Quick 3D Object Detection and Localization by Dynamic Active Search with Multiple Active Cameras

Takahito KAWANISHI Hiroshi MURASE Shigeru TAKAGI
NTT Communication Science Laboratories
3-1 Morinosato Wakamiya Atsugi-shi Kanagawa Pref. Japan
{kawanisi, murase, takagi}@eye.brl.ntt.co.jp

Abstract

This paper proposes a method for detecting known objects in 3D environments and estimating their positions with multiple pan-tilt-zoom cameras. Our search method, Dynamic Active Search, reduces the number of camera operations by predicting the existence of a target in wide angles, zooming-in a promising area, and confirming the target. Even when many reference images taken under various object views and various scales need to be searched for, Dynamic Active Search can detect the object efficiently. With multiple cameras, object detection and object localization become more efficient. Experiments show that Dynamic Active Search with four cameras is 2.5 times faster and 2 times more accurate than with a single camera.

1. Introduction

This paper proposes an algorithm for detecting and locating known objects in room environments with multiple pan-tilt-zoom cameras. This algorithm was devised with practical applications such as robots and surveillance systems in mind. In a wide environment, active camera control [1] search based on multiresolutional approach seems to be effective. Conventional methods that use a substraction method [4] or 3D range sensors [5] for the dicision of next camera zooming are unsatisfactory in terms of performance and flexibility. In 1996, we proposed a quick object detection method based on a color histogram of a target object:"Active Search with color histograms (AS)"[3]. AS greatly reduces the number of matching calculations between a reference histogram and an input image while preserving the matching accuracy. However, applications where illumination conditions, orientation of objects, and camera angles greatly vary will require quite a few (about a hundred) reference images and the search time will be long accordingly even with AS. Moreover, AS was intended for use with static cameras without pan-tilt-zoom functions. To cope with these problems, we enhanced AS in two points. First, we introduced a predict-a-best-direction/zoom-in/verify camera control strategy to reduce the number of pan-tilt-zoom operations. Second, we devised two methods to effectively search for objects with multiple reference histograms. We call this enhanced method Dynamic Active Search (DAS) [2]. A single camera or multiple cameras can be controlled by DAS. When multiple cameras are used, the search time becomes short because the search space for each camera becomes small. When object size is known, object location can be estimated by the direction of the object and its size in the input image. When multiple cameras are used, object position estimation becomes accurate. Experiments show that, compared to a single-camera configuration, DAS with four cameras improves search speed and position estimation accuracy 2.5 and 2 times, respectively.

2. Dynamic Active Search

DAS is a method for object detection in a room environment. If the target object is located far from the camera and pictures are taken at wide angles, the object size in the input image becomes too small to accurately identify and zooming operations are necessary as shown in Fig. 1. However, when zooming is performed, the angle of a view becomes small and many pan-tilt operations are necessary to cover all view as shown in the left of Fig. 2. We introduce the best-direction-first strategy shown in the right of Fig. 2 to reduce the number of camera operations.

2.1. Outline of DAS

Figure 3 outlines the DAS system. The system collects multiple reference images taken under various illumination conditions, object views, and zoom rates. These reference images are categorized into two types according to their sizes: Reference Image for Prediction (RIP) and Reference Image for Verification (RIV). RIP is a small reference image

1051-4651/02 $17.00 (c) 2002 IEEE

Figure 1. Zooming search in a room.

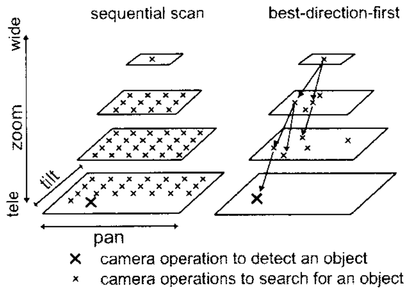


Figure 2. Camera control strategies.

(100∼400 pixels) and discriminative enough for roughly predicting the existence of the object. RIV is a large reference image (larger than 400 pixels) and sufficiently discriminative to decide the existence of the object. At the prediction stage, cameras are set at wide angles and objects are searched with RIPs. Sub-areas whose similarity exceeds a predefined threshold are zoomed-in and the existence of the objects is verified with RIVs. There is usually a large number of RIPs and RIVs and the number of matching operation therefore becomes huge. If two reference histograms, A and B, are similar and A is not similar to an input image C, then B is estimated not to be similar to C. DAS analyzes similarity among RIPs or RIVs and prunes many matching operations.

2.2. Prediction stage

In ordinary template matching, matching operations must be performed for each reference image independently. In the case of color histograms, it is possible to merge multiple reference histograms into a single histogram while guaranteeing that the target will not be missed. In what follows, the principle of AS and the idea of union histogram are described.

Active Search To detect and locate a reference object in the input image, AS [3] calculates the similarity between a reference histogram and a histogram for a focus region (cropped sub area) of an input image by histogram intersection. Intersection $S(R, F)$ of two histograms, R and F, is

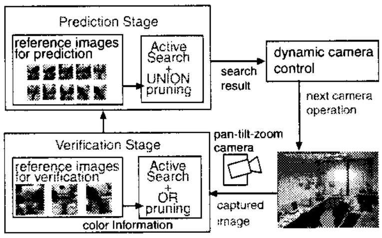


Figure 3. Dynamic Active Search system.

defined as

$$S(R, F) = \sum_{i}^{N} \frac{\min(R^i, F^i)}{|R|}, \quad |R| = \sum_{i}^{N} R^i, \quad (1)$$

where i is the code of the color and N is the number of codes. A similarity value that exceeds the predefined threshold means the object is located in that focus region. When the similarity value $S(R, F_a)$ of the reference histogram R and a focus region F_a is far below the threshold, the upper bound of the similarity value $S(R, F_b)$ of R and another focus region F_b that overlaps F_a is estimated as

$$|R| \cdot S(R, F_b) \leq |R| \cdot S(R, F_a) + |F_b - F_a|, \quad (2)$$

where $|F_b - F_a|$ denotes the number of pixels in F_b but not in F_a. AS calculates the pruning search space by using Eq. 2. That is, it skips focus regions whose calculated upper bound is lower than the threshold.

Union histogram The union histogram of the reference histograms is defined as

$$U^i = \max\left({R_0}^i, {R_1}^i, {R_2}^i, \cdots\right), \quad (3)$$

and has the following property:

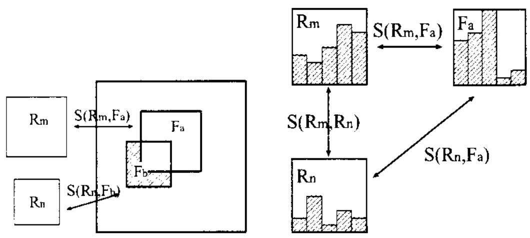
$$S(U, F) \geq S(R_m, F). \quad (4)$$

Therefore, when AS is performed with union histograms, a negative result (i.e. the target object is not found) guarantees the absence of the target object even if AS is performed with every reference histogram. In contrast, a positive result does not guarantee the presence of the target object. Therefore, candidate regions detected with union histograms are zoomed-in and checked with RIVs at the verification stage.

2.3. Verification stage

When two reference histograms, R_m and R_n, are given and their similarity value is not small, the similarity value $S(R_m, F_a)$ of R_m and a focus region F_a is used to estimate

1051-4651/02 $17.00 (c) 2002 IEEE



(a) Overlapping

(b) Relationship between two histograms

Figure 4. Focus regions and similarity relation.

the upper bound of the similarity value $S(R_n, F_b)$ of R_n and the other focus region, F_b, that overlaps F_a as

$$\begin{aligned} |R_n| \cdot S(R_n, F_b) &\leq |R_m| \cdot S(R_m, F_a) \\ &+ |R_n| \cdot (1 - S(R_m, R_n)) \\ &+ |F_b - F_a|, \end{aligned} \tag{5}$$

because $|R_n| \cdot S(R_n, F_a)$ is at most only the rest of R_n not including R_m larger than $|R_m| \cdot S(R_m, F_a)$ as shown in Fig. 4. Therefore, the pruning area for the other reference histograms can be calculated during AS with reference histogram R_m.

3. Object localization

When the target size is known and the target is confirmed in the input image, its position in the room can be estimated. The following subsections describe the calculation method for a single camera and that for multiple cameras.

3.1. Object localization by a single camera

When the object is detected by DAS, its position $\mathbf{T}^i(t_x^i, t_y^i)$ and its size in the input image $\mathbf{W}^i(w_x^i, w_y^i)$ are obtained, where i is the camera number. The direction of the object in the spherical coordinate is calculated as

$$\begin{aligned} \theta_x^i &= \arctan\left(\left(\frac{r_x^i - 2t_x^i}{r_x^i}\right)\tan\frac{f_x^i}{2}\right) \\ \theta_y^i &= \arctan\left(\left(\frac{r_y^i - 2t_y^i}{r_y^i}\right)\tan\frac{f_y^i}{2}\right), \end{aligned} \tag{6}$$

where, θ_x^i and θ_y^i are pan and tilt angles from the camera's optical axes, $\mathbf{F}(f_x^i, f_y^i)$ is the angle of view, and $\mathbf{R}^i(r_x^i, r_y^i)$ is the size of the input image. If the physical size of the object, $\mathbf{H}(h_x, h_y)$, is known, the distance between the camera and object l^i is calculated by

$$l^i = \frac{h_x \cdot r_x^i}{2w_x^i \cdot \tan\frac{f_x}{2}} = \frac{h_y \cdot r_y^i}{2w_y^i \cdot \tan\frac{f_y}{2}}, \tag{7}$$

Finally, the object position $\mathbf{Q}(q_x, q_y, q_z)$ is estimated as

$$\mathbf{Q} = \mathbf{P}^i + l^i \cdot \mathbf{Z}^i(\alpha^i) \cdot \mathbf{Y}(\beta^i) \cdot \mathbf{D}^i, \tag{8}$$

where $\mathbf{P}^i(p_x^i, p_y^i, p_z^i)$ is the camera position, $\mathbf{D}^i(d_x^i, d_y^i, d_z^i)$ the initial direction of the camera, $\mathbf{Z}$ and $\mathbf{Y}$ the rotation operators in the pan and tilt direction ($\alpha^i = c_x^i + \theta_x^i, \beta^i = c_y^i + \theta_y^i$), and $\mathbf{C}^i(c_x^i, c_y^i)$ the current pan-tilt angles from the initial direction.

3.2. Object localization by multiple cameras

If multiple cameras are used, there is no need to know the physical size of the target. The target position can be determined by solving the following triangulation equations:

$$\mathbf{Q} = \mathbf{P}^i + k^i \cdot \mathbf{Z}(\alpha^i) \cdot \mathbf{Y}(\beta^i) \cdot \mathbf{D}^i, \tag{9}$$

where target position $\mathbf{Q}$ and the distance k_i between the target and the camera i are unknown variables. These simultaneous equations can be solved by the singular value decomposition method for least mean square error.

4. Experiments

We conducted experiments to confirm the improvements in search speed by DAS and then experiments to localize the target object. For the object shown in Fig. 3, we prepared about 100 reference images (5 different lighting conditions, 3 views, 3~10 zoom steps). The smaller 50 images (100~400 pixels) were used as RIPs and the larger 50 (400~2000 pixels) as RIVs. Figure 5 shows the room environment.

There were four cameras, one mounted in each corner of the room. Circles $a \sim f$ show the positions we used for locating objects in our experiments. The computer was a SGI O2 (R12000, 400 MHz), and the cameras were SONY EVI–D30s (resolution: 320×240).

4.1. Search time

The CPU time and the camera operation time to search the room environment for an object (Table 1) were measured for search without and with prediction control.

Search time with prediction control was two times faster than that without it. Search times for a single camera and that for the four-camera configuration were also measured. The performance of the four-camera configuration was 2.5 times better than that of a single-camera configuration.

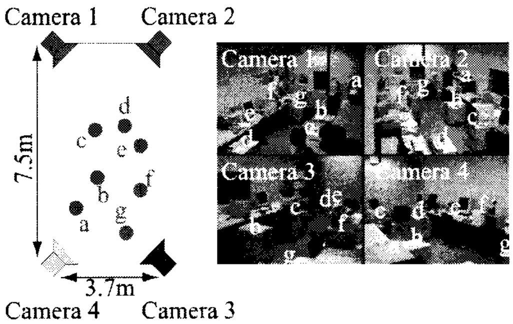
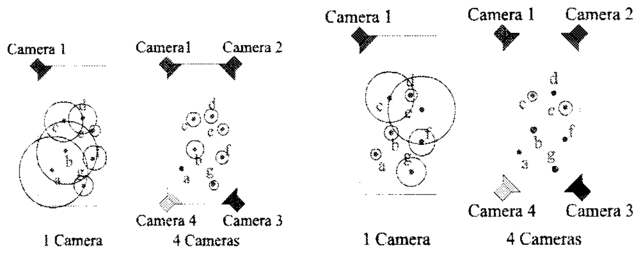


Figure 5. Room environment.



(a) Search time (b) Localization accuracy

Figure 6. Position effect.

4.2. Object localization accuracy

We measured the difference between the objects' estimated position and its correct position. Table 2 shows the average of errors for a single-camera configuration and the four-camera configuration. The object localization accuracy for a single camera was quite poor. This is because the size of an object in an input image cannot be determined precisely by the color histogram matching method. On the other hand, with four cameras, object position could be accurately estimated by the triangulation.

4.3. Position effect

The effect of object position on search time and localization accuracy was investigated. The results are summarized in Fig. 6. The radius of a circle is proportional to the search time or the localization error. With a single camera, there was little object position dependence of search time and localization accuracy. The best-direction-first strategy does not guarantee the existence of objects at the best direction. If the verification stage fails to find the object, the system backtracks and searches the next most promising direction. Backtracking increases search time. When multiple cameras are available, the search time becomes short. This may be because the possibility of a camera being at a good position for detecting objects is increased and the best camera has a small number of backtracks.

With triangulation, object localization becomes quite accurate throughout the room. This is because the accuracy of object directions obtained by DAS is much better than the accuracy of the object distance from the cameras. As a result of the camera zooming by DAS, the direction resolution becomes higher and the errors in the obtained object direction are smaller than without zooming.

5. Conclusion

We described Dynamic Active Search, a method that can search for an object quickly in a 3D room environment. Dynamic Active Search is applicable to both single-camera and multiple-camera configurations. However, with multiple cameras, the search speed and the object localization accuracy are improved 2.5 times and 2 times respectively compared to a single-camera configuration. This method cannot be applied for the objects whose colors are same as the background colors. In future work, we will conbine another feature for DAS to apply it to various room situations.

Table 1. Search time

	4 Cameras	1 Camera
non-DAS	31.9s	89.1s
	(16.9s,15.0s)	(49.8s,39.3s)
DAS	15.3s	40.8s
	(7.2s,8.2s)	(18.6s,22.2s)

(CPU time, camera operation time)

Table 2. Accuracy of localization

	4 Cameras	1 Camera
Localization error	18.9cm	124.8cm

References

[1] J. Aloinomous, A. Bandopdhay, and I. Weiss. Active vision. In *Proc. of ICCV'87*, pages 35–54, 1987.

[2] T. Kawanishi, H. Murase, S. Takagi, and M. Werner. Dynamic Active Search for quick object detection with pan-tilt-zoom camera. In *Proc. of ICIP2001*, volume 3, pages 716–719, Oct. 2001.

[3] V. V. Vinod and H. Murase. Focused color intersection with efficient searching for object extraction. *Pattern Recognition*, 30(10):1787–1797, 1997.

[4] T. Wada and T. Matsuyama. Appearance Sphere: Background model for pan-tilt-zoom camera. In *Proc. of ICPR'96*, pages 718–722, 1996.

[5] Y. Ye and J. Tsotsos. Where to look next in 3d object search. In *Proc. of ISCV*, pages 539–544, 1995.