Learning an Object's Function by Observing the Object in Action*

Zoran Duric[1,2] Ehud Rivlin[1,3] Azriel Rosenfeld[1]
[1]Computer Vision Laboratory, Center for Automation Research
University of Maryland, College Park, MD 20742-3275
[2]Machine Learning and Inference Laboratory
George Mason University, Fairfax, VA 22030
[3]Department of Computer Science
Israel Institute of Technology—Technion
Haifa, Israel

Abstract

One way to learn the function of an object is by watching the object in use. As an example, an observer might "see" a knife being used to slice bread and learn the function of cutting and the context in which it can be used.

This paper demonstrates that the function of an object can be inferred from its motion. We show that the motion of an object, when combined with information about the object's shape, provides strong constraints on possible functions that the object might be performing. In further studies, currently in progress, we will demonstrate that this approach can be used to learn the functionality of an unknown object by observing an image sequence that shows the object performing an action which accomplishes the function.

1 Introduction

Recognizing the functions of objects is often a prerequisite to interacting with them. The functionality of an object can be defined as the usability of the object for a particular purpose [Bogoni and Bajcsy, 1994].

There has been considerable recent research on the problem of recognizing object functionality; for a short survey see [Bogoni and Bajcsy, 1994]. The goal of this research has been to determine functional capabilities of an object based on characteristics such as shape, physics and causation [Stark and Bowyer, 1992]. Little attention has been given to the problem of determining or learning the functionality of an object from its motion. We believe that motion provides a strong indication of function. In particular, velocity, acceleration, and force of impact resulting from motion strongly constrain possible function. As in other approaches to recognition of function, the object (and in our case, its motion) should not be evaluated in isolation, but in context. The context includes the nature of the agent making use of the object and the frame of reference used by the agent.

In this paper, we address the following problem: How can we use the motion of an object, while it is being used to perform a task, to determine its function? Our method of answering this question is based on extraction of a few motion descriptors from the image sequence. These descriptors are compared with stored descriptors that arise in known motion-to-function mappings to obtain function recognition.

In Section 2 we briefly review related work. In Section 3 we review preliminaries on motion and image motion fields. Section 4 considers the problem of determining the functionality of a known object by analyzing an image sequence showing that object performing the function. The motion estimation machinery needed for this task is developed in Section 5. In Section 6 we present experimental results demonstrating that motion analysis can indeed be used in determining functionality. In Section 7 we discuss planned future work in this area.

2 Related Work

Motion and functionality have appeared in the literature in several contexts. Early work on functional recognition can be found in [Freeman and Newell, 1971; Solina and Bajcsy, 1983; Winston et al., 1983]. More recently, Stark and Bowyer [1991a; 1991b; 1992; Stark et al., 1993] used these ideas to solve some of the problems presented by more traditional model-based methods of object recognition. This work deals only with stationary objects; no motion is involved.

*The support of the Air Force Office of Scientific Research under Grants F49620-93-1-0039 and F49620-95-1-0462 is gratefully acknowledged, as is the help of Sandy German in preparing this paper.

In more recent work Green et al. [1994] discuss the recognition of articulated objects, using motion to determine whether the object possesses the appropriate functional properties.

Gould and Shah [1989] use motion characteristics to identify important events corresponding to changes in direction, speed and acceleration in an object's motion. Motion analysis for recognition of activities was described by Polana and Nelson [1993].

These approaches are not adequate for our purposes since many objects can display similar motion characteristics. An object model is necessary to distinguish the functions of objects from their motion characteristics. Our work is based on segmenting the object into primitive parts (see Section 4.1) and analyzing their motions.

3 Preliminaries

In this section we derive equations of motion for observer-centered and object-centered coordinate systems. We then derive projected motion equations for the weak perspective imaging model [Ullman and Basri, 1991]. Finally, we derive the relationship between the image velocities and the projected motion.

3.1 Rigid Body Motion

To facilitate the derivation of the motion equations of a rigid body $\mathcal{B}$, we use two rectangular coordinate frames, one ($Oxyz$) fixed in space, the other ($Cx_1y_1z_1$) fixed in the body and moving with it. The coordinates of any point P of the body with respect to the moving frame are constant with respect to time t, while the coordinates X, Y, Z of the same point P with respect to the fixed frame are functions of t. The position of the moving frame at any instant is given by the position $\vec{d}_c = (X_c\ Y_c\ Z_c)^T$ of the origin C, and by the nine direction cosines of the axes of the moving frame with respect to the fixed frame. For a given position $\vec{p}$ of P in $Cx_1y_1z_1$ we have the position $\vec{r}_p$ of P in $Oxyz$:

$$\vec{r}_p \equiv (X\ Y\ Z)^T \equiv R\vec{p} + \vec{d}_c \qquad (1)$$

where R is the matrix of direction cosines. The velocity of $\vec{r}_p$ is then given by

$$\dot{\vec{r}}_p = \vec{\omega} \times (\vec{r}_p - \vec{d}_c) + \vec{T}$$

where $\vec{\omega} = (A\ B\ C)^T$ is the rotational velocity of the moving frame; $\dot{\vec{d}}_c = (\dot{X}_c\ \dot{Y}_c\ \dot{Z}_c)^T \equiv (U\ V\ W)^T \equiv \vec{T}$ is the translational velocity of the point C. This can be written as

$$\begin{pmatrix} \dot{X} \\ \dot{Y} \\ \dot{Z} \end{pmatrix} = \begin{pmatrix} 0 & -C & B \\ C & 0 & -A \\ -B & A & 0 \end{pmatrix} \begin{pmatrix} X - X_c \\ Y - Y_c \\ Z - Z_c \end{pmatrix} + \begin{pmatrix} U \\ V \\ W \end{pmatrix}. \qquad (2)$$

Let the rotational velocity in the moving frame be $\vec{\omega}_1 = (A_1\ B_1\ C_1)^T$; we can write $\vec{\omega} = R\vec{\omega}_1$ and $\vec{\omega}_1 = R^T\vec{\omega}$.

3.2 The Imaging Model

Let (X, Y, Z) denote the Cartesian coordinates of a scene point with respect to the fixed camera frame, and let (x, y) denote the corresponding coordinates in the image plane. The equation of the image plane is $Z = f$, where f is the focal length of the camera. The perspective projection is given by $x = fX/Z$ and $y = fY/Z$. For weak perspective projection we need a reference point (X_c, Y_c, Z_c). A scene point (X, Y, Z) is first projected onto the point (X, Y, Z_c); then, through plane perspective projection, the point (X, Y, Z_c) is projected onto the image point (x, y). The projection equations are given by

$$x = \frac{X}{Z_c} f, \quad y = \frac{Y}{Z_c} f. \qquad (3)$$

3.3 The Motion Field and the Optical Flow Field

The instantaneous velocity of the image point (x, y) under weak perspective projection can be obtained by taking derivatives of (3) with respect to time and using (2):

$$\begin{aligned} \dot{x} &= f\frac{\dot{X}Z_c - X\dot{Z}_c}{Z_c^2} = f\frac{[-C(Y-Y_c)+B(Z-Z_c)+U]Z_c - XW}{Z_c^2} \\ &= \frac{Uf - xW}{Z_c} - C(y - y_c) + fB\left(\frac{Z}{Z_c} - 1\right), \qquad (4) \\ \dot{y} &= f\frac{\dot{Y}Z_c - Y\dot{Z}_c}{Z_c^2} = f\frac{[C(X-X_c)-A(Z-Z_c)+V]Z_c - YW}{Z_c^2} \\ &= \frac{Vf - yW}{Z_c} + C(x - x_c) - fA\left(\frac{Z}{Z_c} - 1\right) \qquad (5) \end{aligned}$$

where $(x_c, y_c) = (fX_c/Z_c, fY_c/Z_c)$ is the image of the point C. Let $\vec{\imath}$ and $\vec{\jmath}$ be the unit vectors in the x and y directions, respectively; $\dot{\vec{r}} = \dot{x}\vec{\imath} + \dot{y}\vec{\jmath}$ is the projected motion field at the point $\vec{r} = x\vec{\imath} + y\vec{\jmath}$.

If we choose a unit direction vector $\vec{n}_r$ in the image point $\vec{r}$ and call it the normal direction, then the *normal motion field* at $\vec{r}$ is $\dot{\vec{r}}_n = (\dot{\vec{r}} \cdot \vec{n}_r)\vec{n}_r$. $\vec{n}_r$ can be chosen in various ways; the usual choice is the direction of the image intensity gradient.

Let $I(x, y, t)$ be the image intensity function. The time derivative of I can be written as

$$\begin{aligned} \frac{dI}{dt} &= \frac{\partial I}{\partial x}\frac{dx}{dt} + \frac{\partial I}{\partial y}\frac{dy}{dt} + \frac{\partial I}{\partial t} \\ &= (I_x\vec{\imath} + I_y\vec{\jmath}) \cdot (\dot{x}\vec{\imath} + \dot{y}\vec{\jmath}) + I_t \\ &= \nabla I \cdot \dot{\vec{r}} + I_t \end{aligned}$$

where ∇I is the image gradient and the subscripts denote partial derivatives.

If we assume $dI/dt = 0$, i.e. that the image intensity does not vary with time [Horn and

Schunck, 1981], then we have $\nabla I \cdot \vec{u} + I_t = 0$. The vector field $\vec{u}$ in this expression is called the *optical flow*. If we choose the normal direction $\vec{n}_r$ to be the image gradient direction, i.e. $\vec{n}_r \equiv \nabla I / \|\nabla I\|$, we then have

$$\vec{u}_n = (\vec{u} \cdot \vec{n}_r)\vec{n}_r = \frac{-I_t \nabla I}{\|\nabla I\|^2} \quad (6)$$

where $\vec{u}_n$ is called the *normal flow*.

It was shown in [Verri and Poggio, 1987] that the magnitude of the difference between $\vec{u}_n$ and the normal motion field $\vec{r}_n$ is inversely proportional to the magnitude of the image gradient. Hence $\vec{r}_n \approx \vec{u}_n$ when $\|\nabla I\|$ is large. Equation (6) thus provides an approximate relationship between the 3-D motion and the image derivatives. We will use this approximation later in this paper.

4 Function from Motion

4.1 Primitive shapes and primitive motions

Following [Biederman, 1985; Rivlin et al., 1994; Rivlin et al., 1993] we regard objects as composed of primitive parts. On the most coarse level we consider four types of primitive parts: sticks, strips, plates, and blobs, which differ in the values of their relative dimensions. As in [Rivlin et al., 1994] we let a_1, a_2, and a_3 represent the length, width, and height, respectively, of a volumetric part. We can then define the four classes as follows:

$$\textit{Stick}: \quad a_1 \simeq a_2 \ll a_3 \vee a_1 \simeq a_3 \ll a_2 \vee a_2 \simeq a_3 \ll a_2 \quad (7)$$

$$\textit{Strip}: \quad a_1 \neq a_2 \wedge a_2 \neq a_3 \wedge a_1 \neq a_3 \quad (8)$$

$$\textit{Plate}: \quad a_1 \simeq a_2 \gg a_3 \vee a_1 \simeq a_3 \gg a_2 \vee a_2 \simeq a_3 \gg a_2 \quad (9)$$

$$\textit{Blob}: \quad a_1 \simeq a_2 \simeq a_3 \quad (10)$$

If all three dimensions are about the same, we have a blob. If two are about the same, and the third is very different, we have two cases: if the two are bigger than the one, we have a plate, and in the reverse case we have a stick. When no two dimensions are about the same we have a strip. For example, a knife blade is a strip, because no two of its dimensions are similar.

Primitives can be combined to create compound objects. In [Rivlin et al., 1994] the different qualitative ways in which primitives can be combined were described—for example, end to end, end to side, end to edge, etc. In addition to specifying the two attachment surfaces participating in the junction of two primitives, we could also consider the angles at which they join, and classify the joints as perpendicular, oblique, tangential, etc. Another refinement would be to describe qualitatively the position of the joint on each surface; an attachment can be near the middle, near a side, near a corner, or near an end of the surface. We can also specialize the primitives by adding qualitative features such as axis shape (straight or curved), cross-section size (constant or tapered), etc.

Functional recognition is based on compatibility with some action requirement. Some basic "actions" are static in nature (supporting, containing, etc.), but many actions involve using an object while it is moving. To illustrate the ways in which one can interact with a primitive, consider the action of "cutting" with a sharp strip or plate. Here a sharp edge is interacting with a surface. The interaction can be described from a kinematic point of view. The direction of motion of the primitive relative to its axis defines the type of action—for example, stabbing, slicing or chopping. These actions all involve primitive motions, which we define to be motions (translations or rotations) along, or perpendicular to, the main axes of the primitive object.[1] In this paper we will use the detection of primitive motions of an object to infer the object's function.

4.2 Inferring Function from Primitive Motions

Given a moving object as seen by an observer, we would like to infer the function being performed by the object. The object is given as a collection of primitives. For example, a knife can be described as consisting of two primitives: a handle (a stick) and a blade (a strip). Given this model, the system estimates the pose of the object (as in [DeMenthon and Davis, 1995; Rivlin et al., 1994]) and passes this information to the motion estimation module. The model and the results of the motion estimation enable the system to infer the function that is being performed by the object.

The function being performed by the object depends on the object's motion in the object's coordinate system and on its relation to the object it acts on (the "actee"; in [Kise et al., 1993; Kitahashi et al., 1991], called the "functant"). This information gives us the relationship between the direction of motion, the main axis of the object, and the surface of the actee, and these relationships determine the intended function. For example, we would expect the motion of a knife that is being used to "stab" to be parallel to the main axis of the knife, whereas if the knife is being used to "chop" we would expect motion perpendicular to the main axis. In both cases, the motion is perpendicular to the surface of the actee. If the knife is being used to slice,

[1] It is interesting to note that motions along the main axis of a primitive preserve "degenerate views" [Kender and Freudenstein, 1987].

we would expect back-and-forth motion parallel to its main axis and also parallel to the surface of the actee.

5 Motions of Sticks and Strips

5.1 The Motion

Consider a moving object $\mathcal{B}$. There is an *ellipsoid of inertia* associated with $\mathcal{B}$. The center of the ellipsoid is at the center of mass C of $\mathcal{B}$; the axes of the ellipsoid are called the *principal axes*. We associate the coordinate system $Cx_1y_1z_1$ with the ellipsoid and choose the axes of $Cx_1y_1z_1$ to be parallel to the principal axes. Let $\vec{i}_1$ be the unit vector in the direction of the longest axis l_c (this axis corresponds to the smallest principal moment of inertia); let $\vec{k}_1$ be the unit vector in the direction of the shortest principal axis (this axis corresponds to the largest moment of inertia); and let $\vec{j}_1$ be the unit vector in the direction of the remaining principal axis with the direction chosen so that the vectors $(\vec{i}_1, \vec{j}_1, \vec{k}_1)$ form a right-handed coordinate system.

In this paper we consider only objects that are approximately planar, straight strips and sticks. For a planar strip the axis of the maximal moment of inertia is orthogonal to the plane of the strip; if the strip is approximately straight, the axis of the minimal moment of inertia is approximately parallel to the medial axis l_c of the strip. In the case of a straight stick, similarly, l_c corresponds to the longest principal axis of the ellipsoid of inertia; the other two principal axes are orthogonal to l_c and can be chosen arbitrarily. We assume that the motion of the stick or strip is planar and that the plane is "visible" to the observer.[2] When the object is a strip we assume that the motion is in the plane of the strip; the translational velocity is then parallel to the plane of the strip and the rotational velocity is orthogonal to the plane of the strip. When the object is a stick the consecutive positions of the stick define the motion plane; the translational velocity lies in the plane and the rotational velocity is orthogonal to the plane. In this case we choose the axis of minimal moment of inertia to be orthogonal to the plane of the motion.

We choose the center of mass C of a stick or a strip $\mathcal{B}$ as the origin of the object coordinate system $Cx_1y_1z_1$; the coordinates of C expressed in the fixed frame are (X_c, Y_c, Z_c). We choose the unit vector $\vec{i}_1$ along l_c with the orientation chosen to be in the direction of the acting part of the tool; we choose $\vec{k}_1$ to be orthogonal to the plane of motion and pointing away from the observer (camera) so that $\vec{k} \cdot \vec{k}_1 \geq 0$. We choose the direction of $\vec{j}_1$ so that $Cx_1y_1z_1$ is a right-handed orthogonal coordinate system. Let Π_y be the plane in which both the line l_c and $\vec{j}$ (the unit vector in the direction of the y-axis of the camera) lie; we can obtain Π_y by sliding a line parallel to $\vec{j}$ along l_c. Also, let Π_z be the plane in which both the line l_c and $\vec{k}$ (the unit vector in the direction of the z-axis of the camera) lie; we can obtain Π_z by sliding a line parallel to $\vec{k}$ along l_c.

Let the angle between the plane Π_y and the Cy_1 axis of the object be ψ. The rotation $R_{x_1}(-\psi)$ around the Cx_1 axis of the object transforms $\vec{j}_1$ into $\vec{j}_c$ (the unit vector parallel to Π_y) and $\vec{k}_1$ into $\vec{k}_c$. The orthographic image of l_c in the plane $Z = Z_c$ is the line l'_c which is the intersection of the plane $Z = Z_c$ and Π_z; let the angle between l'_c and l_c be φ. The rotation $R_{y_c}(-\varphi)$ around an axis Cy_c (passing through C and parallel to $\vec{j}_c$) transforms $\vec{i}_1$ into $\vec{i}_c$ (the unit vector along l'_c) and it transforms $\vec{k}_c$ into $\vec{k}$ (the unit vector along the z-axis of the camera). Finally, let the angle between the positive direction of the x-axis of the camera and the direction $\vec{i}_c$ be α. The rotation $R_z(-\alpha)$ around the axis Cz (passing through C and parallel to $\vec{k}$) transforms $\vec{i}_c$ into $\vec{i}$ and it transforms $\vec{j}_c$ into $\vec{j}$. The rotation matrix $R = R_z(-\alpha)R_{y_c}(-\varphi)R_{x_1}(-\psi)$ in (1) is then given by

$$R = \begin{pmatrix} \cos\alpha & -\sin\alpha & 0 \\ \sin\alpha & \cos\alpha & 0 \\ 0 & 0 & 1 \end{pmatrix} \cdot \begin{pmatrix} \cos\varphi & 0 & \sin\varphi \\ 0 & 1 & 0 \\ -\sin\varphi & 0 & \cos\varphi \end{pmatrix} \begin{pmatrix} 1 & 0 & 0 \\ 0 & \cos\psi & -\sin\psi \\ 0 & \sin\psi & \cos\psi \end{pmatrix} \quad (11)$$

5.2 The Image Motion Field

By our assumption about the translational velocity of the object and the choice of the object coordinate system we have $\vec{T}_1 = (U_1 \quad V_1 \quad 0)^T$ and $\vec{T} = R\vec{T}_1$. The expression for the translational velocity in the fixed frame is given by

$$\vec{T} = \begin{pmatrix} U \\ V \\ W \end{pmatrix} = R_z(-\alpha) \begin{pmatrix} U_1\cos\varphi + V_1\sin\varphi\sin\psi \\ V_1\cos\psi \\ -U_1\sin\varphi + V_1\cos\varphi\sin\psi \end{pmatrix}. \quad (12)$$

Similarly, for the rotational velocity we have $\vec{\omega}_1 = C_1\vec{k}_1$. The expression for $\vec{k}_1$ in the $Oxyz$ frame is $R\vec{k}_1$. We have from (11)

$$R\vec{k}_1 = \begin{pmatrix} \cos\alpha\sin\varphi\cos\psi + \sin\alpha\sin\psi \\ \sin\alpha\sin\varphi\cos\psi - \cos\alpha\sin\psi \\ \cos\varphi\cos\psi \end{pmatrix} \equiv \begin{pmatrix} N_x \\ N_y \\ N_z \end{pmatrix} \equiv \vec{N}.$$

The expression for the rotational velocity in the fixed frame is given by

$$\vec{\omega} = (A \quad B \quad C)^T = C_1 R\vec{k}_1 = C_1\vec{N}. \quad (13)$$

[2] The "visibility" constraint allows an oblique view as long as the angle between the surface normal and the z-axis of the camera is $\leq 30^\circ$ (say).

We now consider the term $(Z - Z_c)/Z_c$ for the points on the object $\mathcal{B}$. The equations we derive are valid for points in the plane in which l_c lies; the unit vector $\vec{k}_1$ is normal to this plane. The equation (in the $Oxyz$ frame) of the plane orthogonal to $\vec{N} = R\vec{k}_1$ in which the point (X_c, Y_c, Z_c) lies is given by

$$(X - X_c)N_x + (Y - Y_c)N_y + (Z - Z_c)N_z = 0.$$

Multiplying by $f(Z_c N_z)^{-1}$ and using (3) we obtain

$$f\frac{Z - Z_c}{Z_c} = -(x - x_c)N_x/N_z - (y - y_c)N_y/N_z. \quad (14)$$

This is an exact formula for thin planar strips; in the case of sticks this formula is exact for an occluding contour.

From (4)–(5) and (14) we obtain the equations of projected motion for points on $\mathcal{B}$ under weak perspective:

$$\begin{aligned}\dot{x} &= \frac{Uf - xW}{Z_c} - C_1(y - y_c)N_z - C_1 \cdot \\ &\quad [(x - x_c)N_x N_y/N_z + (y - y_c)N_y^2/N_z], \quad (15)\\ \dot{y} &= \frac{Vf - yW}{Z_c} + C_1(x - x_c)N_z + C_1 \cdot \\ &\quad [(x - x_c)N_x^2/N_z + (y - y_c)N_x N_y/N_z]. \quad (16)\end{aligned}$$

Equations (15)–(16) relate the image (projected) motion field and (x_c, y_c) to the scaled translational velocity $Z_c^{-1}\vec{T} = Z_c^{-1}(U\ V\ W)^T$, the rotational parameter C_1, and the normal to the strip $\vec{N} = (N_x\ N_y\ N_z)^T$.

Given the point $\vec{r} = x\vec{\imath} + y\vec{\jmath}$ and the normal direction $\vec{n} = n_x\vec{\imath} + n_y\vec{\jmath}$ we have from (15)–(16) the normal motion field

$$\begin{aligned}\dot{\vec{r}}_n \cdot \vec{n} &= n_x\dot{x} + n_y\dot{y}\\ &= n_x f[U/Z_c + (x_c/f)\,C_1 N_x N_y/N_z]\\ &\quad -n_x x(W/Z_c + C_1 N_x N_y/N_z)\\ &\quad -n_x(y - y_c)C_1(N_z + N_y^2/N_z)\\ &\quad +n_y f[V/Z_c - (y_c/f)C_1 N_x N_y/N_z]\\ &\quad -n_y y(W/Z_c - C_1 N_x N_y/N_z)\\ &\quad +n_y(x - x_c)C_1(N_z + N_x^2/N_z) \quad (17)\end{aligned}$$

Let

$$\mathbf{a} = \begin{pmatrix} a_1\\ a_2\\ a_3\\ a_4\\ a_5\\ a_6 \end{pmatrix} \equiv \begin{pmatrix} n_x f\\ -n_x x\\ -n_x(y - y_c)\\ n_y f\\ -n_y y\\ n_y(x - x_c) \end{pmatrix},$$

$$\mathbf{c} = \begin{pmatrix} c_1\\ c_2\\ c_3\\ c_4\\ c_5\\ c_6 \end{pmatrix} \equiv \begin{pmatrix} U/Z_c + (x_c/f)\,C_1 N_x N_y/N_z\\ W/Z_c + C_1 N_x N_y/N_z\\ C_1(N_z + N_y^2/N_z)\\ V/Z_c - (y_c/f)C_1 N_x N_y/N_z\\ W/Z_c - C_1 N_x N_y/N_z\\ C_1(N_z + N_x^2/N_z) \end{pmatrix}. \quad (18)$$

Using (18) we can write (17) as

$$\dot{\vec{r}}_n \cdot \vec{n} = \mathbf{a}^T\mathbf{c}. \quad (19)$$

Column vector $\mathbf{a}$ consists of observable quantities only, while column vector $\mathbf{c}$ contains quantities which are not directly observable from images. To estimate $\mathbf{c}$ we need estimates of $\dot{\vec{r}}_n \cdot \vec{n}$ at six or more image points.

5.3 Estimating the Motion Parameters from Normal Flow

If we use the spatial image gradient as the normal direction $\vec{n}_r \equiv \nabla I/\|\nabla I\| = n_x\vec{\imath} + n_y\vec{\jmath}$, and assume that $\dot{\vec{r}}_n \approx \vec{u}_n$, we can obtain an approximate equation by replacing the left hand side of (19) by normal flow $-I_t/\|\nabla I\|$. In this way we obtain one approximate equation in the six unknown elements of $\mathbf{c}$. For each point (x_i, y_i), $i = 1, \ldots, m$ of the image at which $\|\nabla I(x_i, y_i, t)\|$ is large we can write one equation. If we have more than six points we have an over-determined system of equations $A\mathbf{c} \approx \mathbf{b}$; the rows of the $m \times 6$ matrix A are the vectors $\mathbf{a}_i$. and the elements of the m-vector $\mathbf{b}$ are $-(\partial I(x_i, y_i, t)/\partial t)/\|\nabla I(x_i, y_i, t)\|$.

We seek the solution for which $\|\mathbf{b} - A\mathbf{c}\|$ is minimal. This solution is the same as the solution of the system $A^T A\mathbf{c} = A^T\mathbf{b} \equiv \mathbf{d}$. We solve the system $A^T A\mathbf{c} = \mathbf{d}$ using the Cholesky decomposition. Since the matrix $A^T A$ is a positive definite 6×6 matrix there exists a lower triangular matrix L such that $L\,L^T = A^T A$. We then have $L\,L^T\mathbf{c} = \mathbf{d}$. We solve two triangular systems $L\mathbf{e} = \mathbf{d}$ and $L^T\mathbf{c} = \mathbf{e}$ to obtain the parameter vector $\mathbf{c}$.

After estimating $\mathbf{c}$ we can use (18) to obtain $\vec{T}/Z_c$ and C_1: Let $c_7 = (c_2 - c_5)/2$; we then have

$$\frac{U}{Z_c} = c_1 - \frac{x_c c_7}{f}, \quad \frac{V}{Z_c} = c_4 + \frac{x_c c_7}{f},$$

$$\frac{W}{Z_c} = \frac{c_2 + c_5}{2}, \quad C_1 = \mathrm{sgn}(c_6)\sqrt{c_3 c_6 - c_7^2}$$

where sgn is the sign function.

We will next show how U_1/Z_c and V_1/Z_c can be estimated from $(U/Z_c, V/Z_c, W/Z_c)$. From (12) we have

$$\begin{aligned} &Z_c^{-1}\begin{pmatrix} U_1\cos\varphi + V_1\sin\varphi\sin\psi\\ V_1\cos\psi\\ -U_1\sin\varphi + V_1\cos\varphi\sin\psi \end{pmatrix}\\ &= R_z(\alpha)\begin{pmatrix} U/Z_c\\ V/Z_c\\ W/Z_c \end{pmatrix} \equiv \begin{pmatrix} d_1\\ d_2\\ d_3 \end{pmatrix} \end{aligned} \quad (20)$$

and by rearrangement we obtain

$$\frac{V_1}{Z_c}\cos\psi = d_2, \quad \begin{pmatrix} U_1/Z_c \\ (V_1/Z_c)\sin\psi \end{pmatrix} = \begin{pmatrix} \cos\varphi & -\sin\varphi \\ \sin\varphi & \cos\varphi \end{pmatrix} \begin{pmatrix} d_1 \\ d_3 \end{pmatrix}. \tag{21}$$

To estimate U_1/Z_c, V_1/Z_c, φ, and ψ we need at least four equations, but (21) provides only three. However, by our assumption about the slant of the plane of the motion relative to the image plane, $\angle(\vec{k}_1, \vec{k})$ is at most 30°. The first and the second rotations in (11) are in orthogonal planes; it follows (from the fact that in a right triangle the longest side is the hypotenuse) that both φ and ψ must be smaller than 30°.

Since we have four variables and only three equations we seek φ and ψ for which $|\varphi| + |\psi|$ is minimal. From (21) we have

$$\begin{aligned} d_2 \tan\psi &= d_1 \sin\varphi + d_3 \cos\varphi \\ &\equiv \sqrt{d_1^2 + d_3^2}\,\sin(\varphi - \varphi_0). \end{aligned} \tag{22}$$

where $\varphi_0 = -\arctan(d_3/d_1)$. The value of φ which satisfies (22) and minimizes $|\varphi| + |\psi|$ belongs to the interval $[0, \varphi_0]$ (the interval can be cropped if it exceeds the 30° bound). Each value of φ corresponds to one value of ψ. Because of the convexity of the constraint the solution to $\min\{|\varphi| + |\psi|\}$ can be found using simple search through all $\varphi \in [0, \varphi_0]$ and corresponding ψs. The values of φ and ψ can then be used in (21) to find U_1/Z_c and V_1/Z_c.

6 Experiments

This section illustrates how our methods can be applied to real image sequences. In each sequence, we observed the motion of a tool (a knife) performing a task. The vision system took images at 25 frames per second for 5 seconds, yielding 125 images per experiment. After each image sequence was recorded, a representative sampling of the 125 images was used for further processing. Eleven evenly spaced samples, each composed of three consecutive images, were used.[3] This resulted in 33 images for each experiment.

In our experiments we assumed a table-top scenario, with a stationary observer on one side of the table. Based on this assumption we used a coordinate system that was fixed to the center of the image, with the X axis horizontal and pointing toward the right side of the image, the Y axis pointing upward, and the Z axis chosen to yield a right-handed coordinate frame (pointing toward the scene). All measurements were made relative to this coordinate system. The focal length f of the camera was 550 (pixels).

In Section 6.1 we describe the method which we use to estimate the direction of the medial axis α and the center of mass (x_c, y_c) of the image of the knife; we also define the parameters used to describe the motion of sticks and strips. In the remaining subsections we illustrate how motion can be used to discriminate between different functionalities of a knife.

6.1 Parameterizing the Motion of a Stick or Strip

We have assumed that an approximate direction (right, left, up, down) of the acting part of the tool (the knife blade) is known. The exact direction of the medial axis is found using the following algorithm:

1 - Make a sorted (circular) list of all edge elements (sorted by their orientations modulo π) for which the normal flow is computed.

2 - Find the shortest segment $[\gamma_1, \gamma_2]$ such that more than 3/4 of the orientations in the list are contained within it.

3 - Find the median orientation α in the sorted sublist chosen in the previous step.

4 - If α does not agree with the general direction of the tool (right, left, up, down) then $\alpha \leftarrow \alpha + \pi$.

5 - Use α as the orientation of the medial axis.

We estimated (x_c, y_c) — the image position of C (the reference point and the center of mass of the object)—as the average of the coordinates of all edge points for which the normal flow was computed.

We define β as the angle between the vector $(U_1 \;\; V_1 \;\; 0)^T$ and the Cx_1 axis of the tool coordinate system; thus

$$\beta = \arctan\frac{V_1}{U_1}. \tag{23}$$

We define θ to be the total rotation angle as a function of time:

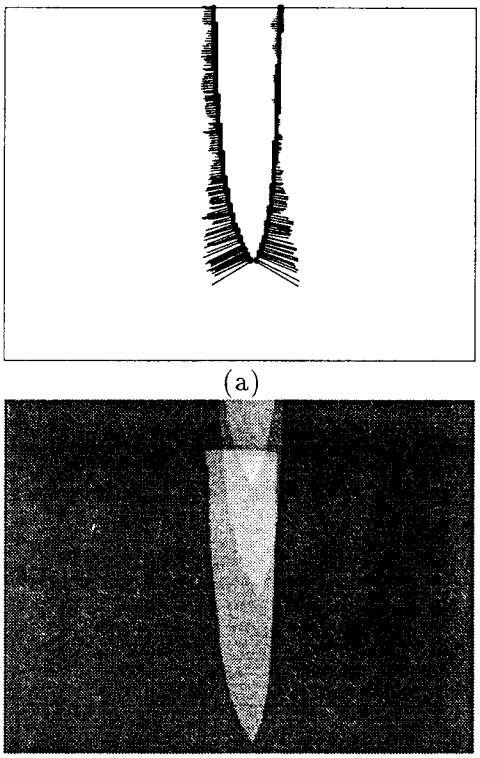
$$\theta = \int_0^t C_1 dt. \tag{24}$$

We use the triples (α, β, θ) to parameterize the motions of sticks and strips.

6.2 Recognition of Stabbing, Chopping, and Slicing

Three simple functions performed by knives are stabbing, chopping, and slicing. We now show how motion can be used to differentiate between the three.

[3] For instance, samples 1 and 2 in any given experiment used images 0–2 and 10–12, respectively.



(a)

(b)

Figure 1: (a) Flow vectors for Stabbing. (b) Stabbing motion.

6.2.1 Stabbing

Stabbing is defined as the cutting motion of a knife in which α (the angle between the projection of l_c onto the plane $Z = Z_c$ and the Ox axis) is close to either $-\pi/2$ or $\pi/2$, β is approximately 0, and θ is small and approximately constant.

Figure 1 shows the flow vectors taken from the 6th sample and a composite image of the knife taken from the 1st, 6th and 11th samples of the stabbing experiment. Figure 2 shows a plot of the triple (α, β, θ) with respect to time (frame numbers). We see that as was expected, the values of α are very close to $-\pi/2$, β is close to 0, and θ is close to 0.

6.2.2 Chopping

Chopping is defined as the cutting motion of a knife in which α (the angle between the projection of l_c onto the plane $Z = Z_c$ and the Ox axis) is close to either 0 or π, β is close to $\pi/2$ ($\alpha \approx \pi$) or $-\pi/2$ (when $\alpha \approx 0$), and θ is small and approximately constant.

Figure 3 shows the flow vectors taken from the 6th sample and a composite image of the knife taken from the 1st, 6th and 11th samples of the chopping experiment. Figure 4 shows a plot of the triple (α, β, θ) with respect to time (frame numbers). We see that, as was expected, the values of α are very close to 0, β is close to $-\pi/2$, and θ is close to 0.

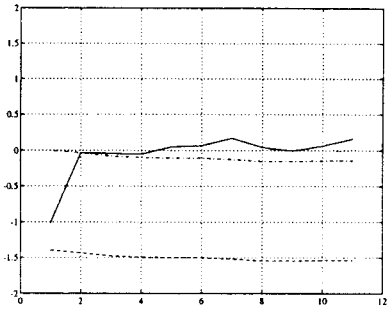


Figure 2: Angles α, β, and θ for Stabbing. α is given by a dashed line, β is given by a solid line, and θ is given by a dash-dot line.

6.2.3 Slicing

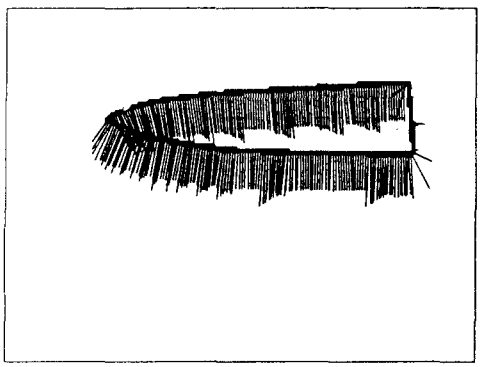
Slicing is defined as the cutting motion of a knife in which α is approximately 0 (or $< \pi/2$), β oscillates between approximately 0 and approximately π, and θ is small and approximately constant.

Figure 5 shows the flow vectors taken from the 6th sample and a composite image of the knife taken from the 1st, 6th and 11th samples of the slicing experiment. (The mass of vectors at the left end of Figure 5(a) come from the motion of the hand, which is visible in the images.) Figure 6 shows a plot of the triple (α, β, θ) with respect to time (frame numbers). We see that, as was expected, the values of α are very close to 0, and that β oscillates between approximately $\pi/2$ and approximately $-3\pi/2$ (note that the two approximate values differ by π).

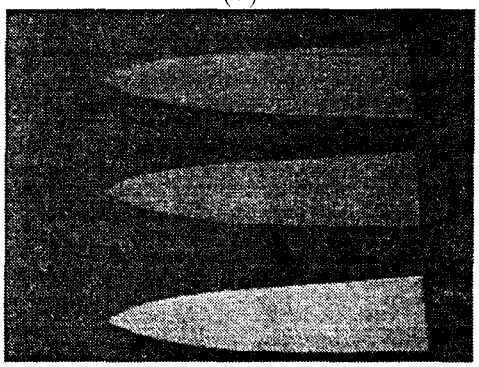
7 Concluding Remarks

Perceiving function from motion provides an understanding of the way an object is being used by an agent. To accomplish this we combined information about the shape of the object, its motion, and its relation to the actee (the object it is acting on). Assuming a decomposition of the object into primitive parts, we analyzed a part's motion relative to its principal axes. Primitive motions (translation and rotation relative to the principal axes of the object) were dominating factors in the analysis. We used a frame of reference relative to the actee. Once such a frame is established, it can have major implications for the functionality of an action.

Several image sequences were used to demonstrate our approach. In the three sequences



(a)



(b)

Figure 3: (a) Flow vectors for Chopping. (b) Chopping motion.

shown in Section 6, motion was used to discriminate between three cutting actions: stabbing, chopping, and slicing. In other sequences, not shown here [Duric et al., 1996], we used motion information to differentiate between two different functionalities of the same object: scooping and hitting with a shovel, and hammering and tightening with a wrench.

Natural extensions of this work include the analysis of more complex objects. Complexity can be expressed in terms of either the shapes of the parts or the way in which the parts are connected. An interesting area is the analysis of articulated objects. The different types of connections between the parts constrain the possible relative motions of the parts. A pair of pliers or a pair of scissors is a simple case, with only a single articulated connection (one degree of freedom in the relative motion of the parts).

Work is in progress in which the methods developed in this paper are used to demonstrate how to learn the functionality of an unknown object by observing image sequences in which the object is performing actions which accomplish its function.

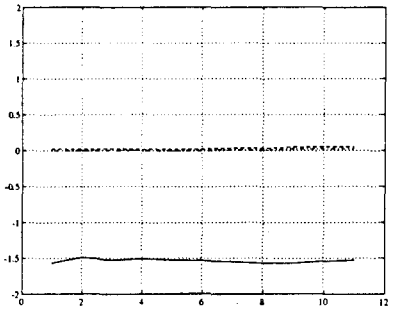


Figure 4: Angles α, β, and θ for Chopping. α is given by a dashed line, β is given by a solid line, and θ is given by a dash-dot line.

References

[Biederman, 1985] I. Biederman. Human image understanding: Recent research and a theory. *Computer Vision, Graphics and Image Processing*, 32:29–73, 1985.

[Bogoni and Bajcsy, 1994] L. Bogoni and R. Bajcsy. Active investigation of functionality. In *Proceedings of the Workshop on Visual Behaviors*, 1994.

[DeMenthon and Davis, 1995] D. DeMenthon and L. Davis. Model-based object pose in 25 lines of code. *International Journal of Computer Vision*, 15:123–141, 1995.

[Duric et al., 1996] Z. Duric, J. Fayman, and E. Rivlin. Function from motion. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, to appear.

[Freeman and Newell, 1971] P. Freeman and A. Newell. A model for functional reasoning in design. In *Proceedings of the International Joint Conference on Artificial Intelligence*, pages 621–640, 1971.

[Green et al., 1994] K. Green, D. Eggert, L. Stark and K. Bowyer. Generic recognition of articulated objects by reasoning about functionality. In *Proceeding of the AAAI-94 Workshop on Represening and Reasoning about Device Function*, pages 56–64, 1994.

[Gould and Shah, 1989] K. Gould and M. Shah. The trajectory primal sketch: A multi-scale scheme for representing motion characteristics. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 79–85, 1989.

[Horn and Schunck, 1981] B.K.P. Horn and B.G. Schunck. Determining optical flow. *Artificial Intelligence*, 17:189–203, 1981.

[Kender and Freudenstein, 1987] J.R. Kender and D.G. Freudenstein. What is a degenerate view? In *Proceedings of the DARPA Image Understanding Workshop*, pages 589–598, 1987.

[Kise et al., 1993] K. Kise, H. Hattori, T. Kitahashi and K. Fukunaga. Representing and recognizing

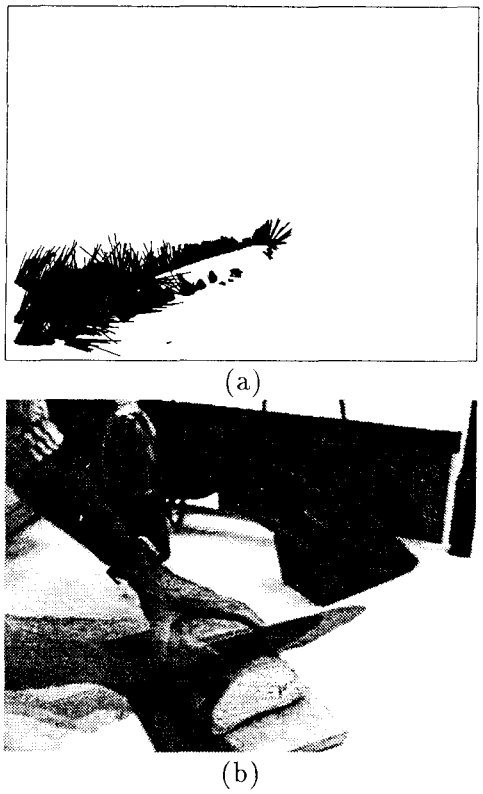


Figure 5: (a) Flow vectors for Slicing. (b) Slicing motion.

simple hand-tools based on their funtions. In *Proceedings of the Asian Conference on Computer Vision*, pages 656–659, 1993.

[Kitahashi et al., 1991] T. Kitahashi, N. Abe, S. Dan, K. Kanada and H. Ogawa. A function-based model of an object for image understanding. In *Advances in Information Modelling and Knowledge Bases*, IOS Press, H. Jaakkola and S. Ohusuga, editors, pages 91-97, 1991.

[Polana and Nelson, 1993] R. Polana and R. Nelson. Detecting activities. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 2–7, 1993.

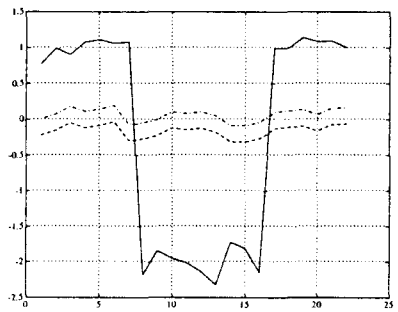


Figure 6: Angles α, β, and θ for Slicing. α is given by a dashed line, β is given by a solid line, and θ is given by a dash-dot line.

[Rivlin et al., 1994] E. Rivlin, S.J. Dickinson, and A. Rosenfeld. Recognition by functional parts. *Computer Vision and Image Understanding*, 62:164–176, 1995.

[Rivlin et al., 1993] E. Rivlin, A. Rosenfeld, and D. Perlis. Recognition of object functionality in goal-directed robotics. In *Proceedings of the AAAI Workshop on Reasoning about Function*, 1993.

[Solina and Bajcsy, 1983] F. Solina and R. Bajcsy. Shape and function. In *Proceedings of the SPIE Conference on Intelligent Robots and Computer Vision*, Volume 726, pages 284–291, 1983.

[Stark and Bowyer, 1991a] L. Stark and K. Bowyer. Achieving generalized object recognition through reasoning about association of function to structure. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 13:1097–1104, 1991.

[Stark and Bowyer, 1991b] L. Stark and K. Bowyer. Generic recognition through qualitative reasoning about 3-D shape and object function. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 251–256, 1991.

[Stark and Bowyer, 1992] L. Stark and K. Bowyer. Indexing function-based categories for generic recognition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 795–797, 1992.

[Stark et al., 1993] L. Stark, A. Hoover, D. Goldgof, and K. Bowyer. Function-based recognition from incomplete knowledge of shape. In *Proceedings of the IEEE Workshop on Qualitative Vision*, pages 11–22, 1993.

[Ullman and Basri, 1991] S. Ullman and R. Basri. Recognition by linear combinations of models. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 13:992–1006, 1991.

[Verri and Poggio, 1987] A. Verri and T. Poggio. Against quantitative optical flow. In *Proceedings of the International Conference on Computer Vision*, pages 171–180, 1987.

[Winston et al., 1983] P.H. Winston, T.O. Binford, B. Katz, and M. Lowry. Learning physical descriptions from functional descriptions, examples, and precedents. In *Proceedings of the National Conference on Artificial Intelligence*, pages 433–439, 1983.