Function From Motion

Zoran Duric, *Member, IEEE Computer Society*, Jeffrey A. Fayman, and Ehud Rivlin

Abstract—In order for a robot to operate autonomously in its environment, it must be able to perceive its environment and take actions based on these perceptions. Recognizing the functionalities of objects is an important component of this ability. In this paper, we look into a new area of functionality recognition: determining the function of an object from its motion. Given a sequence of images of a known object performing some function, we attempt to determine what that function is. We show that the motion of an object, when combined with information about the object and its normal uses, provides us with strong constraints on possible functions that the object might be performing.

Index Terms—Object recognition, action perception, functionality, motion, normal flow.

1 Introduction

In the field of robotics, researchers have long pursued the goal of enabling a robot to act autonomously in its environment. For robots, as for humans, recognizing the functions of objects is a prerequisite to autonomous interaction with them. Functionality can be defined as the usability of an object for a particular purpose [2]. As an example, suppose we would like to open a letter. We seek a sharp object such as a knife or a pair of scissors that would be appropriate for opening the letter. Clearly the knife or scissors are functional in the context of opening a letter, and a robot given the task of opening a letter would at some point be required to recognize such objects as being functional for its task.

Recent research has focused on the problem of recognizing object functionality (for a short survey see [2]). The goal of this research has been to determine functional capabilities of an object based on characteristics such as shape, physics and causation [20]. Little attention has been given to the problem of determining 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 functional recognition, 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 and the frame of reference it uses.

Information derived from motion can be useful in several ways. We expect a robot to take actions based on perceived events. In many instances, the events are perceived visually as motions in the environment. For example, a robot serving as a mechanic's mate [3] might "see" a person tightening a bolt with a pair of pliers and offer the person a wrench which would be more suitable for the task. Here, the robot determines the function of the pliers based on their motion (i.e., tightening) and determines that the wrench would be more suitable for tightening the bolt. This is an example of action perception. This ability can be used by an observing robot for monitoring other agents performing different tasks. Action perception can be useful in other domains as well. For example in automatic video sequence analysis, an observer can search a sequence for a specific action, which is a combination of a known object going through a certain motion profile.

Function based recognition tries to achieve a mapping from function to form. When an agent has some action to carry out an appropriate object is searched for. Observing an acting agent trying to perceive what is the action taking place involve an inversion of this mapping. Since the mapping from function to form is many to many, we need the information provided by motion to enable us to infer what is the mapping that the acting agent did, exact. In the above example, we would like to use a motion of a tool (i.e., tightening) to determine the function of the tool.

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

In Section 2 we review literature that describes related work. In Section 3 we cover some preliminaries related to the problem. 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 the area.

- *Z. Duric is with the Machine Learning and Inference Laboratory, George Mason University, and the Center for Automation Research, University of Maryland. E-mail duric@cfar.umd.edu*
- *J.A. Fayman is a graduate student in the Computer Science Department at the Israel Institute of Technology. E-mail jefff@isaac.cs.technion.ac.il.*
- *E. Rivlin is with the Computer Science Department at the Israel Institute of Technology. E-mail ehud@cfar.umd.edu*

Manuscript received Dec. 14, 1994; revised Mar. 25, 1996. Recommended for acceptance by A. Singh.
For information on obtaining reprints of this article, please send e-mail to: transpami@computer.org, and reference IEEECS Log Number P96029.

2 Related Work

Our research is concerned with the problem of determining the function of an object by analyzing its motion. Motion and functionality have appeared in the literature in several contexts. Early work on functional recognition can be found in [5], [17], [25]. More recently, Stark and Bowyer [18], [19], [20], [21] used these ideas to solve some of the problems presented by more traditional model-based methods of object recognition. In the so-called function-based approach, an object category is defined in terms of properties that an object must have in order to function as an instance of that category [20]. This work deals only with the stationary objects; no motion is involved. In recent work Green et al. [6] discuss the use of motion information for the recognition of articulated objects using function. The motion is used to determine whether the object in view possess the appropriate functional properties. The analysis is done using a full 3D boundary description. The motion is not used to infer function in action.

Gould and Shah [7] use motion characteristics obtained by tracking representative points on an object to identify important events corresponding to changes in direction, speed and acceleration in the object's motion. They believe that, "in many cases where an object has a fixed and predefined motion, the trajectories of several points on the object may serve to uniquely identify the object." This identification would be achieved by analyzing motion characteristics alone without requiring an object model; but no object identification results were given. We believe that since many objects display similar motion characteristics, motion alone is insufficient for function-based analysis, to determine the function of an object in action one needs not only its motion, but also the object's form.

Motion analysis for recognition of activities was described by Polana and Nelson [14]. They use Fourier analysis to detect and localize periodic activities such as walking or flying in a sequence of images. This work is similar in nature to our work in that both use motion as a basis for identifying activities. However, Polana and Nelson are concerned only with detecting the activities, without concern for the source of the motion. This is not adequate for function-based analysis since many objects can display similar motion characteristics. An object model is necessary to distinguish between the functions of objects that display similar motion characteristics.

Our work depends on segmenting the object into primitive parts and analyzing their motions. This kind of segmentation into functional parts was discussed by Rivlin et al. in [15]. They proposed a technique for functional recognition which extends the "Recognition by Parts" paradigm of object recognition to support "Recognition by Functional Parts."

3 Preliminaries

In this section we begin with a discussion of primitive shapes and motions. Next, we derive equations of motion for both the observer-centered and the object-centered coordinate systems. We then derive projected motion equations for the plane perspective imaging model and show how these equations can be simplified by the use of weak perspective projection [22]. Finally, we derive the relationship between the image velocities and the projected motion.

3.1 Primitive Shapes and Primitive Motions

Following [1], [15], [16] 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 [15] we let a_1, a_2, and a_3 represent length, width, and height, respectively, of a volumetric part, we can define the four classes as follows:

$$\textit{Stick}: \quad a_1 \simeq a_2 \ll a_3 \ \vee\ a_1 \simeq a_3 \ll a_2 \ \vee\ a_2 \simeq a_3 \ll a_1 \tag{1}$$

$$\textit{Strip}: \quad a_1 \neq a_2 \wedge a_2 \neq a_3 \wedge a_1 \neq a_3 \tag{2}$$

$$\textit{Plate}: \quad a_1 \simeq a_2 \gg a_3 \ \vee\ a_1 \simeq a_3 \gg a_2 \ \vee\ a_2 \simeq a_3 \gg a_1 \tag{3}$$

$$\textit{Blob}: \quad a_1 \simeq a_2 \simeq a_3 \tag{4}$$

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.

These primitives can be combined to create compound objects. In [15] the different qualitative ways in which these 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 most 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 action—for example, slicing or chopping. We define basic or primitive motions to be motions along, or perpendicular to, the main axes of a primitive object. (It is interesting to note that motions along the main axis of a primitive preserve "degenerate views" [10].) The motion can be a translation or a rotation.

3.2 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 X_1, Y_1, Z_1 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. It is assumed that these functions are differentiable with respect to 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. Let $\vec{i}, \vec{j}$, and $\vec{k}$ be the unit vectors in the directions of the Ox, Oy, and Oz axes, respectively; and let $\vec{i}_1$, $\vec{j}_1$, and $\vec{k}_1$ be the unit vectors in the directions of the Cx_1, Cy_1, and Cz_1 axes, respectively. For a given position $\vec{p}$ of P in $Cx_1y_1z_1$ we have the position $\vec{r}_p$ of P in $Oxyz$

$$\vec{r}_p \equiv \begin{pmatrix} X \\ Y \\ Z \end{pmatrix} = \begin{pmatrix} \vec{i}\cdot\vec{i}_1 & \vec{i}\cdot\vec{j}_1 & \vec{i}\cdot\vec{k}_1 \\ \vec{j}\cdot\vec{i}_1 & \vec{j}\cdot\vec{j}_1 & \vec{j}\cdot\vec{k}_1 \\ \vec{k}\cdot\vec{i}_1 & k\cdot\vec{j}_1 & \vec{k}\cdot\vec{k}_1 \end{pmatrix} \begin{pmatrix} X_1 \\ Y_1 \\ Z_1 \end{pmatrix} + \begin{pmatrix} X_c \\ Y_c \\ Z_c \end{pmatrix} \equiv R\vec{p} + \vec{d}_c \quad (5)$$

where R is the matrix of the direction cosines (the frames are taken as right-handed so that $\det R = 1$). 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

$$\begin{pmatrix} \dot{X} \\ \dot{Y} \\ \dot{Z} \end{pmatrix} = \begin{pmatrix} 0 & -C & B \\ C & 0 & -A \\ -B & A & 0 \end{pmatrix} \begin{pmatrix} X - X_c \\ Y - Y_c \\ Z - Z_c \end{pmatrix} + \begin{pmatrix} U \\ V \\ W \end{pmatrix} \quad (6)$$

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

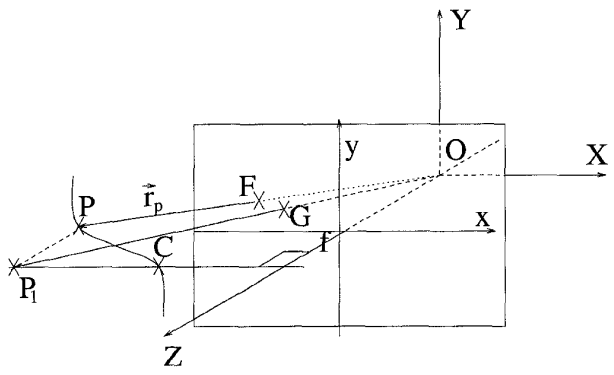


Fig. 1. The plane perspective projection image of P is $F = f(X/Z, Y/Z, 1)$; the weak perspective projection image of P is obtained through the plane perspective projection of the intermediate point $P_1 = (X, Y, Z_c)$ and is given by $G = f(X/Z_c, Y/Z_c, 1)$.

3.3 The Imaging Model

Let (X, Y, Z) denote the Cartesian coordinates of a scene point with respect to the fixed camera frame (see Fig. 1), 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 then given by

$$x = \frac{X}{Z_c} f,\ y = \frac{Y}{Z_c} f. \quad (7)$$

3.4 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 (7) with respect to time and using (6):

$$\dot{x} = f\frac{\dot{X}Z_c - X\dot{Z}_c}{Z_c^2} = f\frac{\left[-C(Y - Y_c) + B(Z - Z_c) + U\right]Z_c - XW}{Z_c^2} = \frac{Uf - xW}{Z_c} - C(y - y_c) + fB\left(\frac{Z}{Z_c} - 1\right), \quad (8)$$

$$\dot{y} = f\frac{\dot{Y}Z_c - Y\dot{Z}_c}{Z_c^2} = f\frac{\left[C(X - X_c) - A(Z - Z_c) + V\right]Z_c - YW}{Z_c^2} = \frac{Vf - yW}{Z_c} + C(x - x_c) - fA\left(\frac{Z}{Z_c} - 1\right) \quad (9)$$

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

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 (as we shall now see) 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

$$\frac{dI}{dt} = \frac{\partial I}{\partial x}\frac{dx}{dt} + \frac{\partial I}{\partial y}\frac{dy}{dt} + \frac{\partial I}{\partial t} = (I_x\vec{i} + I_y\vec{j}) \cdot (\dot{x}\vec{i} + \dot{y}\vec{j}) + I_t = \nabla I \cdot \dot{\vec{r}} + I_t$$

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 [9], 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

$$\vec{u}_n = (\vec{u} \cdot \vec{n}_r)\vec{n}_r = \frac{-I_t \nabla I}{\|\nabla I\|^2} \tag{10}$$

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

It was shown in [24] that the magnitude of the difference between $\vec{u}_n$ and the normal motion field $\dot{\vec{r}}_n$ is inversely proportional to the magnitude of the image gradient. Hence $\dot{\vec{r}}_n \approx \vec{u}_n$ when $\|\nabla I\|$ is large. Equation (10) thus provides an approximate relationship between the 3D motion and the image derivatives. We will use this approximation later in this paper.

4 Function From Motion

Function-based recognition tries to achieve a mapping from function to form. When an agent has some action to carry out, an appropriate object is searched for. The recognition process is an attempt to achieve compatibility with some action requirements [23]. As mentioned earlier some basic "actions" are static by nature (supporting, containing, etc.), but most actions involve using an object while it is moving. Observing an acting agent trying to perceive what is the action taking place involve an inversion of this mapping. Because the mapping from function to form is many to many, if one is interested in revealing the mapping additional constraints are needed. An important such a constraint is motion.

In this work, we are interested in the mapping f: $M \mapsto F$ from motion to function. Given a moving object as seen by an observer we would like to infer the function being performed by the acting agent. The process is described in Fig. 2. The object is given as a collection of primitives. In this example a knife is described as a collection of two primitives. In the figure the different combinations for a stick and a strip are shown. The knife can be composed from a stick and a strip, two sticks or two strips, and the exact combination is given to the system. Given the model the system estimate the pose of the object which is passed to the motion estimation module. The model and the results of the motion estimation phase enable the system to infer the function that is performed by the agent. In this paper we develop and test the motion estimation needed for the mapping.

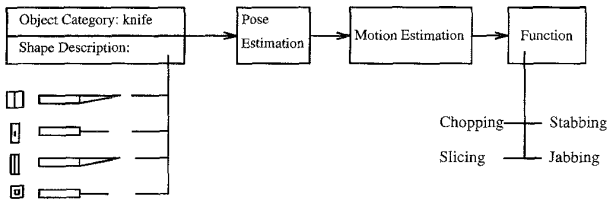


Fig. 2. Mapping motion to function: Given an object, i.e., the shape primitives that constitute the object (front view on the left, side view on the right), the image is processed to achieve a pose estimation. The image sequence is analyzed and motion estimation is carried out. The combination of the given model and the results of the motion estimation enables us to infer the mapping to the function that is carried out by the acting agent.

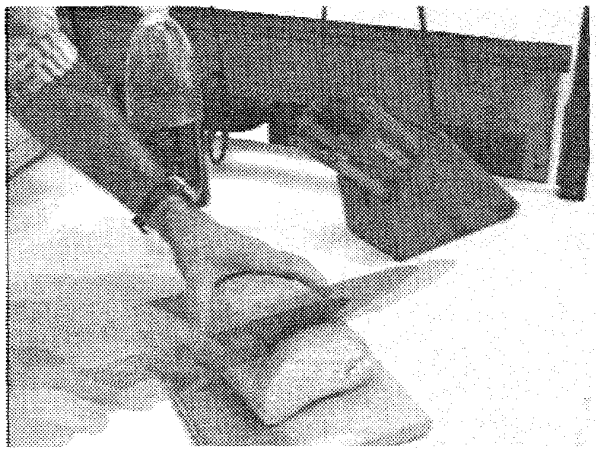


Fig. 3. Slicing: The motion of the strip is planar and periodic.

We are interested in the object's motion over time in the object's coordinate system and its relation to the object it acts on (the actee). Both of these measurements are necessary for the mapping. The object's motion over time in the object coordinate system gives us the relationship between the main axis of the object and its direction of motion. Given an object, these relationships help to determine the intended function. For example, we would expect the motion of a knife that a person is using to "stab" to be parallel to the main axis of the knife, whereas if the person is "chopping" with the knife we would expect motion perpendicular to the main axis.

In our work we give importance to primitive motions. Basic or primitive motions (which can be rotational or translational) are motions relative to the main axes of a primitive object. For a stick, translation and rotation along the main axis are important. When pure rotation is involved we have a screwdriver or a rotor (long and short axis respectively). When pure translation along the main axis is performed we have stabbing or jabbing. (When torsion is also involved we have screwing, drilling, etc.) For a plate, translation along the direction of the normal and rotation around the normal are important. In other directions there is no special component because a plate is isotropic. For a strip there are two important axes (it can be regarded as a plate and stick combined). Note that in all of these examples the motion is in a plane.

When determining function from motion, attention must be paid to the intended recipient. The relation to the actee is essential for establishing the mapping and creating a frame of reference. The importance of the actee in constraining the shape of the acting object was discussed in [11], [12] (where it is termed "functant"). They emphasize the ways in which the acting object shape is constrained by the need to match the shape of the actee. Here we are interested in the spatial relationship between the acting object and the actee—i.e., in establishing the frame of reference. Once this frame is established, motion of a knife in one direction could result in stabbing while motion in perpendicular direction results in slicing. Humans usually employ reference frames in which one axis represents the gravity vector, but this is not neces-

sary. We can slice bread on a wall as well as on a table; what matters is the motion of the knife relative to the actee.

In the next section, we develop the motion estimation machinery needed for this class of examples and we formalize our procedure for obtaining f: $M \mapsto F$. We assume we are given a model of the object in view, that is the type of the primitive part we are observing. We assume a recovery process like the one described in [15] to give us this kind of information. With each object category we relate a set of motion descriptors which map the given model to a function. These descriptors contain values for the different motion parameters as a function of time. Matching the observed motion to the stored descriptors result in function recognition. As was mentioned before the actee has an integral part in the recognition process, as it provides the context. The motion analysis relates to the external coordinate system as given by the observer and the actee. On the other hand the system does not verify a successful execution. For achieving that the actee needs to be observed and analyze. Such an analysis is given in [2] and we did not include it in the system. In its present form the analysis will not differentiate between a real act and a play (pantomime).

In Section 5 we analyze motion of sticks and strips. We assume a pose estimation module like in [4] which we use to establish the object frame. We assume that the motion of the tool is planar and that the plane in which the tool moves is "visible" by the observer (camera). The "visibility" constraint allows an oblique view as long as the angle between the surface normal and the z-axis of the camera is less than or equal to 30°. When the hand tool 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 hand tool 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.

Experimental results inferring function from image sequences are presented in Section 6.

5 Motion of Sticks and Strips

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 (see Fig. 4).

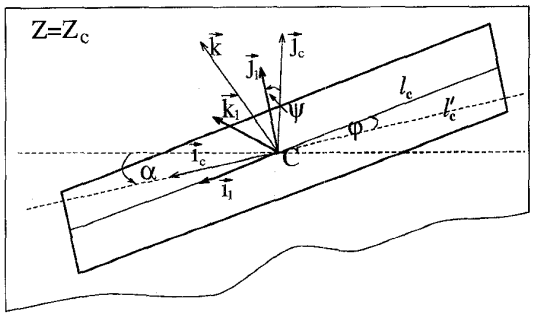


Fig. 4. The object centered coordinate system $Cx_1y_1z_1$: The unit vectors $\vec{i}_1$, $\vec{j}_1$, and $\vec{k}_1$ are parallel to the principal axes of the object.

In this paper we consider only planar and approximately 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; the axis of the maximal moment of inertia is orthogonal to the plane of the strip. In the case of a straight stick we have the center of mass C at the middle of its medial axis l_c: in this case 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 tool is planar and that the plane in which the tool moves is "visible" by the observer.[1] When the hand tool 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 hand tool 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. When the tool is a stick we choose the axis of minimal moment of inertia (it can be chosen arbitrarily) 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) (see Fig. 4). 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 the $\vec{j}$ 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 the $\vec{k}$ along l_c.

1. The "visibility" constraint allows an oblique view as long as the angle between the surface normal and the z-axis of the camera is less than or equal to 30°.

Let the angle between the plane Π_y and the Cy_1 axis of the object be ψ. The rotation $R_{x_i}(-\psi)$ around the Cx_1 axis of the object through the angle $-\psi$ 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} around an axis Cy_c (passing through C and parallel to $\vec{j}_c$) through the angle $-\varphi$ 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 $R_z(-\varphi)$ around the axis Cz (passing through C and parallel to $\vec{k}$) through the angle $-\alpha$ 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 (5) is then given by

$$R = \begin{pmatrix} \cos\alpha & -\sin\alpha & 0 \\ \sin\alpha & \cos\alpha & 0 \\ 0 & 0 & 1 \end{pmatrix} \times \begin{pmatrix} \cos\varphi & 0 & \sin\varphi \\ 0 & 1 & 0 \\ -\sin\varphi & 0 & \cos\varphi \end{pmatrix} \times \begin{pmatrix} 1 & 0 & 0 \\ 0 & \cos\psi & -\sin\psi \\ 0 & \sin\psi & \cos\psi \end{pmatrix} \tag{11}$$

5.1 Image Motion Fields of Sticks and Strips

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

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

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

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

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 and is orthogonal to Π_z; 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 $f(Z_cN_z)^{-1}$ and using (7) we obtain

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

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

From (8), (9), and (14) we obtain the equations of projected motion for points on $\mathcal{B}$ under weak perspective:

$$\dot{x} = \frac{Uf - xW}{Z_c} - C_1(y - y_c)N_z - C_1\left[(x - x_c)N_xN_y/N_z + (y - y_c)N_y^2/N_z\right], \tag{15}$$

$$\dot{y} = \frac{Vf - yW}{Z_c} + C_1(x - x_c)N_z + C_1\left[(x - x_c)N_x^2/N_z + (y - y_c)N_xN_y/N_z\right]. \tag{16}$$

Equations (15) and (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 $\vec{r} = x\vec{i} + y\vec{j}$ and the normal direction $n_x\vec{i} + n_y\vec{j}$ we have from (15) and (16) the normal motion field

$$\begin{aligned} \dot{\vec{r}}_n \cdot \vec{n} = n_x\dot{x} + n_y\dot{y} &= n_xf\left[U/Z_c + (x_c/f)C_1N_xN_y/N_z\right] - \\ & n_x\left(W/Z_c + C_1N_xN_y/N_z\right) \\ & -n_x(y - y_c)C_1\left(N_z + N_y^2/N_z\right) + n_yf\left[V/Z_c - (y_c/f)C_1N_xN_y/N_z\right] \\ & -n_yy\left(W/Z_c - C_1N_xN_y/N_z\right) + n_y(x - x_c)C_1\left(N_z + N_x^2/N_z\right). \end{aligned} \tag{17}$$

Let

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

$$\mathbf{c} = \begin{pmatrix} c_1 \\ c_2 \\ c_3 \\ c_4 \\ c_5 \\ c_6 \end{pmatrix} \equiv \begin{pmatrix} U/Z_c + (x_c/f)C_1N_xN_y/N_z \\ W/Z_c + C_1N_xN_y/N_z \\ C_1\left(N_z + N_y^2/N_z\right) \\ V/Z_c - (y_c/f)C_1N_xN_y/N_z \\ W/Z_c - C_1N_xN_y/N_z \\ C_1\left(N_z + N_x^2/N_z\right) \end{pmatrix} \tag{18}$$

Using (18) we can write (17) as

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

The column vector $\mathbf{a}$ is formed of the observable quantities only, while each element of the 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.

5.2 Estimating Motion Parameters From Normal Flow

If we use the spatial image gradient as the normal direction $\vec{n}_r \equiv \nabla I/\|\nabla I\| = n_x\vec{i} + n_y\vec{j}$ and $\vec{r}_n \approx \vec{u}_n$ 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 $A^TA\mathbf{c} = A^T\mathbf{b} \equiv \mathbf{d}$. We solve the system $A^TA\mathbf{c} = \mathbf{d}$ using the Cholesky decomposition. Since the matrix A^TA is a positive definite 6×6 matrix there exists a lower triangular matrix L such that $LL^T = A^TA$. We then have $LL^T\mathbf{c} = \mathbf{d}$. 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

$$\frac{U}{Z_c} = c_1 - \frac{x_c c_7}{f},$$

$$\frac{V}{Z_c} = c_4 + \frac{x_c c_7}{f},$$

$$\frac{W}{Z_c} = \frac{c_2 + c_5}{2},\quad C_1 = \operatorname{sgn}(c_6)\sqrt{c_3c_6 - c_7^2}$$

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

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

and by rearrangement we obtain

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

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

$$d_2\tan\psi = d_1\sin\varphi + d_3\cos\varphi \equiv \sqrt{d_1^2 + d_3^2}\,\sin(\varphi - \varphi_0) \quad (22)$$

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 30° bound). To each value of φ corresponds 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 captured the motion of an object performing a task. The vision system recorded images at 25 frames per second for five 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.[2] 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).

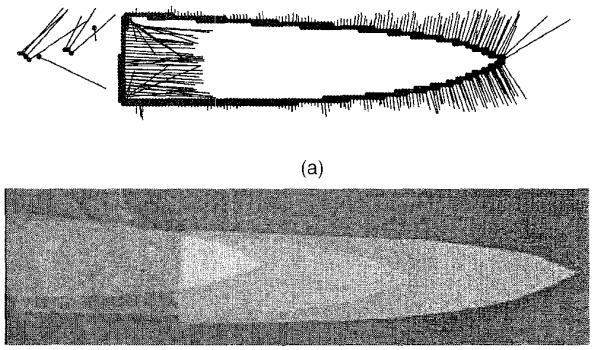
In Section 6.1 we describe the method which we use to estimate the direction of the medial axis α and the center of mass (x_c, y_c) of the image of the tool; we also define the parameters used to describe the motion of sticks and strips. In other subsections we describe the experiments on real image sequences. Two types of experiments were performed. First we show how the motion can be used to discriminate between different functionalities of the cutting tools. All the functionalities belong to the same family of manipulation tasks, namely cutting. In the second scenario we show how the motion information can be used to discriminate between two different functionalities of the same object, but this time for two different families of manipulation tasks. We give two examples of this scenario. In the first example, a shovel is used for scooping or for hitting. In the second example, a wrench is used for tightening or for hammering. In both examples, the first use is the normal one and the second use is an instance of improvisation. The motion gives clear information for a correct interpretation of the action that is taking place.

6.1 Parameterizing Motion of Sticks and Strips

We have assumed that an approximate direction (right, left, up, down) of the acting part of the tool 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 orientation 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 is 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.

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



(a) (b)

Fig. 5. (a) Flow vectors for jabbing. (b) Jabbing motion.

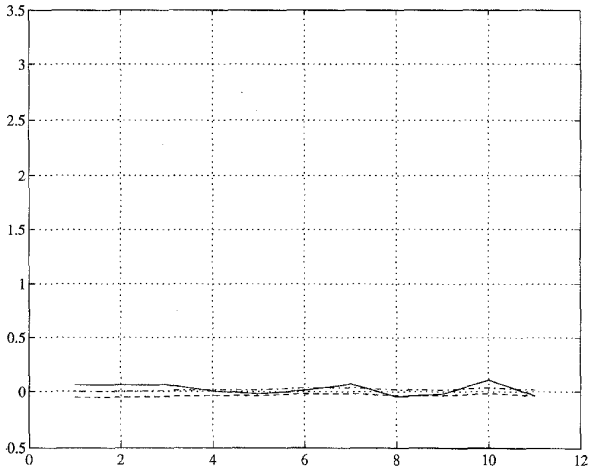


Fig. 6. Angles α, β, and θ for jabbing. α is given by a dashed line, β is given by a solid line, and θ is given by a dash-dot line.

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. We have

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

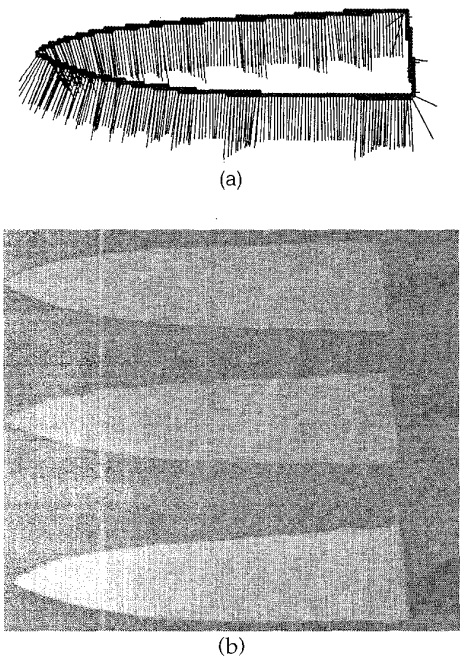
We define θ to be the total rotation angle as a function of time. We have

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

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

6.2 Action Recognition for a Class of Manipulation Tasks: Cutting

We start with three examples of simple functions performed by knives: chopping, jabbing and stabbing. In what follows we demonstrate how motion can be used to differentiate between the three. Finally, we show the cases of slicing with a knife (periodic motion) as well as sawing with a saw.



(a) (b)

Fig. 7. (a) Flow vectors for chopping. (b) Chopping motion.

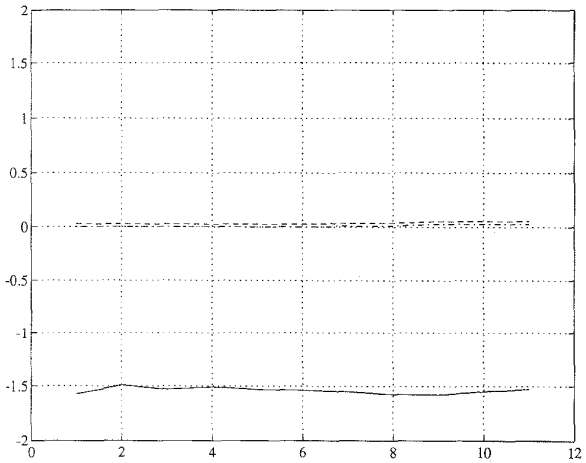


Fig. 8. Angles α, β, and θ for chopping. α is given by a solid line, and θ is given by a dash-dot line.

6.2.1 Jabbing

Jabbing 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 approximately 0, and θ is small and approximately constant.

Fig. 5 shows the flow vectors taken from the sixth sample and a composite image of the knife taken from the first, sixth and eleventh samples of the jabbing experiment. Fig. 6 shows a plot of the triple (α, β, θ) with respect to time (frame numbers). We can see that the values of α are very close to 0, as was expected, β is close to 0, and θ is around 0.

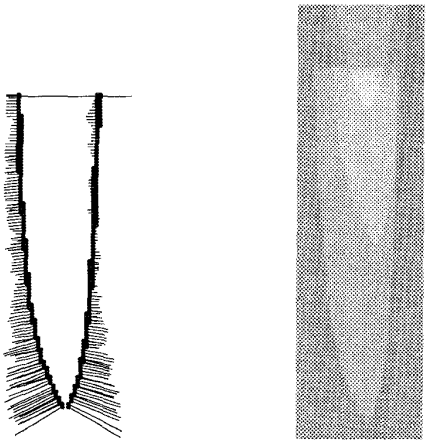


Fig. 9. (a) Flow vectors for stabbing. (b) Stabbing motion.

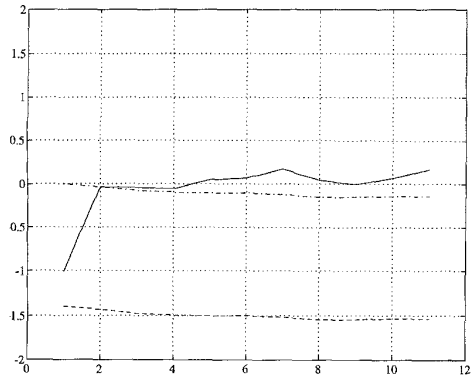


Fig. 10. Angles a, b, and q for jabbing. a is given by a dashed line, b is given by a solid line, and q is given by a dash-dot line.

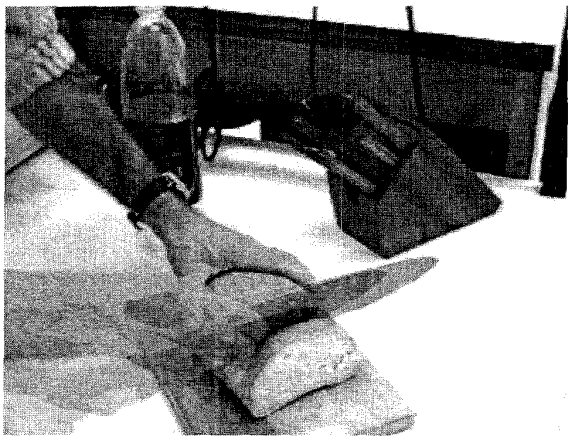
6.2.2 Chopping

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. Fig. 7 shows the flow vectors taken from the sixth sample and a composite image of the knife taken from the first, sixth and eleventh samples of the chopping experiment. Fig. 8 shows a plot of the triple (α, β, θ) with respect to time (frame numbers). We can see that the values of α are very close to 0, as was expected, β is close to $-\pi/2$, and θ is around 0.

6.2.3 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. The difference between jabbing and stabbing is in α. Fig. 9 shows the flow vectors taken from the sixth sample and a composite image of the knife taken from the first, sixth, and eleventh samples of the stabbing experiment. Fig. 10 shows a plot of the triple (α, β, θ) with respect to time (frame numbers). We can see that the values of α are very close to $-\pi/2$, as was expected, β is close to 0, and θ is around 0.

(a)



(b)

Fig. 11. (a) Flow vectors for slicing. (b). Slicing motion.

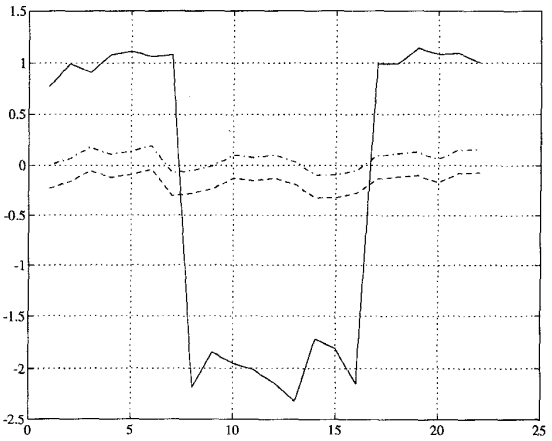
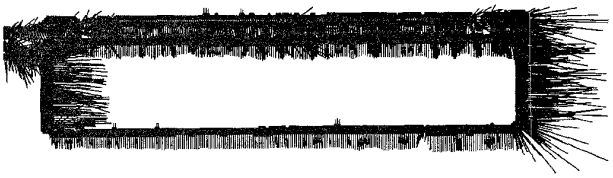


Fig. 12. Angles α, β, *and* θ for slicing. α is given by a dashed line, β is given by a solid line, and θ is given by a dash-dot line.

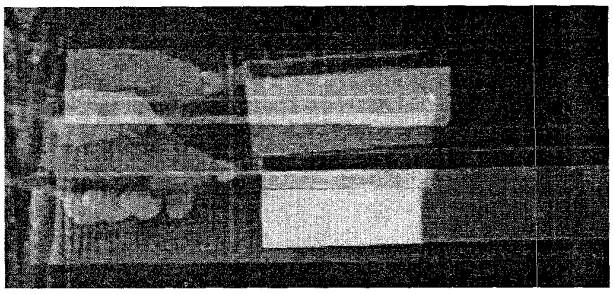
6.2.4 Periodic Motion: Slicing

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

Fig. 11 shows the flow vectors taken from the sixth sample and a composite image of the knife taken from the first, sixth, and eleventh samples of the slicing experiment. Fig. 12 shows a plot of the triple (α, β, θ) with respect to time (frame numbers). We can see that the values of α are very close to 0, as was expected, β oscillates between approximately $\pi/2$ and approximately $-3\pi/2$ (note the two approximate values differ by π).



(a)



(b)

Fig. 13. (a) Flow vectors for sawing. (b) Sawing motion.

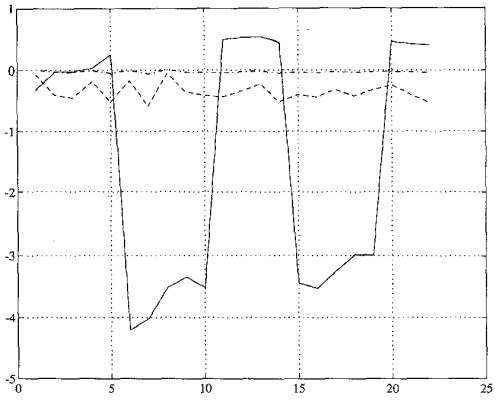
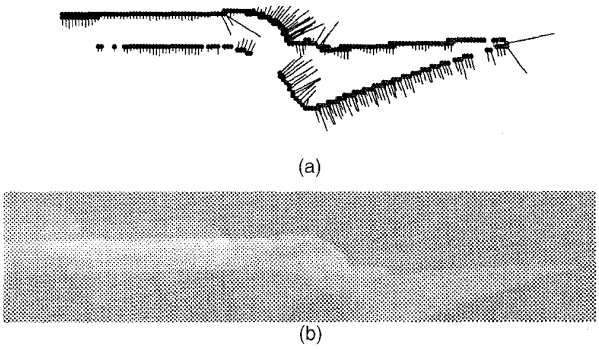


Fig. 14. Angles α, β, *and* θ for sawing. α is given by a dashed line, β is given by a solid line, and θ is given by a dash-dot line.

6.2.5 Periodic Motion: Sawing

Sawing is defined as the periodic cutting motion of a saw in which α is approximately 0, β is oscillating between approximately 0 and approximately π (or any two values which differ by approximately π), and θ is small and approximately constant.

Fig. 13 shows the flow vectors taken from the sixth sample and a composite image of the saw taken from the first, sixth and eleventh samples of the sawing experiment. Fig. 14 shows a plot of the triple (α, β, θ) with respect to time (frame numbers). We can see that the values of α are very close to 0, as was expected, β oscillates between approximately 0 and approximately $-\pi$.



(a)

(b)

Fig. 15. (a) Flow vectors for scooping with a shovel. (b) Scooping motion.

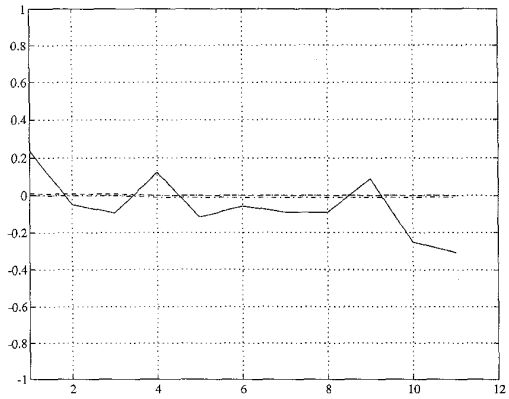


Fig. 16. Angles α, β, *and* θ for scooping with a shovel. α is given by a dashed line, β is given by a solid line, and θ is given by a dash-dot line.

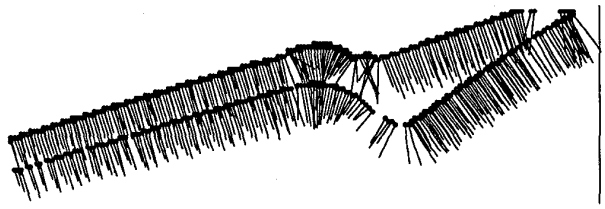
6.3 Multi-Usage Objects

In this Section we have two examples of multiple use of objects. We examine two actions using a shovel and two using a wrench.

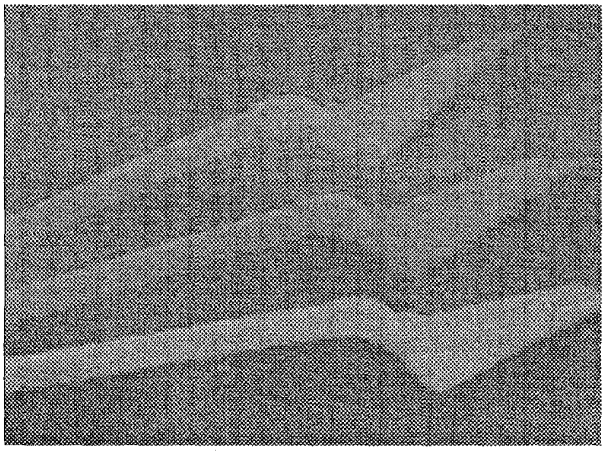
6.3.1 Shovel

Two actions using a shovel were examined. In one experiment, the shovel was used in a scooping action; in the other sequence, it was used in a hitting action. In these cases the same tool is being used for two inherently different functions. This example of double usage is a typical instance of improvisation. In [8] the relationship between the physical properties of an object, its functional representation, and its use in problem solving was explored. This analysis can be used to predict the types of motions one can expect for a given primitive shape. In the following experiment we use motion analysis to differentiate between two possible uses of an object.

Scooping with a shovel is a type of motion in which the rotational angle θ is small and the angle β (which corresponds to the translational direction in the object coordinate frame) is small. Fig. 15 shows the flow vectors taken from the sixth sample and a composite image of the shovel taken from the first, sixth, and eleventh samples of the scooping with a shovel experiment. Fig. 16 shows a plot of the triple (α, β, θ) with respect to time (frame numbers). We can see that the values of θ are small while α and β are close to 0.



(a)



(b)

Fig. 17. (a) Flow vectors for hitting with a shovel. (b) Hitting motion.

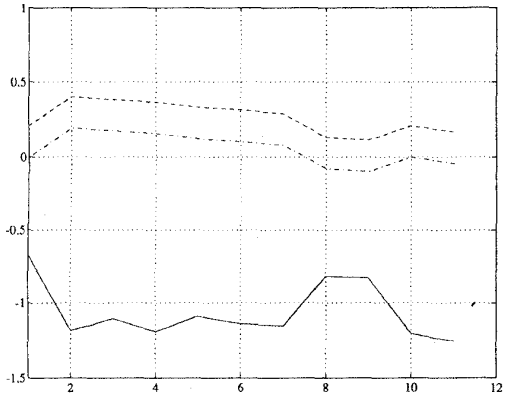
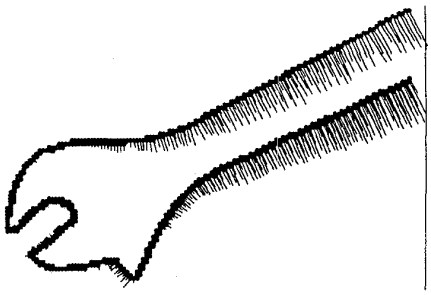
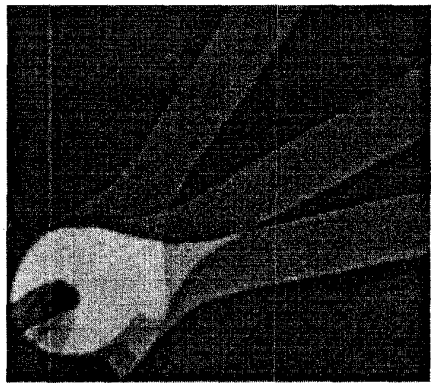


Fig. 18. Angles α, β, *and* θ for hitting with a shovel. α is given by a dashed line, β is given by a solid line, and θ is given by a dash-dot line.

Hitting with the shovel is a type of motion in which the translational part of the motion dominates over the rotational part of the motion and the direction of translation is approximately orthogonal to the direction of the medial axis of the tool. Fig. 17 shows the flow vectors taken from the sixth sample and a composite image of the shovel taken from the first, sixth, and eleventh samples of the hitting with a shovel experiment. Fig. 18 shows a plot of the triple (α, β, θ) with respect to time (frame numbers). We can see that the values of α are small and that $\theta \approx 0$ while β is close to $-\pi/2$.



(a)



(b)

Fig. 19. (a) Flow vectors for tightening with a wrench. (b) Tightening motion.

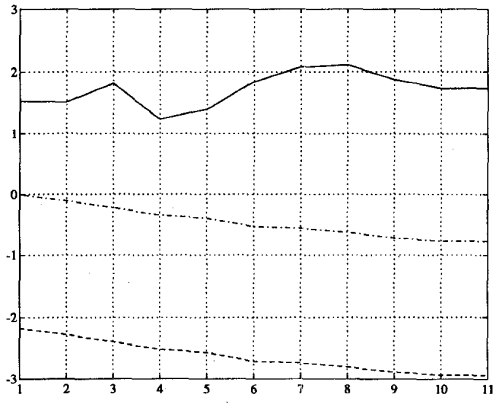
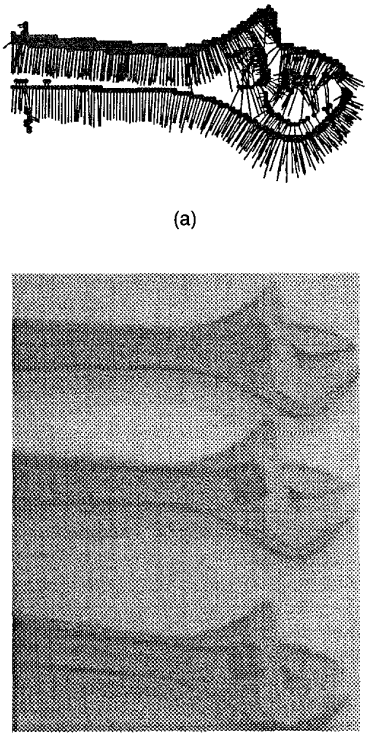


Fig. 20. Angles α, β, *and* θ for tightening with a wrench. α is given by a dashed line, β is given by a solid line, and θ is given by a dash-dot line.

6.3.2 Wrench

Two actions using a wrench were examined. In one experiment, the wrench was used to tighten a bolt; in the other sequence, the wrench was used as a hammer. In these cases the same tool is being used for multiple, inherently different functions. Motion analysis enables us to differentiate between the two.



(a) (b)

Fig. 21. Flow vectors for hammering with a wrench. (b) Hammering motion.

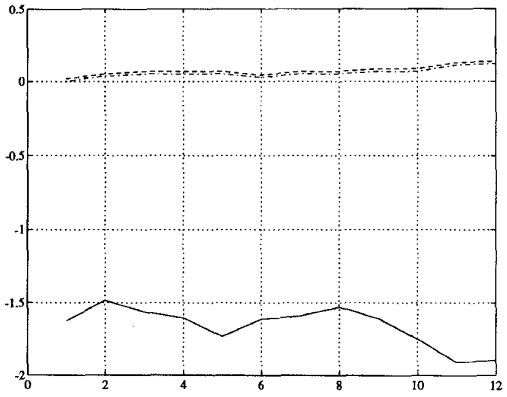


Fig. 22. Angles α, β, *and* θ for hammering with a wrench. α is given by a dashed line, β is given by a solid line, and θ is given by a dash-dot line.

Tightening with the wrench is type of motion in which the rotational angle θ and the angle β (which corresponds to the translational direction in the object coordinate frame) have opposite signs. This is equivalent to saying that the head of the tool is fixed while the handle is moving.

Fig. 19 shows the flow vectors taken from the sixth sample and a composite image of the wrench taken from the first, sixth and eleventh samples of the tightening with a wrench experiment. Fig. 20 shows a plot of the triple (α, β, θ) with respect to time (frame numbers). We can see that the values of α are decreasing (this is equivalent to $\theta < 0$) while β is close to $\pi/2$.

Hammering with the wrench is a type of motion in which the translational part of the motion dominates over the rotational part of the motion and the direction of translation is approximately orthogonal to the direction of the medial axis of the tool. Fig. 21 shows the flow vectors taken from the sixth sample and a composite image of the wrench taken from the first, sixth, and eleventh samples of the hammering with a wrench experiment. Fig. 22 shows a plot of the triple (α, β, θ) with respect to time (frame numbers). We can see that the values of α are small and that $\theta \approx 0$ while β is close to $-\pi/2$.

7 Conclusions

Perceiving function from motion provides an understanding of the way an object is being used by an agent. To accomplish this we combined information on 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.

Seven sequences of images were used to demonstrate the approach. Function understanding from motion was established in all seven cases. In the first three sequences, motion was used to discriminate between three cutting actions: stabbing, chopping and jabbing. In the last two pairs of sequences 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. These examples of double usage are typical instances of improvisation; motion provides clear information for a correct interpretation of the action that is taking place.

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 relative motion of the parts). Learning is another possible extension. A robot can learn object functionality by watching the object in use. As an example, a robot might "see" a knife being used to open a letter and learn the function of cutting and the context in which it can be used. We see our work as a step toward action perception of moving objects, which could lead to a better understanding of perceiving the actions of moving agents.

REFERENCES

[1] I. Biederman, "Human Image Understanding: Recent Research and a Theory," *Comp. Vision, Graphics and Image Processing*, vol. 32, pp. 29-73, 1985.

[2] L. Bogoni and R. Bajcsy, "Active Investigation of Functionality," *Proc. CVPR Workshop on Visual Behaviors*, Seattle, Wash., June 1994.

[3] M. Brady, P.E. Agre, D.J. Braunegg, and J. Connell II, "The Mechanic's Mate," *Proc. Sixth European Conf. on Artificial Intelligence*, pp. 79–94, 1984.

[4] D. Dementhon and L. Davis, "Model-Based Object Pose in 25 Lines of Code," *Int'l J. Comp. Vision*, vol. 15, pp. 123–141, 1995.

[5] P. Freeman and A. Newell, "A Model for Functional Reasoning in Design," *Proc. Int'l Joint Conf. Artificial Intelligence*, pp. 621–640, Aug. 1971.

[6] K. Green, D. Eggert, L. Stark, and K. Bowyer, "Generic Recognition of Articulated Objects by Reasoning About Functionality," *Proc. AAAI-94 Workshop on Representing and Reasoning About Device Function*, pp. 56-64, 1994.

[7] K. Gould and M. Shah, "The Trajectory Primal Sketch: A Multi-Scale Scheme for Representing Motion Characteristics," *Proc. IEEE Conf. Comp. Vision and Pattern Recognition*, pp. 79–85, June 1989.

[8] J. Hodges, "Naive Mechanics—A Computational Model of Device Use and Function in Design Improvisation," *IEEE Expert*, vol. 7, pp. 14–27, 1992.

[9] B.K.P. Horn and B.G. Schunck, "Determining Optical Flow," *Artificial Intelligence*, vol. 17, pp. 189–203, 1981.

[10] J.R. Kender and D.G. Freudenstein, "What Is a Degenerate View?" *Proc. DARPA Image Understanding Workshop*, pp. 589–598, 1987.

[11] K. Kise, H. Hattori, T. Kitahashi, and K. Fukunaga, "Representing and Recognizing Simple Hand-Tools Based on Their Functions," *Proc. Asian Conf. Comp. Vision*, pp. 656–659, 1993.

[12] T. Kitahashi, N. Abe, S. Dan, K. Kanada, and H. Ogawa, "A Function-Based Model of an Object for Image Understanding," H. Jaakkola and S. Ohusuga, eds, *Advances in Information Modeling and Knowledge Bases*, pp. 91–97, 1991.

[13] H. Murase and S.K. Nayar, "Learning Object Models From Appearance," *Proc. National Conf. Artificial Intelligence*, pp. 836–843, Washington, D.C., July 1993.

[14] R. Polana and R. Nelson, "Detecting Activities," *Proc. IEEE Conf. Comp. Vision and Pattern Recognition*, pp. 2–7, New York, June 1993.

[15] E. Rivlin, S.J. Dickinson, and A. Rosenfeld, "Recognition by Functional Parts," *Proc. IEEE Conf. Comp. Vision and Pattern Recognition*, pp. 267–274, Seattle, Wash., June 1994.

[16] E. Rivlin, A. Rosenfeld, and D. Perlis, "Recognition of Object Functionality in Goal-Directed Robotics," *Proc. AAAI Workshop Reasoning About Function*, 1993.

[17] F. Solina and R. Bajcsy, "Shape and Function," *Proc. SPIE Conf. Intelligent Robots and Comp. Vision*, vol. 726, pp. 284–291, 1983.

[18] L. Stark and K. Bowyer, "Achieving Generalized Object Recognition Through Reasoning About Association of Function to Structure," *IEEE Tran. Pattern Analysis and Machine Intelligence*, vol. 13, pp. 1097–1,104, 1,991.

[19] L. Stark and K. Bowyer, "Generic Recognition Through Qualitative Reasoning About 3D Shape and Object Function," *Proc. IEEE Conf. Comp. Vision and Pattern Recognition*, pp. 251–256, Maui, Hawaii, 1991.

[20] L. Stark and K. Bowyer, "Indexing Function-Based Categories for Generic Recognition," *Proc. IEEE Conf. Comp. Vision and Pattern Recognition*, pp. 795–797, Champaign, Ill., June 1992.

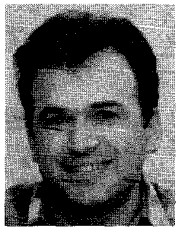
[21] L. Stark, A. Hoover, D. Goldgof, and K. Bowyer, "Function Based Recognition From Incomplete Knowledge of Shape," *Proc. IEEE Workshop Qualitative Vision*, pp. 11–22, New York, 1993.

[22] S. Ullman and R. Basri, "Recognition by Linear Combinations of Models," *IEEE Trans. Pattern Analysis and Machine Intelligence*, vol. 13, pp. 992–1006, 1991.

[23] L. Vaina and M. Jaulent, "Object Structure and Action Requirements: A Compatibility Model for Functional Recognition," *Int'l J. Intelligent Systems*, vol. 6, pp. 313-336, 1991.

[24] A. Verri and T. Poggio, "Against Quantitative Optical Flow," *Proc. Int'l Conf. Comp. Vision*, pp. 171–180, London, England, June 1987.

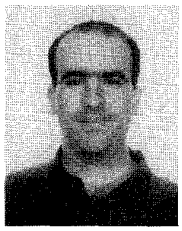
[25] P.H. Winston, T.O. Binford, B. Katz, and M. Lowry, "Learning Physical Descriptions From Functional Descriptions, Examples, and Precedents," *Proc. National Conf. Artificial Intelligence*, pp. 433–439, 1983.



Zoran Duric is assistant research scientist at the Machine Learning and Inference Laboratory at George Mason University and at the Center for Automation Research at the University of Maryland. He received his MS in electrical engineering from the University of Sarajevo, Bosnia and Herzegovina, Yugoslavia, in 1986 and his PhD in computer science from the University of Maryland at College Park in 1995.

From 1982 to 1989 he was a member of the research staff at the Division for Vision and Robotics, Energoinvest Institute for Control and Computer Science, Sarajevo. He was also affiliated with the Electrical Engineering Department of the University of Sarajevo. From 1990 to 1995, he was a graduate research assistant at the Center for Automation Research at the University of Maryland at College Park. In the summer of 1993, he was a summer research assistant at the NEC Research Institute, Princeton, N.J.

His interests include computer vision, video image processing, and applications of machine learning to computer vision. He is a member of the IEEE Computer Society.



Jeffrey A. Fayman received the BS degree in business administration (information systems) from San Diego State University, San Diego, Calif. He received the MS degree in computer science from the same university. He is currently pursuing the PhD degree in computer science at the Israel Institute of Technology, Technion.

His research interests include active vision, functionality, fault tolerance in computer robotic systems, and real-time systems.

Ehud Rivlin received the BSc and MSc degrees in computer science and the MBA degree from Hebrew University in Jerusalem and the PhD degree from the University of Maryland.

Currently he is an assistant professor in the Computer Science Department at the Technion, Israel Institute of Technology. His current research interests include machine vision and robot navigation.