Detection Function and its Application in Visual Tracking

Yiming Ye[a], John K. Tsotsos[b], Karen Bennet[c], and Eric Harley[b]

[a]IBM T.J. Watson Research Center, P.O. Box 704, Yorktown Heights, N.Y. 10598 USA
[b]Department of Computer Science, University of Toronto, Toronto, Ontario M5S 1A4
[c]IBM Centre for Advanced Studies, North York, Ontario, Canada M3C 1V7

ABSTRACT

This paper introduces the concept of detection function for assessing recognition algorithms. A detection function specifies the probability that a particular recognition algorithm will detect the target, given the camera viewing direction and angle size and the target position. The detection function thus determines the region of space in which the target can be detected with high probability using certain specified camera parameters. For this reason, it provides a natural language for discussion of the task of tracking an object moving in three dimensions using a camera with adjustable pan, tilt, and zoom. Most previous studies on visual tracking involve the use of a camera with fixed viewing direction and viewing angle size. We advocate, however, an algorithm wherein these camera parameters are actively controlled to keep the target in the field of view and to maintain its image quality. In this paper we study geometrical issues related to the detection function and describe a novel tracking algorithm.

Keywords: detection function, visual tracking, segmentation, active camera, pan, tilt, zoom

1. INTRODUCTION

The task of visually tracking three-dimensional objects arises in many contexts, including factory or outer space robotics and camera-activated surveillance or security systems. The topic has received considerable attention in the computer vision community over the past few years. The usual approach for tracking a rigid, three-dimensional object is to attempt to follow recognizable features (points or lines) of the object that are projected on the image. A loop is executed with the following major steps: prediction, projection, measurement, and adjustment. In the prediction step, the next three-dimensional pose (position and velocity) of the target is estimated based on its previous positions and velocities. The projection step uses the predicted information to calculate the projection of the target onto a two-dimensional image. The measurement step uses the calculated projection of target features to direct the search for corresponding features in the real image. The adjustment step combines information from measurement and prediction steps to produce new values of the target position and velocities, and these new values are used for the next prediction stage.

Recent work[1–12] examines the task of visually tracking non-rigid objects. For example, Huttenlocher[4] constructed a system for tracking a non-rigid object moving in a complex scene. The method extracts two-dimensional models of an object from a sequence of images. The basic idea is to decompose the image of the target moving in space into two components: a two-dimensional motion and a two-dimensional shape change. The motion component is factored out, and the shape change is represented explicitly by a sequence of two-dimensional models, one corresponding to each image frame. Darrell *et. al.*[1] have implemented vision routines to track people and to identify head/hand locations as they walk about in a room, and to provide foveation cues to guide an active camera to foveate head or hands. The system assumes a fixed background and that the person is facing the camera. Gavrila *et. al.*[2] construct a system to track unconstrained human movement using a three-dimensional model. Image sequences acquired simultaneously from multiple views allows them to recover the three-dimensional body pose at each time instant without the use of markers. Kakadiaris *et. al.*[6] present a method to mitigate the difficulties arising due to occlusion among body parts by employing multiple calibrated cameras in a mutually orthogonal configuration.

It is interesting to note that most previous studies in visual tracking emphasize the problem of tracking the moving image features of the target. On the other hand, issues related to determining the most suitable state parameters (pan, tilt, and zoom) of the camera during tracking have received considerably less attention. The sensor planning

John K. Tsotsos is a Fellow of the Canadian Institute for Advanced Research task for visual tracking, however, is very important since the state parameters of the camera determine the quality of the resulting image and indeed whether the target will be within the image. Thus, an effective sensor control may make the tracking task feasible, as well as facilitate its execution. Furthermore, since image acquisition is usually much less costly than the subsequent image analysis, it seems advantageous to dedicate some of the computational effort to perform the sensor planning task. With such control of the imaging process it may be possible to perform a simpler analysis on the resulting image.

The work reported in this paper focuses on sensor planning — how to control the camera to perform the tracking task, given a target recognition algorithm. We introduce the concept of detection function to assess the performance of given recognition algorithms. A detection function specifies the probability that a particular recognition algorithm will detect the target in an image, given the camera viewing direction and angle size and the target position. This function in turn determines the region of space in which the target can be detected with high probability using certain specified camera parameters. For this reason, it provides a natural language for discussion of the task of tracking an object using a camera with adjustable pan, tilt, and zoom. The goal of our tracking strategy is to use the best camera parameter settings while tracking the target in the environment. We describe a novel tracking algorithm wherein the detection function is used to help select the state parameters of the camera during the tracking process, so that the target is kept within the field of view of the camera and the quality of the image is maintained.

The paper is organized as follows. Section 2 describes the detection function and related geometrical issues. Section 3 describes how to select a minimal set of camera states (MCPS) such that if the target is present in the environment then it can be detected using one of these states. Section 4 describes a tracking algorithm which uses the MCPS first to locate and then to track an object moving in a fixed three-dimensional environment.

2. DETECTION FUNCTION

The *detection function* specifies the ability of the recognition algorithm to detect the target, averaged over various factors and conditions that affect its performance. In particular, the detection function $\mathbf{b}(\langle w,h\rangle,\langle\theta,\delta,l\rangle)$ gives the probability of detecting the target by the given recognition algorithm when the viewing angle size of the camera is $\langle w,h\rangle$ and the position of the target relative to the camera is $\langle\theta,\delta,l\rangle$; where $\theta = arctan(\frac{x}{z})$, $\delta = arctan(\frac{y}{z})$, $l = z$, and (x,y,z) are the coordinates of the target center in camera coordinate system. The value of $\mathbf{b}(\langle w,h\rangle,\langle\theta,\delta,l\rangle)$ can be obtained empirically: the target is placed at (θ,δ,l) and experiments are performed under various conditions of light intensity, background situation, target orientation with respect to the camera, etc. The value of $\mathbf{b}$ is given by the number of successful recognitions divided by the total number of experiments.

It is not necessary to record the detection function values for all the different camera viewing angle sizes. Rather, we only need the detection values of one camera angle size (we call it the reference angle), since those for other camera angle sizes can be approximated by the following simple transformation. Suppose we know the detection function values for viewing angle size $\langle w_0,h_0\rangle$. We want to find the detection function values for viewing angle size $<w,h>$. To get the value of $\mathbf{b}(\langle w,h\rangle,\langle\theta,\delta,l\rangle)$ for a given $\langle\theta,\delta,l\rangle$, we need to find values of $\langle\theta_0,\,\delta_0,\,l_0\rangle$ for angle size $\langle w_0,h_0\rangle$ that satisfy the following approximation:

$$\mathbf{b}(\langle w,h\rangle,\langle\theta,\delta,l\rangle) \approx \mathbf{b}(\langle w_0,h_0\rangle,\langle\theta_0,\delta_0,l_0\rangle). \tag{1}$$

The approximation relation (Formula (1)) means that when we use the recognition algorithm to analyze the picture taken with parameters $(\langle w,h\rangle,\langle\theta,\delta,l\rangle)$ and the picture taken with parameters $(\langle w_0,h_0\rangle,\langle\theta_0,\delta_0,l_0\rangle)$, we should get almost the same result. To guarantee this, the images of the target object should be almost the same in both cases, i.e., they must be approximately equal in at least two geometric factors, namely the scale factor and the position factor. The scale factor refers to the size of the projection of the target object on the image plane. The position factor refers to the position on the image plane of the projection of the center of the target object. Typically, the position factor has much less influence than the scale factor.

We use the scale factor to find the value of l_0 when l is given. The sizes of the projection of the target object on the image planes for $(\langle w,h\rangle,\langle\theta,\delta,l\rangle)$ and $(\langle w_0,h_0\rangle,\langle\theta_0,\delta_0,l_0\rangle)$ are approximately determined by l and l_0, respectively. Equality of the scale factors means that for a target patch that is parallel to the image plane, the area of its projection on the image plane for $(\langle w,h\rangle,\langle\theta,\delta,l\rangle)$ should be same as the area of its projection on the image plane for $(\langle w_0,h_0\rangle,\langle\theta_0,\delta_0,l_0\rangle)$. Let W and H denote the width and height of the image plane, respectively. Since the size of the image plane remains constant for different focal lengths, W and H will be same for any focal length. (We assume here that the image plane and the focal plane of the camera are always coincident).

Let S be the area of the patch on the target facing the camera. Let $S^{'}$ be the area of the projected target image for the desired arguments $(\langle w, h\rangle, \langle \theta, \delta, l\rangle)$, and $S_0^{'}$ be the area of the projected target image for the reference arguments $(\langle w_0, h_0\rangle, \langle \theta_0, \delta_0, l_0\rangle)$. From the similarity relation between the target patch and its projected image, it is easy to show that

$$S^{'} = \frac{f^2}{l^2}S. \tag{2}$$

Since

$$tan(\frac{w}{2}) = \frac{\frac{W}{2}}{f},$$

and

$$tan(\frac{h}{2}) = \frac{\frac{H}{2}}{f},$$

we have

$$S^{'} = \frac{f^2}{l^2}S = \frac{WH}{4l^2 tan(\frac{w}{2})tan(\frac{h}{2})}S. \tag{3}$$

Similarly,

$$S_0^{'} = \frac{WH}{4l_0^2 tan(\frac{w_0}{2})tan(\frac{h_0}{2})}S. \tag{4}$$

To guarantee $S^{'} = S_0^{'}$, we get:

$$l_0 = l\sqrt{\frac{tan(\frac{w}{2})tan(\frac{h}{2})}{tan(\frac{w_0}{2})tan(\frac{h_0}{2})}}. \tag{5}$$

We use the position factor to find the values of θ_0, δ_0 when θ and δ are given. Let D denote the center of target patch with respect to $(\langle w, h\rangle, \langle \theta, \delta, l\rangle)$, and let $D^{'}(x^{'}, y^{'}, z^{'})$ denote the image of D on the image plane with respect to $(\langle w, h\rangle, \langle \theta, \delta, l\rangle)$. Similarly, let D_0 be the center of target patch with respect to $(\langle w_0, h_0\rangle, \langle \theta_0, \delta_0, l_0\rangle)$, and let $D_0^{'}(x_0^{'}, y_0^{'}, z_0^{'})$ represent the image of D_0 on the image plane with respect $(\langle w_0, h_0\rangle, \langle \theta_0, \delta_0, l_0\rangle)$, where $(x^{'}, y^{'}, z^{'})$ and $(x_0^{'}, y_0^{'}, z_0^{'})$ are in camera coordinate system. Then we have

$$x^{'} = ftan(\theta) = \frac{\frac{W}{2}}{tan(\frac{w}{2})}tan(\theta) = \frac{W}{2}\frac{tan(\theta)}{tan(\frac{w}{2})}. \tag{6}$$

Similarly,

$$y^{'} = \frac{H}{2}\frac{tan(\delta)}{tan(\frac{h}{2})}, \tag{7}$$

$$x_0^{'} = \frac{W}{2}\frac{tan(\theta_0)}{tan(\frac{w_0}{2})}, \tag{8}$$

and

$$y_0^{'} = \frac{H}{2}\frac{tan(\delta_0)}{tan(\frac{h_0}{2})}. \tag{9}$$

To guarantee $x^{'} = x_0^{'}$ and $y^{'} = y_0^{'}$, we get

$$\theta_0 = arctan[tan(\theta)\frac{tan(\frac{w_0}{2})}{tan(\frac{w}{2})}], \tag{10}$$

and

$$\delta_0 = arctan[tan(\delta)\frac{tan(\frac{h_0}{2})}{tan(\frac{h}{2})}]. \tag{11}$$

Therefore, when we want to find the detection function value for parameters $\langle\theta, \delta, l\rangle$ with respect to the camera angle size $\langle w, h\rangle$, we first find the corresponding $\langle\theta_0, \delta_0, l_0\rangle$, and then retrieve the detection function value for $\mathbf{b}(\langle w_0, h_0\rangle, \langle\theta_0, \delta_0, l_0\rangle)$ from the look up table or from the analytical formula.

Finally, we note that the detection function $\mathbf{b}(\langle w, h\rangle, \langle\theta, \delta, l\rangle)$ involves averaging over variations in the target orientation and the background, when these variations are treated as random or uncontrolled factors. In cases where the effect of target orientation may be great, however, and it is desirable to explicitly account for this effect, we generate a detection function for each distinguishable aspect or face of the target. Suppose the target has m different aspects, $a_1 \ldots a_m$, then we define the value of $\mathbf{b}^a(\langle w, h\rangle, \langle\theta, \delta, l\rangle)$ as the probability of detecting the target when the viewing angle size of the camera is $\langle w, h\rangle$, the relative position of the target is $\langle\theta, \delta, l\rangle$ and the aspect of the target facing the camera is a ($a \in \{a_1, \ldots, a_m\}$).

3. EFFECTIVE VOLUME AND MINIMUM CAMERA SETTINGS

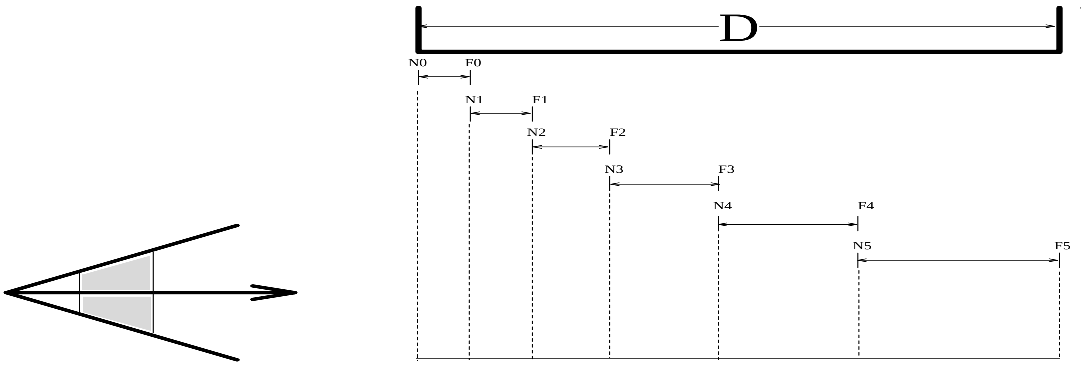
The detection function is used in this section to introduce the concept of effective volume, which defines for given camera parameters the region of space where the target can easily be detected. The notion of effective volume is then used to design a strategy for selection of a minimum set of camera viewing parameters such that their combined effective volumes will cover the whole environment. This set of camera states limits the number of moves the camera must make to locate the target.

3.1. Camera Angle Sizes

First consider the effective range of the camera. Assume the camera is aimed at the target, and let the target aspect facing the camera be a. We are interested in the range of distances from the camera to the target such that the target will be recognized by the recognition algorithm. For a given camera viewing angle size $\langle w, h\rangle$, the value of the detection function $\mathbf{b}^a(\langle w, h\rangle, \langle\theta, \delta, l\rangle)$ depends only on parameters $\langle\theta, \delta, l\rangle$. Since we assume that the target is in the image, the angles θ and δ only determine the target position within the image. It is well known that the target image position has no or very little influence on the recognition results. Thus, we omit the influence of θ, δ and only consider the influence of l. Usually the recognition algorithm can successfully recognize the target only when the image size of the target is such that the whole target can be brought into the field of view of the camera and the features can be detected with the required precision. For a given recognition algorithm, a fixed viewing angle size, and a given target aspect, the probability of successfully recognizing the target is high only when the target is within a certain range. Therefore, different sizes of the viewing angle $\langle w, h\rangle$ will be associated with different **effective ranges** of distance l.

Suppose D is the maximum distance of the camera center to any point in the environment. Our goal is to select a set of angles whose effective ranges will cover the entire depth D of the environment without overlap. Suppose that the biggest viewing angle for the camera is $\langle w_0, h_0\rangle$, and its effective range for the given aspect is $[N_0, F_0]$. We can use geometric constraints to find other required viewing angle sizes $\langle w_1, h_1\rangle$, …, $\langle w_{n_0}, h_{n_0}\rangle$ and their corresponding effective ranges $[N_1, F_1], \ldots, [N_{n_0}, F_{n_0}]$, such that $[N_0, F_0] \bigcup \ldots \bigcup [N_{n_0}, F_{n_0}] \supseteq [N_0, D]$, and $[N_i, F_i) \bigcap [N_j, F_j) = \emptyset$ for $i \neq j$. These $n_0 + 1$ angle sizes are enough to examine the whole depth of the environment with high probability. Figure 1 illustrates the above idea in two-dimensions.

The effective range $[N_{i+1}, F_{i+1}]$ of the next viewing angle $\langle w_{i+1}, h_{i+1}\rangle$ should be adjacent to the effective range of the current viewing angle $\langle w_i, h_i\rangle$, i.e., $N_{i+1} = F_i$. To guarantee that the areas of the images of the target for $\langle w_i, h_i\rangle$ at N_i and F_i are equal to the areas of the images of the target for $\langle w_{i+1}, h_{i+1}\rangle$ at N_{i+1} and F_{i+1}, respectively, we obtain (using Equations (5)):



(a) (b)

Figure 1. Schematic showing selection of the camera angle size in two dimensions. (a) The effective range for a given angle size. (b) The viewing angle sizes should be selected such that their effective ranges can cover the entire depth D without overlap.

$$w_i = 2arctan[(\frac{N_0}{F_0})^i tan(\frac{w_0}{2})] \tag{12}$$

$$h_i = 2arctan[(\frac{N_0}{F_0})^i tan(\frac{h_0}{2})] \tag{13}$$

$$N_i = F_0(\frac{F_0}{N_0})^{i-1}; F_i = F_0(\frac{F_0}{N_0})^i \tag{14}$$

Since $N_i \leq D$, we obtain $i \leq \frac{ln(\frac{D}{F_0})}{ln(\frac{F_0}{N_0})} - 1$. Let $n_0 = \lfloor \frac{ln(\frac{D}{F_0})}{ln(\frac{F_0}{N_0})} - 1 \rfloor$, then the angles that are needed to cover the whole tracking environment for the given aspect are $\langle w_0, h_0 \rangle$, $\langle w_1, h_1 \rangle$, ..., $\langle w_{n_0}, h_{n_0} \rangle$.

3.2. Camera Viewing Directions

The effective ranges of $\langle w_0, h_0 \rangle$, $\langle w_1, h_1 \rangle$, ..., $\langle w_{n_0}, h_{n_0} \rangle$ divide the space around the camera center into a layered sphere. Each layer can be successfully examined by the corresponding effective angle. For a given effective angle size $\langle w, h \rangle$, there are a huge number of viewing directions that can be considered. Each direction $\langle p, t \rangle$ (pan, tilt) corresponds to a rectangular pyramid, which is the viewing volume determined by parameters $\langle w, h, p, t \rangle$. Within this viewing volume, only a slice of the pyramid can be examined with high detection probability by the given recognition algorithm. This slice of the pyramid is the **effective volume** for camera state $\langle w, h, p, t \rangle$. The union of the effective volumes of all the possible $\langle p, t \rangle$ given $\langle w, h \rangle$ will cover the given layer. To examine this layer, it is not necessary to try every possible $\langle p, t \rangle$ one by one — we only need to consider those directions such that the union of their effective volumes cover the whole layer with little overlap. This idea is illustrated in Figure 2.

Let $\alpha = min\{w, h\}$, i.e., the smaller of the width or height of the camera viewing angle. We study in the following a method to select a set of camera viewing directions for angle size $< \alpha, \alpha >$ such that the selected actions can cover the whole sphere. Suppose the viewing direction of the camera (pan and tilt) is $< p, t >$. A patch, on the surface of the sphere, covered by the viewing volume of the camera can be specified by a tilt range $t_b \leq tilt \leq t_e$ and a pan range $p_b \leq pan \leq p_e$. We wish to express these boundaries t_b, t_e, p_b, and p_e in terms of p, t and α.

First, we need to find the range of pan $\Delta_{pan} = p_e - p_b$ for the given viewing direction of the camera. From Figure 3, we can see that any point J on arc $\widehat{AB}$ has a corresponding point E on arc $\widehat{CD}$ such that they have the same value of tilt. The smallest difference of pan for a point on $\widehat{AB}$ and its corresponding point on $\widehat{CD}$ occurs at point A and point D. We take the difference of the range of pan for points A and D as the value of Δ_{pan}. After a detailed calculation, we obtain

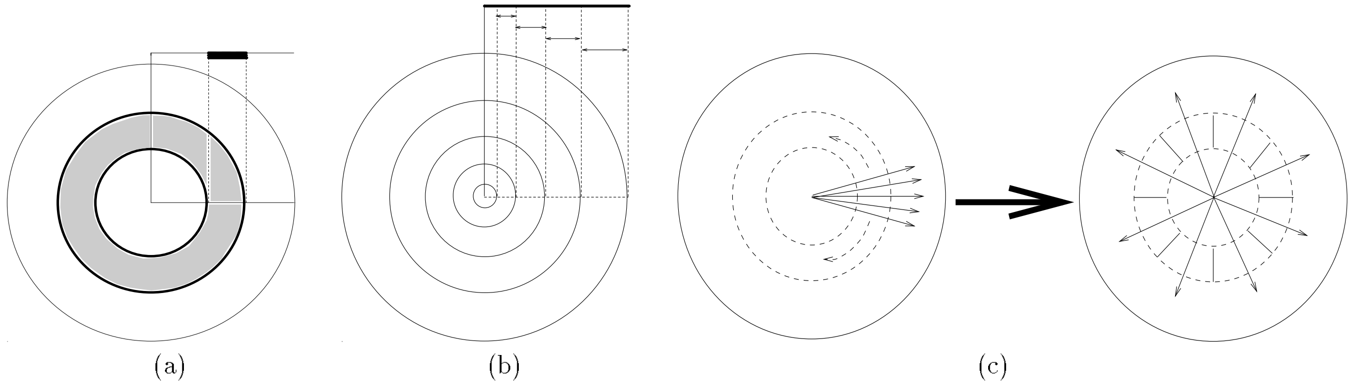


Figure 2. (a) Each viewing angle size can check a layer in space. (b) The selected camera viewing angle sizes divide the space into a layered sphere. (c) Out of the infinite number of possible directions, a limited number suffice for examining a given layer.

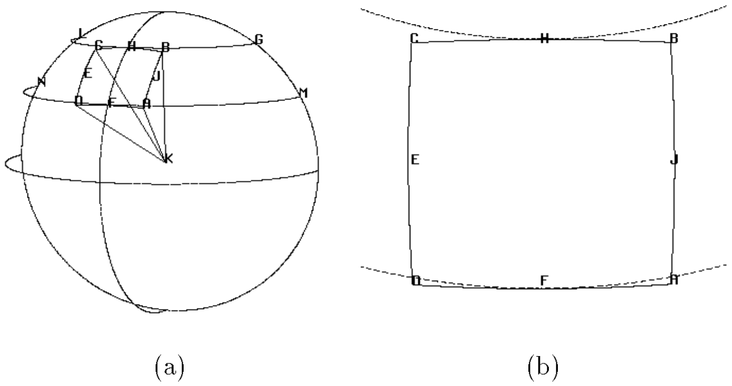


Figure 3. The surface patch cut by the viewing volume. (a) The viewing volume of the camera. Pyramidal section $KABCD$ is the intersection of a sphere centered at the camera center K and the viewing volume of the camera. Big circle arcs $\widehat{AB}$, $\widehat{BC}$, $\widehat{CD}$ and $\widehat{DA}$ are cut by the bounding planes of the viewing volume. Big circle arc $\widehat{HF}$ is cut by the vertical plane of the viewing volume (YZ plane in Camera Coordinate System). Point H is at the middle of big circle arc $\widehat{BC}$, and point F is at the middle of big circle arc AD. Arc $\widehat{NFM}$ is part of the locus of points on the sphere whose tilt is $t+\frac{\alpha}{2}$, and arc $\widehat{LHG}$ is part of the locus of points on the sphere whose tilt is $t-\frac{\alpha}{2}$. Arcs $\widehat{LHG}$ and $\widehat{BC}$ only have one common point H. Arcs $\widehat{NFM}$ and $\widehat{AD}$ only have one common point F. (b) An enlarged view of the surface patch $ABCD$ that is contained in the viewing volume of the camera.

$$\Delta_{pan} = 2arctan\{\frac{sin(\frac{\alpha}{2})}{sin(t+\frac{\alpha}{2})}\}. \tag{15}$$

The viewing volume of a camera aimed at $\langle p, t\rangle$ with angle size $\langle \alpha, \alpha\rangle$, however, does not cover the entire surface patch determined by pan interval Δ_{pan} and the tilt interval between $t-\frac{\alpha}{2}$ and $t+\frac{\alpha}{2}$. From Figure 3(b), we can see that the big circle arc $\widehat{HB}$ gradually goes down (decreases in Z value) as a point moves from H to B. So, there is a part between big circle arc $\widehat{CHB}$ and arc $\widehat{LHG}$ that can not be covered by the viewing volume of the camera when the intended pan range is Δ_{pan}. We need to determine the tilt value of a point as it approaches B from H along the big circle arc $\widehat{HB}$ with the change of pan equal to $\frac{\Delta_{pan}}{2}$. After some calculations, we obtain the tilt value

$$arccos\{\frac{cos(t-\frac{\alpha}{2})}{\sqrt{1+\frac{sin^2(\frac{\alpha}{2})sin^2(t-\frac{\alpha}{2})}{sin^2(t+\frac{\alpha}{2})}}}\}.$$

From the above calculations, it is easy to show that the following surface patch on the sphere can be covered by the viewing volume of a camera with direction $\langle p, t\rangle$ and angle size $\langle \alpha, \alpha\rangle$:

$$p - arctan\{\frac{sin(\frac{\alpha}{2})}{sin(t+\frac{\alpha}{2})}\} \leq pan \leq p + arctan\{\frac{sin(\frac{\alpha}{2})}{sin(t+\frac{\alpha}{2})}\}, \tag{16}$$

$$arccos\{\frac{cos(t-\frac{\alpha}{2})}{\sqrt{1+\frac{sin^2(\frac{\alpha}{2})sin^2(t-\frac{\alpha}{2})}{sin^2(t+\frac{\alpha}{2})}}}\} \leq tilt \leq t + \frac{\alpha}{2}. \tag{17}$$

Suppose the final set of viewing directions is $S_{candidate}$. According to the above calculations, the following algorithm enumerates the viewing directions required for covering the whole sphere.

1. $S_{candidate} = \phi$
2. $p \longleftarrow 0$, $t \longleftarrow 0$, $S_{candidate} = S_{candidate} \bigcup < p, t >$.
3. $t_e \longleftarrow \frac{\pi}{2}$
4. $t_b \longleftarrow arccos\{\frac{cos((t_e-\frac{\alpha}{2})-\frac{\alpha}{2})}{\sqrt{1+\frac{sin^2(\frac{\alpha}{2})sin^2((t_e-\frac{\alpha}{2})-\frac{\alpha}{2})}{sin^2((t_e-\frac{\alpha}{2})+\frac{\alpha}{2})}}}\}$
5. Cover the slice of the spherical surface whose tilt is within the range $[t_b, t_e]$ and the slice whose tilt is within the range of $[\pi - t_e, \pi - t_b]$.
   (a) Let $t \longleftarrow t_e - \frac{\alpha}{2}$.
   (b) let $\Delta_{pan} \longleftarrow 2arctan\{\frac{sin(\frac{\alpha}{2})}{sin((t-\frac{\alpha}{2})+\frac{\alpha}{2})}\}$
   (c) Use Δ_{pan} to divide the range $[0, 2\pi]$ for the given slice into a a series of intervals $[p_b, p_e]$, as follows: $[0, \Delta_{pan}]$, $[\Delta_{pan}, 2\Delta_{pan}]$, …, $[k\Delta_{pan}, 2\pi]$. Note: the length of the last interval may not be Δ_{pan}.
   (d) For each interval, let $p \longleftarrow \frac{p_b+p_e}{2}$ and $S_{candidate} = S_{candidate} \bigcup < p, t > \bigcup < p, \pi - t >$.
6. Let $t_e \longleftarrow t_b$
7. If $t_e \leq \alpha$, stop the process. Otherwise Goto 4.

3.3. Minimum Camera Parameter Settings

For a given target aspect a_i, we can find a set of n_i effective viewing angle sizes $\langle w_{i,1}, h_{i,1}\rangle$, …, $\langle w_{i,n_i}, h_{i,n_i}\rangle$ such that the system can examine the entire depth of the environment. For each angle size $\langle w_{i,j}, h_{i,j}\rangle$, we can find a set of $n_{i,j}$ viewing directions $\langle p_{i,j,1}, t_{i,j,1}\rangle$, …, $\langle p_{i,j,n_{i,j}}, t_{i,j,n_{i,j}}\rangle$ such that the corresponding layer of the sphere is examined for the target. Each triple $V_{ijk} =< a_i; \langle w_{i,j}, h_{i,j}\rangle; \langle p_{i,j,k}, t_{i,j,k}\rangle >$ determines an effective volume such that when the aspect of the target facing the camera is a_i, and the target to be tracked is within V_{ijk}, and the camera state is $\langle w_{i,j}, h_{i,j}\rangle$ and $\langle p_{i,j,k}, t_{i,j,k}\rangle$, then the given recognition algorithm can be expected to detect the target. Let $V_i = \cup_{j=1}^{n_i} \cup_{k=1}^{n_{i,j}} V_{ijk}$, then V_i is the union of all the effective volumes for aspect a_i. If the target is within the environment and has aspect a_i facing the camera, then there exists at least one effective volume $V_{ijk} \in V_i$ that contains the target. The corresponding camera parameters $\langle w_{i,j}, h_{i,j}\rangle$ and $\langle p_{i,j,k}, t_{i,j,k}\rangle$ determine the camera state such that the recognition system can detect the target.

Suppose the target has m aspects. For each aspect a_i there exists a set of effective volumes V_i. Let $V = \cup_{i=1}^{m} V_i$. Then V is the set of the effective volumes for the target. No matter where the target might be in the environment or what aspect it presents to the camera, there exists at least one effective volume that contains the target. The camera parameters corresponding to this effective volume allow the system to detect the target.

When discussing the effective volume, we assumed that the camera must cover the entire spherical region surrounding it. In most situations, however, the environment is only a portion of the sphere with radius equal to the depth of the environment. Let Ω denote the region occupied by the environment. Then during the tracking process, we only need to consider those effective volumes V_{ijk} that have common regions with Ω.

Thus, we define the Minimum Camera Parameter Settings ($MCPS$) as:

$$MCPS = \{\langle w_{i,j}, h_{i,j}, p_{i,j,k}, t_{i,j,k}\rangle \mid 1 \leq i \leq m, 1 \leq j \leq n_i, 1 \leq k \leq n_{i,j}, V_{ijk} \cap \Omega \neq \emptyset\}$$

$MCPS$ is the minimum set of camera parameters needed to track the target within the environment.

4. TRACKING WITH PRE-RECORDED IMAGE DATABSE

4.1. Segmentation by Image Difference

In order to track a target, we must be able to segment the target from the background in the image. Generally this is a very difficult task, but here we propose a simple strategy which will suffice in many applications. The essence of the idea is first to create a database of selected images of the environment in the absence of the target, and then during tracking compare each new image with its corresponding image in the database. Significant differences in this comparison reveal presence of the target.

The comparison between the tracking image and the database image involves computing a series of difference images. We start with a color difference image, from which is calculated a greyscale difference and finally we use a threshold to convert to a binary black and white image. Calculation of the initial color difference image compensates for small mechanical errors (associated with moving the camera to the desired state) by comparing each pixel of the tracking image with its corresponding and surrounding pixels in the database image, and taking the smallest of these differences. To further reduce noise, the final binary image is subjected to a process called erosion and then to dilation to recover signal size. Blobs of connected white pixels in the thus improved binary image are classified according to size and center of mass. Blobs below a certain size are ignored as noise, and the remaining blobs, if any, are considered to be the target.

4.2. Tracking with MCPS Image Database

MCPS gives the minimum set of camera parameters required to track the target within the environment. Before the target enters the scene, we take images of the environment for all camera settings $\langle w, h; p, t\rangle \in MCPS$. These images form the image database IB_{MCPS}. This image database is used when segmenting the target from the background as described above. The target tracking algorithm based on the minimal set of camera states MCPS and the corresponding image database IB_{MCPS} is outlined below.

1. Generate IB_{MCPS}
   (a) $IB_{MCPS} = \emptyset$.
   (b) Foreach $\langle w, h; p, t\rangle \in MCPS$ do
      i. Manipulate the camera to the state $\langle w, h; p, t\rangle$ and take an image $I_{\langle w,h;p,t\rangle}$.
      ii. Add $I_{\langle w,h;p,t\rangle}$ into the image database IB_{MCPS}:
      $$IB_{MCPS} = IB_{MCPS} \cup I_{\langle w,h;p,t\rangle}$$
2. Select the first $\langle w, h; p, t\rangle \in MCPS$.
3. Manipulate the camera to the state $\langle w, h; p, t\rangle$ and take an image $I^*_{\langle w,h;p,t\rangle}$.
4. Compare the image $I^*_{\langle w,h;p,t\rangle}$ just taken with the image $I_{\langle w,h;p,t\rangle}$ in the database.
5. If the two images are same, (i.e., no blobs of significant size are detected in the final difference image), then select the next $\langle w, h; p, t\rangle \in MCPS$ and GoTo 3. Otherwise, the target is detected.
6. Track the target.

(a) Take an image $I^*_{\langle w,h;p,t\rangle}$ according to $\langle w, h; p, t\rangle$.

(b) Compare the image $I^*_{\langle w,h;p,t\rangle}$ just taken and the image $I_{\langle w,h;p,t\rangle}$ in the database IB_{MCPS}.

(c) If the target is detected (the target is still in the image), GoTo 6(a) to repeat the process. Otherwise (there is no difference between $I_{\langle w,h;p,t\rangle}$ and $I^*_{\langle w,h;p,t\rangle}$) the target is not within the effective volume $V_{\langle w,h,p,t\rangle}$

(d) Perform the 'Where to Look Next' task as described below to decide which effective volume might next contain the target.

4.3. Where to Look Next

"Where to Look Next" refers to the task of selecting the next $\langle w^*, h^*, p^*, t^*\rangle \in MPCS$ such that the target can be brought into the field of view of the camera for detection by the given recognition algorithm. This requires predicting the next target aspect and the next effective volume that will contain the target.

Here we assume that the speed to perform the image processing task is much faster than the movement (translation or rotation) of the target. This is true in most situations because the operation to calculate the image difference is very fast. If this assumption is true, then the target can only appear in those effective volumes that surround the current effective volume for the given aspect, and the aspect of the target can only be one of those aspects that surround the current aspect in the aspect graph.[13]

Suppose the current effective volume is $V_{ijk} =< a_i; \langle w_{i,j}, h_{i,j}\rangle; \langle p_{i,j,k}, t_{i,j,k}\rangle >$. The effective volumes that surround $V_{i,j,k}$ are those that are on the same layer of the layered sphere as V_{ijk} (that is, have the same viewing angle size $\langle w_{i,j}, h_{i,j}\rangle$), those that are on the next layer outside the current layer (that is, with angle size $\langle w_{i,j+1}, h_{i,j+1}\rangle$), and those that are on the next layer inside the current layer (that is, with angle size $\langle w_{i,j-1}, h_{i,j-1}\rangle$). These surrounding effective volumes can be easily identified before the tracking process. We call the region occupied by these surrounding effective volumes the **surrounding region** for V_{ijk}.

As discussed above, each aspect gives rise to a set of effective volumes. Suppose that on the aspect graph for the target, the current aspect a_i is adjacent to aspects $a_1^{'}, \ldots, a_r^{'}$. Each aspect $a_l^{'}$ corresponds to a group of effective volumes that tessellate the space surrounding the camera. Some of them have regions in common with the **surrounding region** for V_{ijk}. We call these the related effective volumes of $a_l^{'}$ for V_{ijk} of a_i. The union of the related effective volumes of all the $a_1^{'}, \ldots, a_r^{'}$ is called the related effective volumes for V_{ijk} of aspect a_i. The union of the corresponding camera state parameters is called the **related camera settings** (RCS) for V_{ijk} of a_i, written as RCS_{ijk}. If the target disappears when the effective volume is V_{ijk} then the next camera setting should be selected from the RCS for V_{ijk}.

Before the tracking process, we generate RCS_{ijk} for all the possible effective volumes V_{ijk}. During the tracking process, if the effective volume that lost the target is V_{ijk}, then we check all the actions in RCS_{ijk} to look for the target. It might be possible in some circumstances to give a preferred order for the actions in RCS_{ijk}, based on available information. If the target is found, while scanning RCS_{ijk}, then GoTo 6(a) and continue the tracking process. Otherwise, go back to step (2) to locate anew the target.

5. Conclusion

The task of tracking an object moving in three dimensions is both very important and very challenging. Higher animal life forms, of course, accomplish this task with amazing skill and ease. As yet researchers in computer science are at an early stage in developing a similar ability in robots.

Whereas previous studies have approached this problem from the standpoint of fixed cameras and complex segmentation routines, this paper shifts attention to the topic of active control of the camera viewing direction and angle size. By first considering the detection capability of the recognition algorithm for the various camera states and target aspects, we are able to tessellate the space surrounding the camera into a finite set of effective volumes. The corresponding set of camera states (MCPS) defines a list of actions that can be used to detect and follow the target, if present.

Target detection requires segmentation of the target image from the background image. We propose a simple segmentation strategy which says that a target is present in the current image if and only if there is a sizable difference between the current image and the corresponding image stored in the image database. The image database must, therefore, contain a control image for every camera state in MCPS, but since the size of MCPS is minimal, the image storage requirement is not prohibitive. This strategy makes segmentation almost trivial and implies that the system will track any moving object. Nevertheless, there are applications such as security and surveillance systems where such nondiscriminate tracking is appropriate. For other applications where the appearance of the target is important, the segmentation routine can be augmented to discriminate size, colors and shape. We are in the process of conducting experiments to test the effectiveness of our tracking algorithm, and the results are promising.

REFERENCES

1. B. M. T. Darrell and A. P. Pentland, "Active face tracking and pose estimation in an interactive room," in *CVPR*, pp. 67–71, 1996.
2. D. Gavrila and L. Davis, "3-d model based tracking of humans in action: a multi-view approach," in *CVPR*, pp. 73–79, 1996.
3. H. Graf, E. Cosatto, D. Gibbon, M. Kocheisen, and E. Petajan, "Multi-modal system for locating heads and faces," in *International Conference on Automatic Face and Gesture Recognition*, pp. 88–93, (Killington, Vermont), October 1996.
4. J. N. D. Huttenlocher and W. Rucklidge, "Tracking non-rigid objects in complex scenes," in *ICCV93*, pp. 93–101, 1993.
5. S. X. Ju, M. J. Black, and Y. Yacoob, "Cardboard people: A parameterized model of articulated image motion," in *International Conference on Automatic Face and Gesture Recognition*, pp. 38–44, (Killington, Vermont), October 1996.
6. I. Kakadiaris and D. Metaxas, "Model-based estimation of 3d human motion with occlusion based on active multi-viewpoint selection," in *CVPR*, pp. 81–87, 1996.
7. C. Kervrann and F. Heitz, "A hierarchical statistical framework for the segmentation of deformable objects in image sequences," in *CVPR*, pp. 724–728, 1994.
8. J. Kuch and T. Huang, "Vision based hand modeling and tracking," in *Proceedings of International Conference on Computer VIsion*, pp. 81–87, 1996.
9. B. Moghaddam and A. Pentland, "Probabilistic visual learning for object detection," in *ICCV95*, pp. 786–793, 1995.
10. K. Rohr, "Towards model-based recognition of human movements in image sequences," *CVGIP: Image Understanding* **59**(1), pp. 94–115, 1986.
11. J. Weng, N. Ahuja, and T. Huang, "Learning recognition and segmentation using the cresceptron," in *ICCV93*, pp. 121–128, 1993.
12. C. Wren, A. Azarbayejani, T. Darrell, and A. Pentland, "Pfinder: Real-time tracking of the human body," in *International Conference on Automatic Face and Gesture Recognition*, pp. 51–60, (Killington, Vermont), October 1996.
13. J. Koenderink and A. van Doorn, "The internal representation of solid shape with respect to vision," *Biological Cybernetics* **32**, pp. 211–216, 1979.