined as a set of images, $F^t = \{I_1^t, I_2^t, \dots, I_M^t\}$, where M is the number of focal planes in the stack. As a result, the plenoptic image sequence \mathbb{I} of N time frames and M focal planes has NM images, $\{I_k^t\}$ where $t \in \{1, 2, \dots, N\}$ and $k \in \{1, 2, \dots, M\}$.

The object tracking on 2D image sequence is a well-defined problem that estimates a trajectory of an object over the time. Specifically, the estimated locations of the target object at the frame index t is generally described by a bounding box B^t centered at (x^t, y^t) whose width and height are w^t and h^t , respectively. Given an initial bounding box B^1 of the target object, the object tracking is to find a sequence of bounding boxes B^2, \dots, B^N containing the target object over the frames.

Object-tracking problems on a plenoptic image sequence and a simple 2D image sequence share the same goal locating the target object over frames. However, the problem on a plenoptic sequence is more complex than an ordinary image sequence because it requires additional consideration to identify appropriate focal planes in the focal stacks. In conventional object-tracking problem, it is assumed that the target object is well-focused, and its appearance is fairly clear in 2D images. On the other hand, in the plenoptic case the region of a target object can be blurry or even occluded if we select an inappropriate focal plane. Thus, in each frame t we need to first carefully identify an optimal focal plane where the object region is visible and clearly presented prior to estimate the location of the object. The visibility and clarity are directly related to the success-failure rate and accuracy of the tracking result.

3.2. Overview

To estimate the trajectory of a target object in the plenoptic image sequence, we first select an optimal focal plane from a focal stack F^t at each frame t . We propose a novel algorithm to identify an optimal focal plane based on both the focus measure [25] and visual feature similarity (Section 4). Compared to the prior attempts for the optimal image selection based on the focus measure [2], our proposed technique exploiting visual feature similarity provides higher robustness and accuracy. Once an optimal image is identified, we generate several candidate bounding boxes based on two factors, the history of the object trajectory and distance between the object and camera. We then select the best one among the candidates according to the image patch similarity to the tracking result at the previous frame (Section 5). Finally, the selected bounding box is passed to the object-tracking module designed for conventional 2D image sequences to obtain a refined tracking result.

4. Focus Index Selection

As discussed in Section 3.1, identifying an optimal focal plane from a focal stack is crucial in the object tracking on plenoptic image sequences since the visibility and clarity of appearance of the target object are highly correlated with the success-failure rate and accuracy of the tracking task. Moreover, a correctly selected focal plane reduces the difficulty in the tracking when the target object

is partially or even fully occluded by other environmental objects. In this section, we introduce our novel algorithm to select an optimal image from a given focal stack.

4.1. Focus Measure Based Focus Index Selection

To select an optimal image from the focal stack, one reasonable approach is using the focus measure. In this strategy an optimal image is defined to be the maximally focused at the region where the target object is assumed to be located. Specifically, given a region of interest (RoI) defined as a set of pixels $\Omega(x, y)$ where the target object highly likely to be located (e.g., the bounding box B^{t-1} estimated in the previous frame $t - 1$), the optimal image I_k^t is selected among a set of images $\{I_1^t, \dots, I_M^t\}$ in the focal stack F^t to have the highest focus measure on $\Omega(x, y)$. Pertuz et al. [4] have conducted extensive evaluation of many successful focus measures and reported that Sum-Modified-Laplacian (SML) operator [25] provided high performance and was robust to various factors such as the quality and content of the image. The focus measure based on SML operator for an image I and RoI $\Omega(x, y)$ is defined as follows:

$$\phi(I, \Omega(x, y)) = \sum_{(i,j) \in \Omega(x, y)} \Delta_m I(i, j), \quad (1)$$

where $\Delta_m I(i, j)$ is a modified Laplacian with two convolution filters $\mathcal{L}_X = [-1 \ 2 \ -1]$ and $\mathcal{L}_Y = \mathcal{L}_X^T$ as expressed as follows ($*$ is the convolution operator):

$$\Delta_m I(i, j) = |I * \mathcal{L}_X| + |I * \mathcal{L}_Y|. \quad (2)$$

Let us define an image cropping function $c(I, B)$ that extracts an image patch corresponding to a bounding box $B(x, y, w, h)$ from the image I . For instance, $c(I, (15, 30, 10, 20))$ produces a rectangular image patch from I whose coordinates of left upper and right lower are (10, 20) and (20, 40), respectively. We then can define the *focus score*, f-score(I_k^t), of an image I_k^t in the focal stack F^t by rewriting the focus measure (Equation (1)) by using the cropping function $c(\cdot, \cdot)$ as follows:

$$\text{f-score}(I_k^t) = \sum_{(i,j) \in c(I_k^t, B^{t-1})} \Delta_m I_k^t(i, j) \quad (3)$$

The focus index selected according to the focus score can provide a reasonably accurate focal plane as shown in [2,3]. Using only the focus score to select an index, however, can produce an undesired focus index as the optimum in cases where the target object is occluded by environmental objects which is maximally focused. In [2], a hill-climbing technique is proposed to resolve the issue. The heuristic starts from the focus index \hat{k}_{t-1} of the previous frame $t - 1$, and searching for the focus index having a local peak of focus score by moving to higher scores. Since the algorithm prevents significant changes in the focus index by starting from the focus index of the previous frame, it could reduce the aforementioned problem. However, it still suffers from such incorrect index selection when there is the occluder close to the target object but has higher focus score than the target.

4.2. Content-Based Score for Focus Index Selection

The aforementioned issue discussed in Section 4.1 is mainly because the focus index selection only based on the focus measure (Equation (1)) which is totally unaware of the content in the RoI. When evaluating the optimality of given focus index, it is also crucial factor whether the target object appears or not. We, hence, propose to use visual similarities in the focus index selection to capture the content in the RoI.

We define an image feature extractor $f(I)$ that produces a high-dimensional feature representation describing the content of a given image I . Notable examples of such feature extractor $f(I)$ include the Bag of Visual Word (BoVW) [39] and Neural Codes [37].

Given a focal stack F^t at the frame index t , we evaluate images I_k^t from the focal stack by examining image patches $c(I_k^t, B^{t-1})$ corresponding to the estimated bounding box B^{t-1} at the previous frame. We define a *content score*, c-score(I_k^t) of the k th image in the focal stack at frame t to be the average similarity of $c(I_k^t, B^{t-1})$ to the initial bounding box at the first frame and the estimated bounding box at the previous frame $t - 1$, as follows:

$$\text{c-score}(I_k^t) = \frac{1}{2} \left\{ \text{f-sim}\left(f\left(c(I_{\hat{k}_1}^1, B^1)\right), f\left(c(I_k^t, B^{t-1})\right)\right) + \text{f-sim}\left(f\left(c(I_{\hat{k}_{t-1}}^{t-1}, B^{t-1})\right), f\left(c(I_k^t, B^{t-1})\right)\right) \right\}, \quad (4)$$

where $\text{f-sim}(f_x, f_y)$ defines the similarity of the features f_x and f_y . Please note that our content score can be simply accepting any feature similarity measures including Euclidean distance, cosine similarity, and even more complex metrics such as Mahalanobis distances. This implies that we can take advantage of any advanced image feature representation since our method is orthogonal to the features or their similarity. For instance, our content score becomes the followings if the feature similarity is defined by dot products (i.e., cosine similarity):

$$\begin{aligned} \text{c-score}(I_k^t) &= \frac{1}{2} \left(f\left(c(I_{\hat{k}_1}^1, B^1)\right) \cdot f\left(c(I_k^t, B^{t-1})\right) + f\left(c(I_{\hat{k}_{t-1}}^{t-1}, B^{t-1})\right) \cdot f\left(c(I_k^t, B^{t-1})\right) \right) \\ &= \frac{f(c(I_k^t, B^{t-1}))}{2} \cdot \left(f\left(c(I_{\hat{k}_1}^1, B^1)\right) + f\left(c(I_{\hat{k}_{t-1}}^{t-1}, B^{t-1})\right) \right), \end{aligned} \quad (5)$$

In the experiments, we have used VGG-16 [27], SqueezeNet [40] and Inception-ResNet v2 [30] networks to extract features. We crop the image patch and resize to suitable sized image for the network. To capture the semantic similarity among image patches, the features are extracted from the last layer except the classification layer. The extracted features are then normalized according to their L_2 norms. In this case, we can use Equation (5) to determine the similarity.

4.3. Combined Score and Focus Index Selection

Using the content similarity alone to select an optimal image from the focal stack can cause a problem when there are occluders which have similar appearance to the target object. Suppose that the target object which is maximally focused at focus index k_{target} moves to the behind another object which has very similar or even identical appearance to the target object. In this scenario, our content similarity measure can be failed to identify the focus index k_{target} by selecting the focus index where the occluder is maximally focused, if the feature similarity is the highest with the occluder.

To resolve such problem, we finally combine the aforementioned two scores, focus and content scores (Equations (3) and (4)), to examine the optimality of focal planes. Since the values computed by those two scoring methods are in different ranges, we convert them to the relative value with respect to the scores of the selected focus index in the previous frame $I_{\hat{k}_{t-1}}^{t-1}$. As a result, our combined score for the focus index, fi-score(I_k^t) is defined as follows:

$$\text{fi-score}(I_k^t) = \frac{1}{2} \cdot \left(\frac{\text{f-score}(I_k^t)}{\text{f-score}(I_{\hat{k}_{t-1}}^{t-1})} + \frac{\text{c-score}(I_k^t)}{\text{c-score}(I_{\hat{k}_{t-1}}^{t-1})} \right). \quad (6)$$

To find an optimal focus index where we perform the object tracking, one straightforward approach is taking the index having maximum $\text{fi-score}(\cdot)$. This strategy, however, can cause undesired focus index selection as we discussed in Section 4.1. Moreover, the hill-climbing heuristic proposed in [2] can be missing the optimum. We thus propose to search two peaks in terms of $\text{fi-score}(\cdot)$, towards decreasing and increasing focus indices. We also start from the optimal focus index \hat{k}_{t-1} of the previous

frame, and search two focus indices k_L and k_R which have locally maximum fi-score. Specifically, we search for two indices k_L and k_R that satisfy the follows:

$$\begin{aligned} \text{fi-score}(I_{k_L-1}^t) &\leq \text{fi-score}(I_{k_L}^t), & \text{fi-score}(I_{k_L}^t) &\geq \text{fi-score}(I_{k_L+1}^t), \\ \text{fi-score}(I_{k_R-1}^t) &\leq \text{fi-score}(I_{k_R}^t), & \text{fi-score}(I_{k_R}^t) &\geq \text{fi-score}(I_{k_R+1}^t), \end{aligned} \quad (7)$$

where $k_L \leq \hat{k}_{t-1}$ and $k_R \geq \hat{k}_{t-1}$. Among the indices k_L satisfying the conditions, we select the largest index which is the closest to the \hat{k}_{t-1} . Similarly, we pick the smallest one for k_R . We then select one of K_L and K_R closer to \hat{k}_{t-1} as \hat{k}_t . If K_L and K_R have the same distance from \hat{k}_{t-1} , then we take one having higher fi-score(\cdot). Figure 3 illustrates the overall process of our focus index selection algorithm. And Figure 4 shows a practical example of the optimal focus index selection. Once fi-score is computed for given focal stack, the peak selection algorithm looks for two local peaks in both left and right directions from the focus index selected in the previous frame. Among two peak indices, the one closer to the focus index at the previous frame is then selected as an optimal focus index of the current frame.

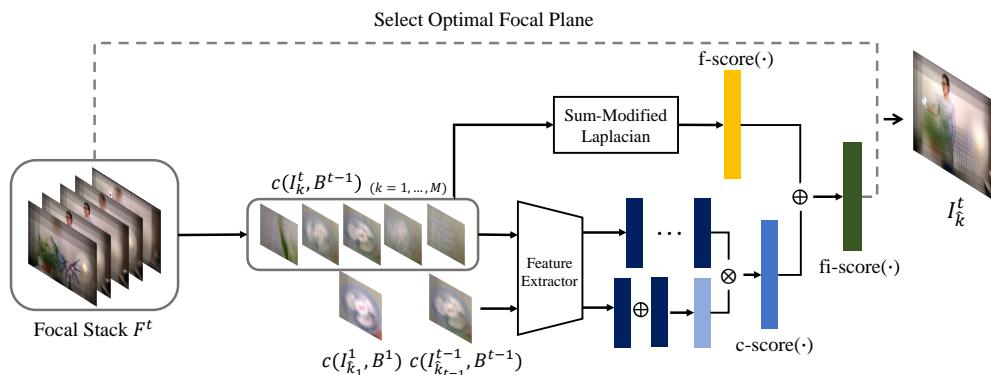


Figure 3. Illustration of our focal plane selection method.

Please note that, extensive comparisons between the f-score and fi-score in both quantitative and qualitative manners are available in Section 6.

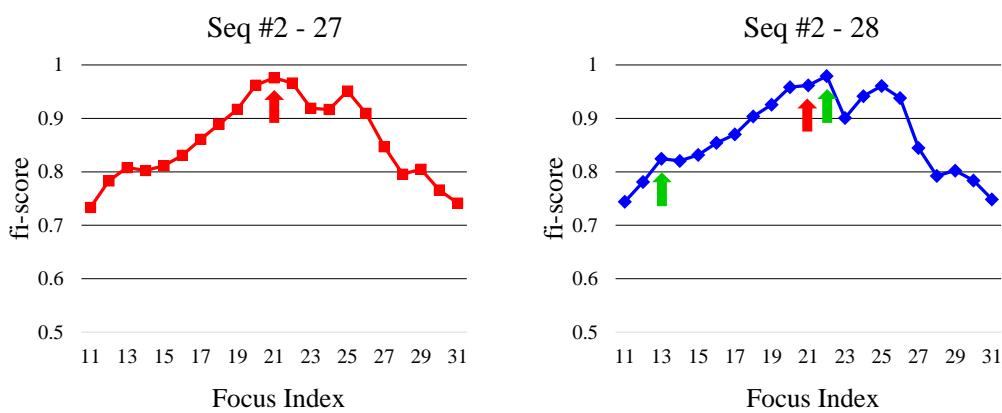


Figure 4. The left and right graphs show the fi-scores with respect to the focus index at 27th and 28th frames of Seq #2 (Section 6), respectively. The red arrowed index is the selected focus index at the 27th frame. In the 28th frame, fi-scores are first computed, and two local peaks are searched from the focus index of the previous frame. The green arrowed indices are the identified two indices. We then select closer index from the focus index of the previous frame among two peaks as an optimal focus index. In this example, the right green arrowed index is selected.

5. Bounding Box Proposal Based on Focus Index

To overcome the difficulty in the object-tracking task caused by the motion and size changes of the object, many of existing tracking techniques first conduct a bounding box proposal stage. The goal of this proposal stage is that producing a candidate bounding box which is nicely fit to the target object.

In the object tracking on plenoptic image sequences, Bae et al. [3] have presented a simple heuristic for bounding box proposal. They suggest generating proposals based on three fixed scales (3% smaller, the same, and 3% larger) from the selected proposal at the previous frame. They then feed each proposal to the Boosting tracker [5] and select the best one among those boxes based on the tracking confidence score determined by the Boosting tracker. Although this heuristic has been reported to provide accuracy enhancement over the baseline, using fixed scaling ratio is still too ad-hoc and their method is highly depending on the Boosting tracker. In other words, it is hard to take advantage of other advanced trackers. One may think of alternative proposal methods that use object detectors. In such approaches, the proposal is selected among the detected object regions according to the similarity to the target object. We, however, observed that such proposal methods provide unstable results on the plenoptic image sequences since the detectors often failed to find the target object even if we have used a reliable detector such as Faster R-CNN [32]. This is mainly because such detectors are not robust to blurs that the plenoptic images intrinsically have and are confused from the other objects.

The distance between the target object and camera is directly related to the size of the target object in the image plane. For instance, the size of the target object in the image becomes larger as the object approaches to the fixed camera. Since the focus index is tightly related to the distance between the object and camera, it is a strong cue to estimate the scale of the object in the image. In this section, we introduce a novel bounding box proposal method that takes both the optimal focus index and the motion of the object into account.

5.1. Scaling Factors

To derive appropriately sized proposals, we explore the focal stack parameters. Among various parameters we concentrate three variables, (μ_{min}, μ_{max}) and $\mu_{interval}$, related to the distance of the focal planes from the camera. Specifically, μ_{min} and μ_{max} indicate the range of the distances between focal planes and the camera, and $\mu_{interval}$ is a gap between two consecutive focal planes. Let us introduce a scaling factor s^t at t th frame that determines the relative size change from the selected proposal in the previous frame $t - 1$, based on difference in the focus indices \hat{k}^{t-1} and \hat{k}^t computed by our focus index selection algorithm described in Section 4.

Since size of the target object is proportionally decreasing to its distance from the camera, the relative scale between two frames t and $t - 1$ can be expressed as follows:

$$\text{size}(t) = \frac{d_0 + \mu_{interval} \times \hat{k}^{t-1}}{d_0 + \mu_{interval} \times \hat{k}^t} \times \text{size}(t - 1), \quad (8)$$

where d_0 is the distance from camera to its closest focal plane. Figure 5a illustrates a geometric interpretation of the object size changes with respect to the focus index. However, it is hard to have the exact value of d_0 in practice, so we approximate d_0 as $\mu_{max} - \mu_{min}$. We then finally define our scaling factor s^t as follows:

$$s^t = \frac{(\mu_{max} - \mu_{min}) + \mu_{interval} \times \hat{k}^{t-1}}{(\mu_{max} - \mu_{min}) + \mu_{interval} \times \hat{k}^t}. \quad (9)$$

Intuitively, we produce smaller proposals when the focus index is increased from the previous frame and larger ones when the focus index is decreased. As a result, the width \tilde{w} and heights \tilde{h} of our generated proposals are shrunk or enlarged as follows:

$$\tilde{w} = \hat{w}^{t-1} + z_s(s^t - 1)\hat{w}^{t-1} \text{ and } \tilde{h} = \hat{h}^{t-1} + z_s(s^t - 1)\hat{h}^{t-1}, \quad (10)$$

where \hat{w}^{t-1} and \hat{h}^{t-1} are the width and height of the selected proposal at the previous frame $t - 1$, respectively. The parameter z_s is multiplied to generate proposals in multi-scale. When we propose a single scaled box, the z_s should be 1. In evaluation (Section 6), we have proposed boxes with three different scales, $z_s = 0, 1$, and 2.

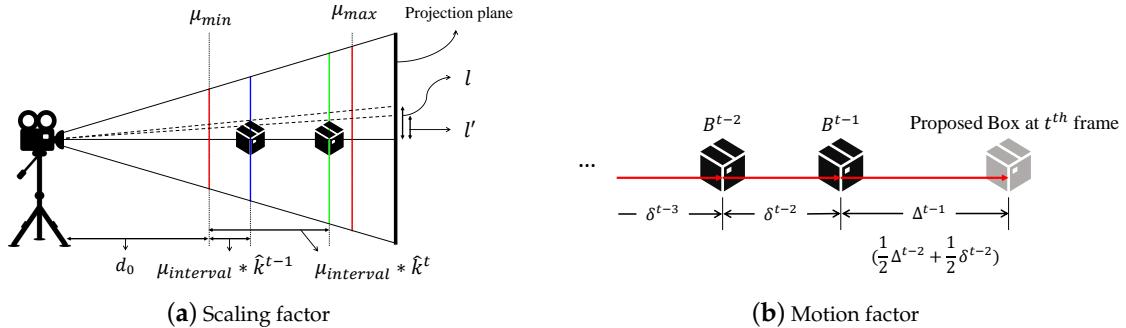


Figure 5. (a) This figure shows the geometric relationship between the object size in the image plane and the focus index. Since the focus index reflects the distance between the object and camera, the size of the object can be approximated based on the focus index changes. Our detailed derivation is given in Section 5.1. (b) This figure illustrates our motion factor formulation. Our underlying assumption is that the motion of object is not significantly changing from past a few frames. The displacement from the previous frame is then predicted as a linear combination of past displacements, where the weights are exponentially decreasing with respect to frame index distance.

5.2. Motion Factors

In addition to predict the scale changes, estimating the location of the target object is also crucial to produce more accurate proposals. To estimate the location of the target object, we take the trajectory of target objects in the past frames into account. Specifically, we predict the 2D location $c^t = [\bar{x}^t, \bar{y}^t]^T$ at a frame t based on the center positions $c^1 = [x^1, y^1]^T, \dots, c^{t-1} = [x^{t-1}, y^{t-1}]^T$ of past tracking results B^1, B^2, \dots, B^{t-1} . Let us first define the displacement δ_i between i and $i + 1$ frames as the follows:

$$\delta^i = c^{i+1} - c^i = [x^{i+1} - x^i, y^{i+1} - y^i]^T \quad (11)$$

Our assumption on this task is that motion of the target object is not significantly changed from past few frames and highly likely to be similar to recent history of the trajectory. We design the following model by reflecting the assumption:

$$c^t = c^{t-1} + z_m \Delta^{t-1}, \quad (12)$$

where

$$\Delta^{t-1} = \frac{1}{2} \delta^{t-2} + \frac{1}{4} \delta^{t-3} + \frac{1}{8} \delta^{t-4} + \dots + \frac{1}{2^{t-2}} \delta^1 = \sum_{i=1}^{t-2} \frac{1}{2^{t-i-1}} \delta^i. \quad (13)$$

and z_m is a parameter for generating multiple bounding box proposals as we have in the scale factor (Section 5.1). The displacement of recent frames influences more on our prediction, and the factors are exponentially decreasing with respect to the distance of frame index. Since our estimated displacement is defined recursively, it can be efficiently computed as follows (Refer Figure 5b):

$$\Delta^{t-1} = \frac{1}{2} \Delta^{t-2} + \frac{1}{2} \delta^{t-2}. \quad (14)$$

Once multiple bounding box proposals are generated based on various z_s and z_m , we are required to select the best one among the candidates. We evaluate the proposals according to the image patch

similarity. Specifically, the proposal which is the most similar to the tracking result at the previous frame $c(I_{k_{t-1}}^{t-1}, B^{t-1})$ is selected and feed to tracker. To compute the similarity among patches we employ the SSIM score [41].

6. Evaluation

In this section, we evaluate our proposed techniques for the object tracking on plenoptic image sequences.

6.1. Benchmarks

We first introduce the following benchmarks used in the experiments:

- **Seq #1:** The simplest one among three benchmarks. The target object is simply moving horizontally in the image plane and partially occluded. This sequence consists of 150 frames and the size of the focal stack is 14.
- **Seq #2:** The target object is moving far from the camera, so its size in the image is getting smaller. The occlusion is more serious than Seq #1. The number of frames and the size of the focal stack are 220 and 40, respectively.
- **Seq #3:** Contrary to Seq #2, the target object is approaching to the camera, so its size is getting larger. This sequence has 230 frames and the size of the focal stack is the same with Seq #2. The occlusion is the most serious among three sequences.

Please note that, the target objects of Seq #1, #2 and #3 are a paper box, an electric hand fan, and a cup, respectively. The cup in Seq #3 is a simple solid colored cup which does not have rich visual features. It is smaller than the other target objects while the Seq #3 has the most serious occlusions. These reasons lead the Seq #3 the most difficult scenario among tested benchmarks. Table 1 describes the parameters and Figure 6 shows several frames of benchmarks. We have used a plenoptic camera consisting of a 5×5 camera array to construct our benchmarks. Details about the equipment is described in [2,3].

Table 1. Parameters of the plenoptic image sequences used in the experiments.

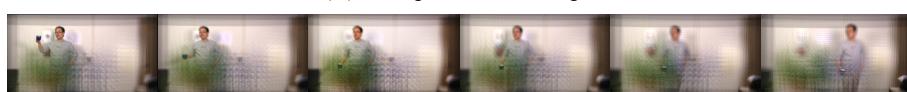
	Seq #1	Seq #2	Seq #3
μ_{min}	−0.9	−6.8	−6.7
μ_{max}	−0.2	0.8	0.9
μ_{int}	0.05	0.2	0.2
Focal Stack Size M	14	40	40
Frames	150	220	230



(a) Example frames of Seq #1



(b) Example frames of Seq #2



(c) Example frames of Seq #3

Figure 6. Several frames from our benchmark. Among images in the focal stack, we show the images selected by our proposed focus index selection method described in Section 4.

6.2. Evaluation Protocol

All the tested methods are evaluated using two accuracy measures for the object tracking. For a sequence of estimated bounding boxes B^1, \dots, B^N , a distance score is defined as the average of distances between centers of estimated and ground truth bounding boxes as follows:

$$\text{dist-score}(B^1, \dots, B^N) = \frac{1}{N} \sum_{t=1}^N \sqrt{(x^t - x_{GT}^t)^2 + (y^t - y_{GT}^t)^2}, \quad (15)$$

where (x^t, y^t) and (x_{GT}^t, y_{GT}^t) are the center positions of estimated and ground truth boxes at frame index t , respectively. The second accuracy measure, an overlap score, is evaluating the intersection of the estimated and ground truth boxes. In the experiments we used a constant-sized bounding box for both estimation and ground truth, so the overlap score is defined as follows:

$$\text{overlap score}(B^1, \dots, B^N) = \frac{1}{N} \sum_{t=1}^N \frac{|B^t \cap B_{GT}^t|}{|B^t \cup B_{GT}^t|}, \quad (16)$$

where $|\cdot|$ is the number of pixels contained in the region. Please note that the ground truth bounding boxes are labeled by human. We conducted the experiment more than 10 times for each sequence with varying initial bounding boxes. Specifically, the number of trials on Seq #1, #2, and #3 are 18, 12, and 14 cases, respectively. The reported accuracy and timings are the averages of such many trials.

We mainly compare the following five methods:

- **Baseline #1** [2]: This method uses only the f-score to select a focus index and Boosting tracker [5].
- **Baseline #2** [3]: An extension from Baseline #1. This method additionally performs image sharpening process and generates proposals based on a fixed ratio of sizes. We set the ratio as 3% as suggested in [3].
- **Ours(B)**: Our method using fi-score (Section 4) to select a focus index, producing proposals based on the scaling and motion factors (Section 5), and using Boosting tracker [5].
- **Ours(M)**: Similar to Ours(B) but using MOSSE tracker [6].
- **Ours(MC)**: Similar to Ours(M) but the MOSSE tracker is using CNN features [7].

All the tested methods are implemented in Python language, and all the experiments are conducted on a machine consisting of Intel i7-8700K CPU, 16 GB main memory, and NVIDIA GTX 1080 GPU.

In our focus index selection algorithm (Section 4), we employed image features to estimate visual similarity among image patches. We specifically used deep features extracted from the last fully connected layer of pre-trained version of VGG-16 [27] model which is publicly available. To train correlation filters of MOSSE tracker with CNN features, we used the activated feature map of the first convolution layer as suggested in [7]. Our proposal technique can produce multiple candidate boxes by applying different parameters z_s and z_m . In experiment we mostly used 3 values, 0, 1, and 2, for each parameter, thus we considered 9 proposals in total.

6.3. Experimental Results

Table 2 shows the object-tracking accuracies for the tested method on three different benchmarks. Among techniques using the Boosting tracker, Baseline #1, #2, and Ours(B), Ours(B) provided the highest accuracy in all the tested datasets. Moreover, the second-best method, Baseline #2, is tightly coupled with the Boosting tracker. Its proposal algorithm is designed and optimized for the Boosting tracker, so Baseline #2 cannot accept other trackers such as MOSSE. On the other hand, our method can use any advanced trackers since there is not any assumption on the tracker.

Table 2. Tracking accuracies (D and O indicate the distance and overlap scores, respectively.)

Method	Focus	Proposal	Tracker	Seq #1		Seq #2		Seq #3	
				D	O	D	O	D	O
Baseline #1 [2]	f-score	×	Boosting	44.502	0.852	59.997	0.712	142.796	0.576
Baseline #2 [3]	f-score	Fixed ratio	Boosting	39.293	0.867	55.916	0.735	41.101	0.822
Ours(B)	fi-score	Ours	Boosting	28.256	0.904	34.980	0.825	15.014	0.858
Ours(M)	fi-score	Ours	MOSSE	8.400	0.959	26.584	0.882	11.394	0.889
Ours(MC)	fi-score	Ours	MOSSE+CNN	11.978	0.946	27.469	0.879	9.148	0.915

By applying MOSSE tracker and its extension with CNN features to our technique, Ours(M) and Ours(MC), consistently outperformed the other tested methods including Ours(B). These experimental results verify that our method can provide even higher accuracy if we use more accurate 2D tracker.

To validate the benefits of our focus index selection technique (Section 4) we compared the tracking results based on two different focus index scoring schemes, f-score (Section 4.1) and fi-score (Section 4.3), and results of the experiments are given in Table 3. As reported in Table 3, our fi-score based on both the focus measure and visual similarity provides significantly higher tracking accuracy over the f-score used in Baseline #1 and #2. Please note that, we did not use any proposal algorithms in this experiment for a fair comparison.

Table 3. Comparison between f-score and fi-score with various trackers in Seq #3. Each pair of numbers denote the distance score and overlap score. Focus index selection based on our fi-score significantly outperform the f-score used in [2,3]. Please note that, the proposal algorithms are excluded for a fair comparison.

	Boosting	MOSSE	MOSSE+CNN
f-score	28.665/0.853	31.630/0.829	13.601/0.873
fi-score	19.984/0.842	11.534/0.888	9.807/0.909

We also report the qualitative results corresponding to Table 3. Figure 7 shows the image patches at the position of the target object from the focus indices selected by f-score and fi-score. Since the f-score only relies on the focus measure, the selected focus index based on the f-score index can be wrong once the target object got behind an occluder. As shown in Figure 7a, the f-score-based technique identified the focus index where the occluder is maximally focused. On the other hands, our proposed fi-score robustly found appropriate focus indices over frames, since it considers not only the focus measure but also the visual consistency. The tracking results corresponding to this comparison is reported in Figure 8. The tracking process based on the f-score failed to locate the target object once it is occluded, while one with our fi-score was correctly estimating the target object. Please note that, in those experiments the proposal algorithms are excluded and MOSSE tracker was used.

We also conducted experiments of our method with and without proposals, and its results are shown in Table 4. As reported, our candidate box proposal algorithm consistently provided accuracy improvements. The improvement on Seq #2 was the largest among three sequences mainly because Seq #2 has more varied motions and object sizes compared to Seq #1, and #3. These results validate that our proposal algorithm based on scaling and motion factors assists trackers to accurately locate the target object.

Let us report qualitative analysis of our proposal algorithm in Figure 9 by comparing Baseline #1 and Ours(B) in Seq #1. We mainly investigate our motion factor in the candidate box proposal stage. Figure 9a,b shows the tracking results of Baseline #1 and Ours(B), respectively. Baseline #1 failed to locate the target object when it is moving fast. Since most of trackers including Boosting-based techniques incrementally learn the appearance of the object, they hardly recover the tracking capability once they lost the object. On the other hand, our proposal algorithm considering the motion factor

can assist the trackers to recover its tracking capability. Our tracker has lost the object between first and second columns in Figure 9b; however, it immediately relocated the target object based on the proposal bounding box as shown in the third columns. Please note that, the selected proposals are illustrated in red boxes.

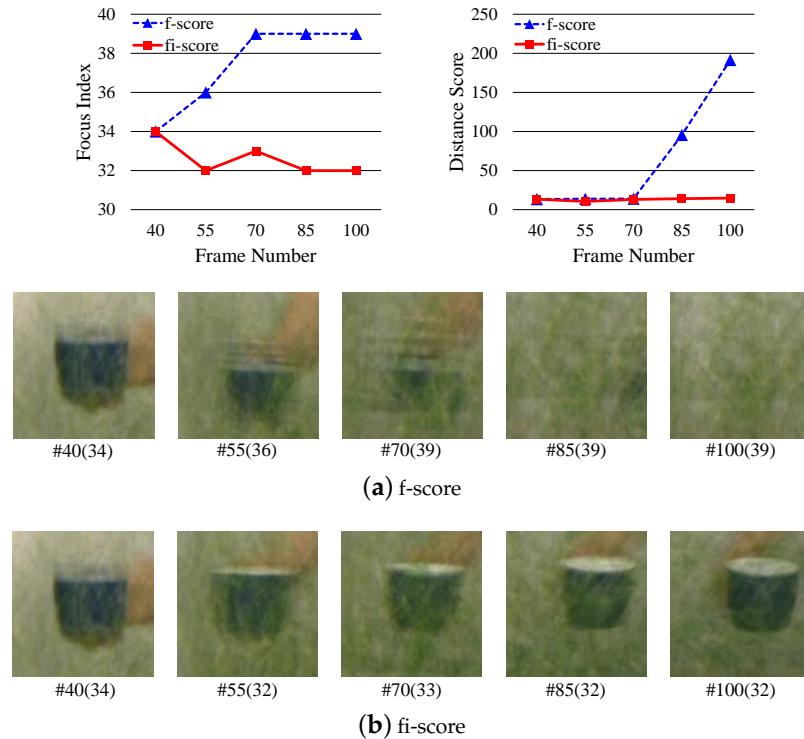
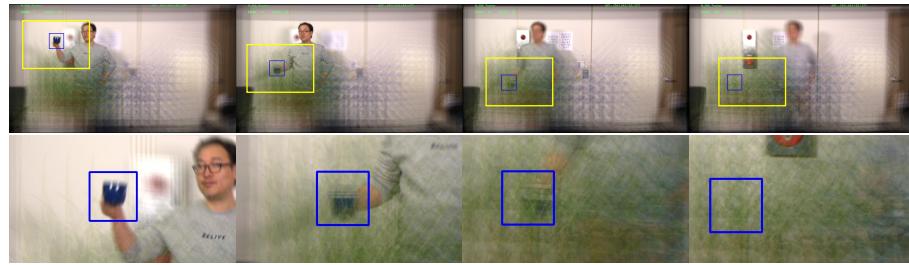


Figure 7. Qualitative comparison in Seq #3 between two scoring methods for the focus index selection, f-score and fi-score. A pair of values for each image patch denotes the frame index and the computed optimal focus index. Since the f-score only considers the focus measure, it can identify a wrong focus index when the target object gets behind the occluder. As shown in (a), the focus index selection based on f-score suggested an optimal focus index where the occluders are maximally focused. On the other hands, our fi-score-based focus index selection algorithm stably identifies appropriate focus indices over frames. Please note that, the quantitative results corresponding this figure is given in Table 3, and the proposal algorithms are excluded in this experiment.

Table 4. Accuracies of our methods with/without proposals.

	Seq #1		Seq #2		Seq #3	
	Distance	Overlap	Distance	Overlap	Distance	Overlap
Ours(M) w/o Proposal	8.789	0.957	43.868	0.807	11.534	0.888
Ours(M)	8.400	0.959	26.584	0.882	11.394	0.889
Ours(MC) w/o Proposal	14.523	0.936	43.683	0.810	9.807	0.909
Ours(MC)	11.978	0.946	27.469	0.879	9.148	0.915

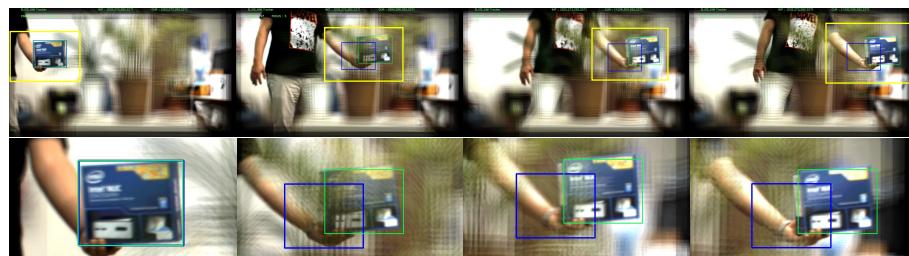


(a) Tracking results of Ours(M) based on f-score in Seq #3

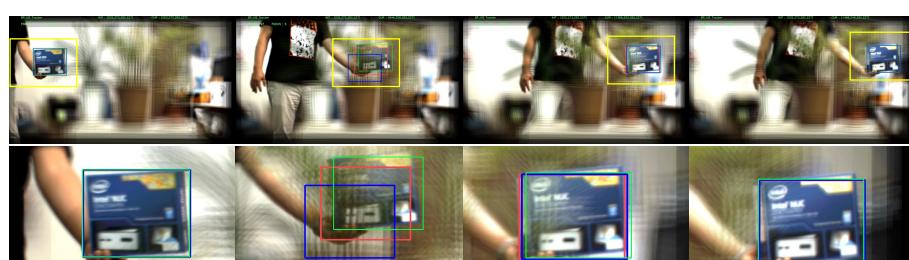


(b) Tracking results of Ours(M) based on fi-score in Seq #3

Figure 8. A comparison between the tracking results based on f-score and fi-score in Seq #3. The tracking process based on the f-score missed the target object once it was occluded because a wrong focus index is selected, while one with our fi-score was providing stable tracking performance. Please note that, we did not perform any candidate box proposal stage and MOSSE tracker is used in this experiment.



(a) Tracking results of Baseline #1 in Seq #1

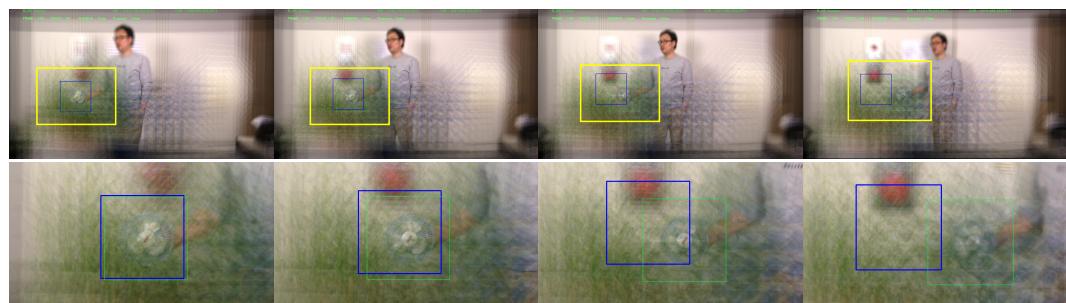


(b) Tracking results of Ours(B) in Seq #1

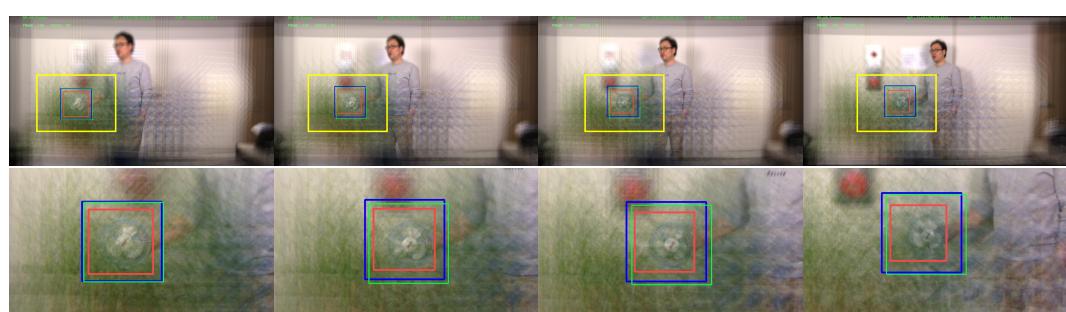
Figure 9. Qualitative analysis of the motion factor in our proposal algorithm. The yellow boxes in the first rows of (a,b) are magnified to the second rows for a clearer presentation. The green boxes are the ground truth bounding boxes while blue boxes are the tracking results. In the second columns, both methods lost the target object while it is moving fast. (a) Baseline #1 failed to recover its tracking capability in the following. (b) Ours(B), however, immediately relocated the target object by assistant of the proposal (red boxes). These qualitative results clearly validate the benefits of our proposal algorithm.

Another qualitative results to evaluate our proposal algorithm is given in Figure 10. Figure 10a,b show tracking results of Ours(B) without and with our proposal algorithm in Seq #2, respectively. When the size of the target object is getting smaller, as shown in the third column of the Figure 10a, Ours(B) without the proposals failed to estimate a correct location. On the other hand, our proposal algorithm generated suitable sized proposals (red boxes) and thus we could precisely locate the target object even with the scale changes.

Since our proposed focus index selection and candidate proposal techniques use image features to capture the content of the images, a choice of the feature extractor is fairly important. In all the aforementioned experiments, we used the VGG-16 network [27] as the feature extractor. Specifically, we used the last fully connected layer of VGG-16 in the focus index selection and the first convolution layer to train MC (MOSSE+CNN) tracker. The VGG-16 provides high accuracy in the image classification; however, it is known that its feature-extraction speed is quite slow. In order to see the performance of our proposed techniques when using lightweight feature extractor, we conducted experiments with SqueezeNet [40]. We also included the Inception-ResNet v2 [30] which showed one of the state-of-the-art classification performances recently. We report the tracking results with three different feature extractors, VGG-16, SqueezeNet and Inception-ResNet v2, in Table 5. The tracking accuracies are not significantly different. Since we have used pre-trained models publicly available for the feature extraction, the feature-extraction capability for particular objects can be diverse. Such diversity in the capability causes inconsistent performance tendency; however, our method still significantly outperformed the baseline techniques.



(a) Tracking results of Ours(B) without the proposal algorithm in Seq #2



(b) Tracking results of Ours(B) with the proposal algorithm in Seq #2

Figure 10. Qualitative evaluation of object tracking with/without the proposal algorithm in Seq #2. As the target object is getting smaller, our proposal algorithm generates the proposals which fit to the object. The same colors are used with Figure 9 for representing ground truth, tracking results, and proposals.

Table 5. Average Distance scores with three different feature extractors, VGG-16, SqueezeNet and Inception-ResNet v2. P denotes our proposal algorithm.

	Seq #1			Seq #2			Seq #3		
	VGG	SQZ	IncRes	VGG	SQZ	IncRes	VGG	SQZ	IncRes
Ours(M) w/o P	8.789	8.624	8.627	43.868	4.957	15.967	11.534	11.509	23.347
Ours(M)	8.400	8.322	8.271	26.584	4.653	15.954	11.394	11.358	12.485
Ours(MC) w/o P	14.523	85.041	44.089	43.683	23.516	34.385	9.807	11.956	15.021
Ours(MC)	11.978	18.765	13.939	27.469	18.566	11.685	9.148	11.828	9.450

Let us finally report the average tracking speed. Baseline #1 and #2 took 0.1 and 0.16 seconds per frame, respectively. On the other hands, Ours(B), Ours(M), and Ours(MC) took 0.37, 0.32, and 0.70 s per frame on average when we used VGG-16 network as the feature extractor, respectively. As reported in Table 6, the tracking times of MOSSE tracker with VGG-16 and Inception-ResNet are similar. The MC tracker with the Inception-ResNet v2 provides the slowest tracking speed, because it has huge sized first convolution layer so that the longer time is consumed for updating correlation filters. On the other hand, our proposed techniques can be accelerated if we use computationally lighter models such as SqueezeNet. For instance, when we combined our proposed techniques with the MOSSE tracker and SqueezeNet as the feature extractor, our method provided both higher accuracy and faster speed compared to Baseline #2. Moreover, the fastest method among the tested combinations, our focus index selection method combined with SqueezeNet, provided significantly higher accuracy over Baseline #1 and #2.

Table 6. Tracking times per frame with three different feature extractors, VGG-16, SqueezeNet and Inception-ResNet v2. The timings are measured with Seq #2 and #3.

	Time(s)/Frame		
	VGG	SQZ	IncRes
Ours(M) w/o Proposal	0.275	0.085	0.281
Ours(M)	0.319	0.155	0.322
Ours(MC) w/o Proposal	0.658	0.565	1.111
Ours(MC)	0.706	0.633	1.186

7. Conclusions

We have presented a novel object-tracking method for plenoptic image sequences. To find an appropriate focal plane where the target object clearly appears, we have developed a content-aware focus index selection technique based on the visual consistency and focus measure. In addition to the focal plane selection, we have provided candidate box proposal scheme that considers two factors, a scale factor based on the change of the focus index and a motion factor estimated from the past trajectory of the target object, and selects an optimal proposal according to the content similarity. We have extensively evaluated our technique over the prior tracking methods for the plenoptic image sequences. Since our method is orthogonal to the choice of tracker for 2D images, we can take advantage of any advanced tracking algorithms designed for ordinary 2D image sequences. For the experiments, we have combined three popular trackers to our technique. According to the experimental results, our method has provided significantly higher accuracy over the compared methods. The performance improvements were consistent and significant. Those results clearly confirm the benefits of our proposed algorithms.

Author Contributions: Conceptualization, D.B. and J.H.; Data curation, D.B. and J.K.; Funding acquisition, J.K. and J.H.; Investigation, D.B. and J.H.; Methodology, D.B. and J.H.; Project administration, J.H.; Software, D.B., J.K., and J.H.; Supervision, J.H.; Validation, D.B. and J.H.; Visualization, D.B. and J.H.; Writing—original draft, D.B. and J.H.; Writing—review & editing, D.B. and J.H.

Funding: This work was supported by Institute for Information & Communications Technology Promotion(IITP) grant funded by the Korean government(MSIT). [2016-0-00009, Authoring Platform Technology for Next-Generation Plenoptic Contents].

Conflicts of Interest: The authors declare no conflict of interest.

References

1. Cho, M.; Javidi, B. Three-dimensional tracking of occluded objects using integral imaging. *Opt. Lett.* **2008**, *33*, 2737–2739. [[CrossRef](#)] [[PubMed](#)]
2. Kim, J.W.; Bae, S.J.; Park, S.; Kim, D.H. Object tracking using plenoptic image sequences. In Proceedings of the Three-Dimensional Imaging, Visualization, and Display, Anaheim, CA, USA, 9–13 April 2017; International Society for Optics and Photonics: Bellingham, WA, USA, 2017; Volume 10219, p. 1021909.
3. Bae, D.H.; Kim, J.W.; Noh, H.C.; Kim, D.H.; Heo, J.P. Plenoptic imaging techniques to improve accuracy and robustness of 3D object tracking. In Proceedings of the Three-Dimensional Imaging, Visualization, and Display, Orlando, FL, USA, 15–19 April 2018; International Society for Optics and Photonics: Bellingham, WA, USA, 2018; Volume 10666, p. 106660H.
4. Pertuz, S.; Puig, D.; Garcia, M.A. Analysis of focus measure operators for shape-from-focus. *Pattern Recognit.* **2013**, *46*, 1415–1432. [[CrossRef](#)]
5. Grabner, H.; Grabner, M.; Bischof, H. Real-time tracking via on-line boosting. In *Proceedings of the British Machine Vision Conference*; BMVA Press: Blue Mountains, ON, USA, 2006; Volume 1, pp. 47–56.
6. Bolme, D.S.; Beveridge, J.R.; Draper, B.A.; Lui, Y.M. Visual object tracking using adaptive correlation filters. In Proceedings of the 2010 IEEE Computer Society Conference on Computer Vision and Pattern Recognition, San Francisco, CA, USA, 13–18 June 2010; pp. 2544–2550.
7. Danelljan, M.; Hager, G.; Shahbaz Khan, F.; Felsberg, M. Convolutional features for correlation filter based visual tracking. In Proceedings of the IEEE International Conference on Computer Vision Workshops, Santiago, Chile, 7–13 December 2015; pp. 58–66.
8. Smeulders, A.W.; Chu, D.M.; Cucchiara, R.; Calderara, S.; Dehghan, A.; Shah, M. Visual tracking: An experimental survey. *IEEE Trans. Pattern Anal. Mach. Intel.* **2014**, *36*, 1442–1468.
9. Hu, W.; Tan, T.; Wang, L.; Maybank, S. A survey on visual surveillance of object motion and behaviors. *IEEE Trans. Syst. Man Cybern. Part C* **2004**, *34*, 334–352. [[CrossRef](#)]
10. Yilmaz, A.; Javed, O.; Shah, M. Object Tracking: A Survey. *ACM Comput. Surv.* **2006**, *38*. [[CrossRef](#)]
11. Adam, A.; Rivlin, E.; Shimshoni, I. Robust fragments-based tracking using the integral histogram. In Proceedings of the 2006 IEEE Computer Society Conference on Computer Vision and Pattern Recognition, New York, NY, USA, 17–22 June 2006; Volume 1; pp. 798–805.
12. Liu, B.; Huang, J.; Yang, L.; Kulikowsk, C. Robust tracking using local sparse appearance model and k-selection. In Proceedings of the 2011 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Colorado Springs, CO, USA, 20–25 June 2011; pp. 1313–1320.
13. Zhang, K.; Zhang, L.; Yang, M.H. Fast compressive tracking. *IEEE Trans. Pattern Anal. Mach. Intel.* **2014**, *36*, 2002–2015. [[CrossRef](#)] [[PubMed](#)]
14. Grabner, H.; Matas, J.; Van Gool, L.; Cattin, P. Tracking the invisible: Learning where the object might be. In Proceedings of the 2010 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), San Francisco, CA, USA, 13–18 June 2010; pp. 1285–1292.
15. Babenko, B.; Yang, M.H.; Belongie, S. Robust object tracking with online multiple instance learning. *IEEE Trans. Pattern Anal. Mach. Intel.* **2011**, *33*, 1619–1632. [[CrossRef](#)] [[PubMed](#)]
16. Cai, Z.; Wen, L.; Lei, Z.; Vasconcelos, N.; Li, S.Z. Robust deformable and occluded object tracking with dynamic graph. *IEEE Trans. Image Process.* **2014**, *23*, 5497–5509. [[CrossRef](#)] [[PubMed](#)]
17. Kristan, M.; Matas, J.; Leonardis, A.; Felsberg, M.; Cehovin, L.; Fernandez, G.; Vojir, T.; Hager, G.; Nebehay, G.; Pflugfelder, R. The visual object tracking vot2015 challenge results. In Proceedings of the IEEE International Conference on Computer Vision Workshops, Santiago, Chile, 7–13 December 2015; pp. 1–23.
18. Henriques, J.F.; Caseiro, R.; Martins, P.; Batista, J. High-Speed Tracking with Kernelized Correlation Filters. *IEEE Trans. Pattern Anal. Mach. Intel.* **2015**, *37*, 583–596. [[CrossRef](#)] [[PubMed](#)]

19. Danelljan, M.; Shahbaz Khan, F.; Felsberg, M.; van de Weijer, J. Adaptive Color Attributes for Real-Time Visual Tracking. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Columbus, OH, USA, 24–27 June 2014.
20. Danelljan, M.; Hager, G.; Shahbaz Khan, F.; Felsberg, M. Learning Spatially Regularized Correlation Filters for Visual Tracking. In Proceedings of the IEEE International Conference on Computer Vision (ICCV), Santiago, Chile, 7–13 December 2015.
21. Lukezic, A.; Vojir, T.; Zajc, L.C.; Matas, J.; Kristan, M. Discriminative Correlation Filter with Channel and Spatial Reliability. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Honolulu, HI, USA, 21–26 July 2017; Volume 6, p. 8.
22. Shi, G.; Xu, T.; Guo, J.; Luo, J.; Li, Y. Consistently Sampled Correlation Filters with Space Anisotropic Regularization for Visual Tracking. *Sensors* **2017**, *17*, 2889. [[CrossRef](#)] [[PubMed](#)]
23. Li, F.; Zhang, S.; Qiao, X. Scene-Aware Adaptive Updating for Visual Tracking via Correlation Filters. *Sensors* **2017**, *17*, 2626. [[CrossRef](#)] [[PubMed](#)]
24. Xue, X.; Li, Y.; Shen, Q. Unmanned Aerial Vehicle Object Tracking by Correlation Filter with Adaptive Appearance Model. *Sensors* **2018**, *18*, 2751. [[CrossRef](#)] [[PubMed](#)]
25. Nayar, S.K.; Nakagawa, Y. Shape from Focus. *IEEE Trans. Pattern Anal. Mach. Intel.* **1994**, *16*, 824–831. [[CrossRef](#)]
26. Krizhevsky, A.; Sutskever, I.; Hinton, G.E. ImageNet Classification with Deep Convolutional Neural Networks. In *Advances in Neural Information Processing Systems* 25; Pereira, F., Burges, C.J.C., Bottou, L., Weinberger, K.Q., Eds.; Curran Associates, Inc.: New York, NY, USA, 2012; pp. 1097–1105.
27. Simonyan, K.; Zisserman, A. Very deep convolutional networks for large-scale image recognition. *arXiv* **2014**, arXiv:1409.1556.
28. He, K.; Zhang, X.; Ren, S.; Sun, J. Deep Residual Learning for Image Recognition. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Las Vegas, NV, USA, 27–30 June 2016.
29. Szegedy, C.; Vanhoucke, V.; Ioffe, S.; Shlens, J.; Wojna, Z. Rethinking the inception architecture for computer vision. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Las Vegas, NV, USA, 27–30 June 2016; pp. 2818–2826.
30. Szegedy, C.; Ioffe, S.; Vanhoucke, V.; Alemi, A.A. Inception-v4, inception-resnet and the impact of residual connections on learning. In Proceedings of the AAAI, San Francisco, CA, USA, 4–9 February 2017; Volume 4, p. 12.
31. Girshick, R.; Donahue, J.; Darrell, T.; Malik, J. Rich feature hierarchies for accurate object detection and semantic segmentation. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Columbus, OH, USA, 24–27 June 2014; pp. 580–587.
32. Ren, S.; He, K.; Girshick, R.; Sun, J. Faster r-cnn: Towards real-time object detection with region proposal networks. In *Advances in Neural Information Processing Systems*; Curran Associates, Inc.: Montreal, Canada, 2015; pp. 91–99.
33. Redmon, J.; Divvala, S.; Girshick, R.; Farhadi, A. You Only Look Once: Unified, Real-Time Object Detection. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Las Vegas, NV, USA, 27–30 June 2016.
34. Wang, N.; Yeung, D.Y. Learning a deep compact image representation for visual tracking. In *Advances in Neural Information Processing Systems*; Curran Associates, Inc.: Lake Tahoe, Nevada, 2013; pp. 809–817.
35. Held, D.; Thrun, S.; Savarese, S. Learning to track at 100 fps with deep regression networks. In *European Conference on Computer Vision*; Springer: Cham, Switaerland, 2016; pp. 749–765.
36. Wu, Y.; Lim, J.; Yang, M.H. Online Object Tracking: A Benchmark. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Portland, OR, USA, 23–28 Jun 2013.
37. Babenko, A.; Slesarev, A.; Chigorin, A.; Lempitsky, V. Neural codes for image retrieval. In *European Conference on Computer Vision*; Springer: Cham, Switaerland, 2014; pp. 584–599.
38. Babenko, A.; Lempitsky, V. Aggregating local deep features for image retrieval. In Proceedings of the IEEE International Conference on Computer Vision, Santiago, Chile, 7–13 December 2015; pp. 1269–1277.
39. Fei-Fei, L.; Perona, P. A bayesian hierarchical model for learning natural scene categories. In Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition, San Diego, CA, USA, 20–25 June 2005; Volume 2, pp. 524–531.

40. Iandola, F.N.; Han, S.; Moskewicz, M.W.; Ashraf, K.; Dally, W.J.; Keutzer, K. SqueezeNet: Alexnet-level accuracy with 50x fewer parameters and <0.5 mb model size. *arXiv0* **2016**, arXiv:1602.0736.
41. Wang, Z.; Bovik, A.C.; Sheikh, H.R.; Simoncelli, E.P. Image quality assessment: From error visibility to structural similarity. *IEEE Trans. Image Process.* **2004**, *13*, 600–612. [CrossRef] [PubMed]



© 2018 by the authors. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license (<http://creativecommons.org/licenses/by/4.0/>).