Time to Contact from Active Tracking of Motion Boundaries

Shanon X. Ju
Department of Computer Science
University of Toronto
Toronto, Ont M5S 1A4 Canada

Michael J. Black[1]
Xerox Palo Alto Research Center
3333 Coyote Hill Road
Palo Alto, CA 94304

ABSTRACT

This paper addresses the problem of recovering time-to-contact by actively tracking motion boundaries. Unlike previous approaches which use image features, we use the camera's own motion to both detect and track object boundaries. First we develop a framework in which the boundaries of objects are automatically detected using the motion parallax caused by the motion of an active camera. We use a correlation-based method to locate motion boundaries and our work has focused on detecting the motion boundaries *early* and *robustly*. A confidence field, which expresses the likelihood that a point lies on a motion boundary, is constructed from the shape of the correlation surface. Spatial coherence of object boundaries is modeled with dynamic contours which are automatically initialized using an attentional mechanism. Then, as the camera moves, the shapes of the dynamic contours are held fixed and they are tracked under an assumption of affine deformation. The affine parameters are recovered from the active tracking over time and are used to compute time-to-contact. We illustrate the behaviour of this active approach with experiments on both synthetic and real image sequences.

Keywords: Motion boundaries, time-to-contact, robust estimation, snakes, affine motion, rigid contour tracking, active boundary detection.

1 INTRODUCTION

In a complex and dynamic environment, a robot must be able to detect the boundaries of objects for common tasks like obstacle avoidance and grasping. An active robot can use its own motion by exploiting the effect of *motion parallax* to aid in the recovery of these object boundaries. Previous researchers have shown how an active observer can use its own motion to compute differential image invariants and use these to estimate surface slant and time-to-contact (TTC).[6] These approaches have focused on the tracking of features in the brightness image which may not correspond to the physical boundaries of objects in the scene.

In this paper we show how these results can be extended so that the camera's own motion can be used to estimate the location of object boundaries. These object boundaries are then tracked as the camera moves using an assumption of affine deformation of the object contour. We illustrate the behavior of this active approach to boundary detection and rigid affine model tracking with experiments which show how the recovered affine parameters are used to compute time-to-contact. Time-to-contact is defined as the current estimate of the time until an observer either collides with, or passes, an object in its path if it continues with the same relative translational velocity. This information is important to a moving robot that must avoid obstacles, pass through doors, or grasp objects. The primary measure of TTC is derived from the first order differential invariants of the image velocity field.[6,12,18]

We propose a new framework for estimating TTC by active tracking of motion boundaries. The framework

[1] This work was supported by a grant from the Natural Sciences and Engineering Research Council of Canada.

has four levels. First, motion boundaries are detected in a sequence of images using a correlation-based approach. Most previous approaches to motion discontinuity detection have assumed that the optical flow field has already been recovered.[16] Another class of techniques recovers discontinuities and flow simultaneously using line processes[13] or weak constraints.[10,14] These approaches are computationally expensive and currently are not appropriate for discontinuity detection in an active environment. Our approach is novel in that we recover motion boundaries *early*; that is, before the computation of optical flow.[4] Second, the discontinuities correspond to surface boundaries in the world, and hence in practice, we can assume that such boundaries have spatial coherence. Spatial coherence is enforced using controlled continuity splines (ie. Snakes).[11] Unlike previous approaches which have relied on manual initialization of the snakes[6] we exploit an automatic initialization scheme based on an attentional mechanism.[7] Third, the motion boundaries of the rigid objects are tracked over time. Assuming an affine transformation of the feature shapes, we present a novel method to recover the affine parameters by rigid contour tracking. While standard snakes are used to model the spatial coherence of object contours, only affine deformations of the snake are allowed between frames. Finally, time-to-contact of the object is estimated from the affine parameters for deliberately forward motion.

Cipolla and Blake[6] present a method to track a closed image contour using B-spline snakes and to estimate TTC of the object from the temporal changes in the area of the closed snake. Their approach avoids estimating a dense optical flow field and its partial derivatives. Although the tracking of intensity features is simple and accurate, there is no guarantee that these features correspond to physical properties of the scene. If information about time-to-contact is to be used for obstacle avoidance or grasping, we would like to track physical boundaries in the scene. Black and Anandan[4] exploit five constraints to achieve early detection of motion discontinuities. Their work can be divided into two parts. First, they construct a confidence field corresponding to the measurement of the surface discontinuities by taking into account three properties of the correlation surface. Next, a snake is initialized manually near the discontinuity, and the snake will automatically be attracted to the local maxima, which correspond to the motion boundaries. The first and second levels of our framework are similar to their approach.

Our approach is novel in that the goal of our tracking process is not only to find the new position of the contours, but also to recover the approximate linear transformation of the shape between each frame. This transformation is related to the divergence and deformation of the image velocity field, as well as the 3D structure of the scene and the motion of the viewer. The remainder of this paper is organized as follows: From Section 2 to Section 5, each of the four levels of the framework is developed in detail and illustrated with experimental results on synthetic data. We then present the results of a real image sequence. Finally, we will discuss limitations of the current approach, future work, and draw a brief conclusion.

2 EARLY DETECTION OF MOTION BOUNDARIES

2.1 Correlation Based Approach – Review

Correlation-based matching exploits the assumption of *data conservation*; that is, the local brightness distribution remains unchanged although its location may change. The basic idea can be expressed as the minimization of the following error measure:

$$\mathcal{E}(u,v) = \sum_{(x,y)\in\mathcal{R}} \rho(I(x,y,t) - I(x+u\delta t, y+v\delta t, t+\delta t)) \tag{1}$$

where [u, v] is the displacement, $\mathcal{R}$ stands for the correlation window, ρ is an error norm, I is the brightness function at time t, and δt is a small time step. When $\rho(x) = x^2$, Equation 1 is the standard Sum-of-Squared-Differences (SSD) measure.[2] The *correlation surface* is defined over a search window with the height of the surface corresponding to the error measure, $\mathcal{E}(u,v)$, of a particular displacement.

The correlation approach assumes the flow field within the correlation window can be approximated as a uniform translational motion. When multiple motions exist within a correlation region, the data conservation constraint is violated, and the correlation surface may contain multiple minima corresponding to the different motions. Black and Anandan[4] pointed out that the presence of multiple minima in the correlation surface indicates the possible presence of a motion discontinuity. Therefore, with some additional constraints, it is often possible to detect the motion discontinuities before the computation of optical flow.

2.2 Robust Estimation

Stated simply, the goal of robust statistics is to estimate the parameters of a model that best fit a set of measurements, when a minority of the data may be outliers.[9] Consider the minimization problem:

$$\min_{\mathbf{a}} \sum_{s \in \mathcal{S}} \rho(d_s - \mathbf{u}(s, \mathbf{a})),$$

where $\mathbf{u}(s, \mathbf{a})$ is the model, $\mathbf{a}$ is the parameter vector, and $d_s, s \in \mathcal{S}$ is a set of observations of the data. When the errors in the measurements are normally distributed, the optimal *maximum-likelihood* estimate is obtained when ρ is quadratic, i.e., *least-squares* estimation. However, the model assumptions may be violated; for example, in computing optical flow, the uniform motion model within a region is violated at motion boundaries. To cope with these problems, robust estimators have been used instead of least-squares estimation.[5]

An estimator is said to be robust if it is insensitive to outliers. The problem with the least-squares approach (Figure 1a) is that an arbitrarily bad outlier can produce an arbitrarily bad estimate regardless of the sample size. Hampel *et al.*[9] introduced the approach based on *influence functions*. Loosely speaking, the influence function **IF** is proportional to the first derivative of ρ-function and measures the asymptotic bias caused by contamination in the observations. For least-squares estimation, the influence of outliers increases linearly and without bound (Figure 1b). To increase robustness, the robust *redescending* estimators in maximum likelihood estimation are introduced.[9] These estimators have the property that, beyond a threshold, the influence of outliers decreases. Figure 1c and 1e show two examples of robust ρ-functions.

Correlation assumes a single motion within the correlation region. When computing the correlation at a motion boundary, the measurements from one surface can corrupt those of the other surface. To reduce the effect of outlying measurements we replace the quadratic with a robust ρ-function. Figure 2b and 2c compare the correlation surface generated using the least-squares estimator and truncated quadratic estimator. The surfaces are computed at the corner of the metal bracket (Figure 2a) from the NASA Coke can sequence. The two peaks in Figure 2c correspond to the two motions present in the window. It is clear that the robust estimator makes the peaks more visible.

2.3 Confidence Measures

The shape of the correlation surface is typically quite complex. It not only contains information about the motion of the surface, but is also related to the brightness patterns of the region.[2] In the case of repetitive structures, the surface is ridge-like or multi-ridge-like. In a homogeneous area, the surface shows very little variation. We need a pointwise confidence field to show the presence of a motion boundary. This field will be computed before the computation of the dense optical flow field. We choose three measures:

$\mathcal{C}_{peak}$: the height of the best peak: The lower the measure, the less the match error. At a motion boundary, the best match error of the point should be a local maximum in a neighborhood.

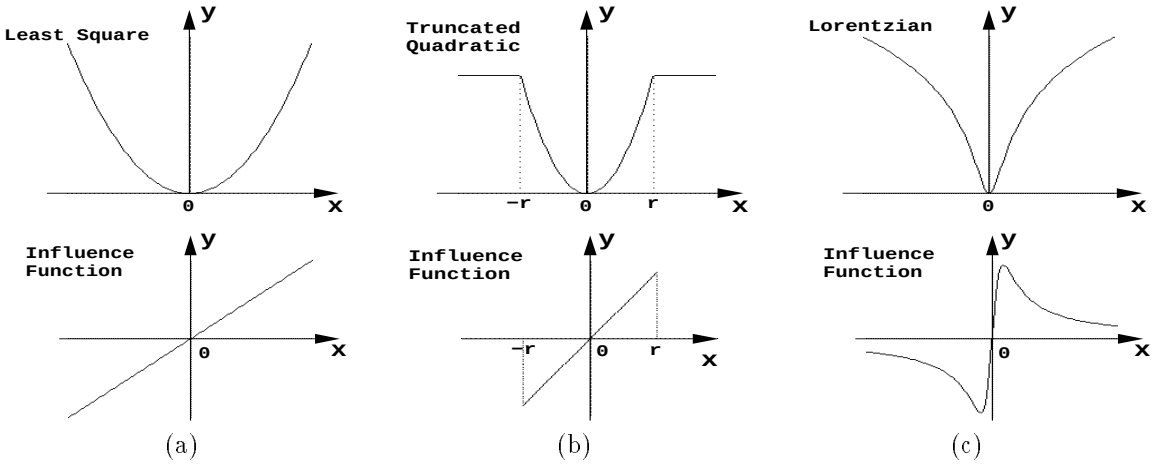


Figure 1: (a) Least-squares error norm and its influence function. (b) Truncated quadratic ρ-function and its influence function. (c) Lorentzian ρ-function: $log(1+\frac{x^2}{2\sigma^2})$, and its influence function: $\frac{2x}{2\sigma^2+x^2}$, where $\sigma^2=\frac{1}{8}$.

$\mathcal{C}_{steep}$**: the steepness of the best peak:** Assuming the best match is (u,v), the steepness of the peak is computed as:

$$\sum_{i=-1}^{1}\sum_{j=-1}^{1}\omega_{i,j}\ \mathcal{E}(u+i,v+j) \qquad \begin{array}{|c|c|c|}\hline \omega_{-1,-1} & \omega_{0,-1} & \omega_{1,-1} \\ \hline \omega_{-1,0} & \omega_{0,0} & \omega_{1,0} \\ \hline \omega_{-1,1} & \omega_{0,1} & \omega_{1,1} \\ \hline \end{array} = \begin{array}{|c|c|c|}\hline 1 & 4 & 1 \\ \hline 4 & -20 & 4 \\ \hline 1 & 4 & 1 \\ \hline \end{array}$$

where $\mathcal{E}(u,v)$ is defined in Equation 1 and $\omega_{i,j}$ is the weight. The higher the measure, the more likely a steep peak exists. Therefore, at the motion boundary, this measure should be the lowest.

$\mathcal{C}_S$**: the ratio of the height of the two best peaks:** When there are two different motions within the correlation window, the surface will contain two apparent peaks. Intuitively, if the second peak is nearly as good as the first one, the likelihood of a discontinuity is high. A simple measure is defined as the ratio of the height of the two best peaks: ρ_0/ρ_1. Where ρ_0, ρ_1 are the match errors for the first and second peaks respectively. In certain well defined cases, the measure has a maximum of 1.0 at a motion boundary and falls off as distance from the boundary increases. If the relative motion of the surfaces is small, then due to the discretization, the peaks may merge together and this measure will be unreliable.

We can combine these three measures to form the confidence field Ψ where the maximum points correspond to the area where there is high confidence that a discontinuity is present:

$$\Psi = \lambda_1\mathcal{C}_{peak} - \lambda_2\mathcal{C}_{steep} + \lambda_3\mathcal{C}_S\,,$$

where λ_i are scalar weights. Figure 2d shows one image from a synthetic sign sequence, in which a slanted traffic sign is translating towards the viewer with respect to a stationary background. The correlation surface is computed when the search window is 11x11 and the correlation window is 15x15. Figure 2e, 2f, 2g, and 2h show the $\mathcal{C}_{peak}$, $\mathcal{C}_{steep}$, $\mathcal{C}_S$, and Ψ confidence measure fields respectively. $\mathcal{C}_{peak}$ and $\mathcal{C}_{steep}$ are scaled so that the lowest value is 0.0, while highest is 1.0, and $\mathcal{C}_S$ is scaled to 0.0-2.55[2]. The corresponding weights used to compute Ψ are: $\lambda_1 = 0.06$, $\lambda_2 = 0.01$, and $\lambda_3 = 3.0$.

[2]Those ranges are the same for the real image sequences in section 6.2.

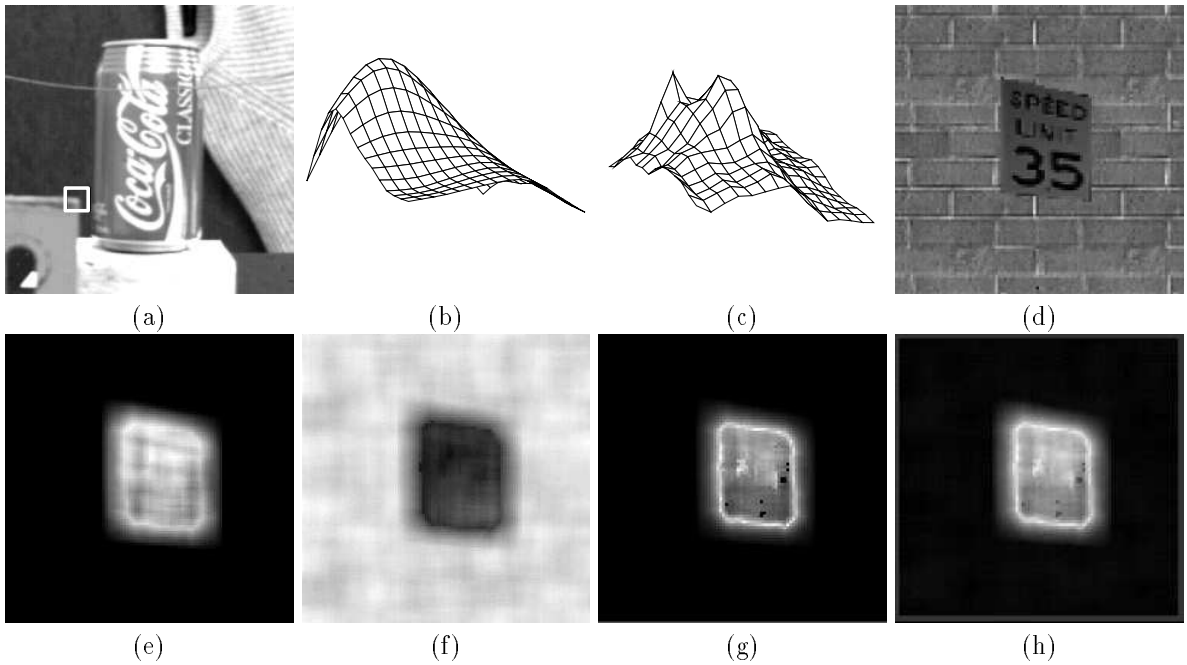


Figure 2: (a) A correlation window (white square) on a motion boundary. (b) SSD correlation surface of the region shown in (a). (c) Robust correlation surface of the same region. (Figures (b) and (c) are inverted for display). (d) one image from the traffic sign sequence. (e) $\mathcal{C}_{peak}$ confidence measure field. (f) $\mathcal{C}_{steep}$ confidence measure field. (g) $\mathcal{C}_S$ confidence measure field. (h) Ψ combined confidence field.

3 SPATIAL COHERENCE

3.1 Snake Model – Review

Motion discontinuities correspond to the boundaries of objects and hence have spatial extent. We model the spatial coherence of motion boundaries with *controlled continuity splines*, or snakes.[11] Snakes simulate the fitting of an elastic contour, providing a continuous boundary, to an image feature. Once the snake is interactively initialized near an object contour in the first frame, it will automatically track the contour from frame to frame as long as the feature does not move too fast. The behavior of a snake is controlled by internal and external forces. The internal forces serve as a smoothness constraint, and the external forces guide the active contour towards image features. Following the notation from the original model proposed by Kass *et al.*,[11] given a parametric representation of an image curve $v(s) = (x(s), y(s))$, the energy function is defined as

$$\mathcal{E}_{snake} = \int_0^1 \mathcal{E}_{int}(v(s)) + \mathcal{E}_{ext}(v(s))ds. \tag{2}$$

The function $\mathcal{E}_{int}$ represents the internal energy of the active contour and is composed of a first and second order terms:

$$\mathcal{E}_{int} = (\alpha|v_s(s)|^2 + \beta|v_{ss}(s)|^2)/2, \tag{3}$$

where the subscripts indicate differentiation with respect to s. Adjusting the weights α and β controls the relative importance of the first and second terms. $\mathcal{E}_{ext}$ represents the external potential $P(x, y) = c[G_\sigma * \Psi(x, y)]$, where c is the weight, $G_\sigma * \Psi$ denotes the image convolved with a Gaussian smoothing filter, Ψ is the confidence field on motion boundaries. $P(x, y)$ is a scalar potential function defined over image plane, which attract snake to intensity maxima. Minimizing the energy function of Equation (2) gives rise to two independent Euler equations.[11] The

tracking behavior of the snake is achieved by numerical, iterative solution of these two equations using techniques from variational calculus.

3.2 Initialization

It is well known that one problem with current snake models is that the recovered solution is sensitive to the initial snake position. In previous work, snakes are either interactively initialized[1,4,11] or automatically initialized using some prior knowledge about the position of the object.[6] We present an approach which uses an attentional mechanism to automatically initialize the snake. The role of attention mechanisms in computer vision is to select only the information essential to the current task and ignore the irrelevant details. Thus, attention, or task guidance, simplifies computation and reduces the amount of processing.

The attention procedure is based on the Culhane and Tsotsos attentional prototype,[7] which is composed of a processing hierarchy and an attention beam that guides selection of portions of the hierarchy. We construct a spatially filtered and subsampled pyramid in which the lowest level of the processing hierarchy is a region in the input image, and each successive level is a simple average of the previous level. A winner-take-all process (WTA), which specifies the "brightest" pixel as the winner, is performed at the top of the hierarchy. The *pass zone* is defined as the region that includes the winner and those elements at the lowest level that contribute to the winner. At each successively lower level, the WTA is only executed within the pass zone from the previous level. Once a winner has been located in the input level, it is considered as the final attentional point, and that pixel and a small surrounding zone, the *inhibited zone*, in the input image are inhibited. The approach is illustrated in Figure 3.

Snake initialization algorithm: (i) Perform the attention procedure in the motion discontinuity field to locate the first snake node. (ii) Run the attention procedure within a *search window* $\mathcal{R}$ around the current snake node, find a new snake node. (iii) In the input image, inhibit all the points on the path from the previous node to the present one. (iv) Repeat step (ii) and (iii) until the stop criteria have been satisfied. Figure 4a shows the result of the automatic initialization of a closed snake. The size of search window $\mathcal{R}$ is 17x17 and the inhibited zone is 3x3. The stop criteria are: (a) when the number of snake nodes found exceeds a threshold *and* (b) when the distance between the new node and the start node is less than a threshold. Figure 4b shows the local extrema found by the snake from the initial position.

4 ACTIVE TRACKING

Unconstrained snakes suffer from the drawback that they may be attracted by spurious contours if the environment is too complex or the motion between two frames is not small enough. This is mainly due to the excessive flexibility of the snake model. Ueda et al.[17] propose a moderate "stiffness" that preserves the shape of the tracked object contour in the previous frame as much as possible. Berger[3] adds a term to preserve at best the initial curvature during the snake process. Curwen et al.[8] combine the original B-spline snake model with a parametric template model.[19] The dynamic contour is first trained by an uncoupled snake, then "frozen" and becomes the template model. Ueda and Berger do not make any assumption about the motion of the objects. Curwen assumes constant velocity of features, and hence the approach has problems tracking divergent or rotational objects.

4.1 Rigid Affine Snake Model

The rigid affine snake model is built under three assumptions. (i) The target object is rigid. (ii) The distance between the object surface patch and the camera is large with respect to the surface extension, therefore the motion of the object between two frames can be approximated by an affine transformation. (iii) The change of

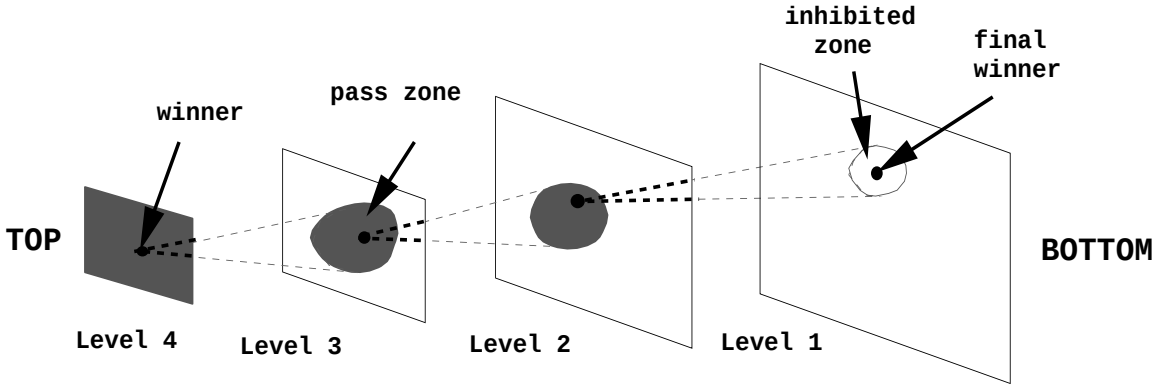


Figure 3: Processing hierarchy of locating the attentional point

the shape is not too great. An affine transformation can be described as:

$$\begin{aligned} u(x,y) &= a_1 + a_2 x + a_3 y, \\ v(x,y) &= a_4 + a_5 x + a_6 y, \end{aligned}$$

where u and v are the horizontal and vertical image velocity respectively and the a_i are the affine parameters. Using vector notation this can be rewritten as

$$\dot{\mathbf{X}} = \begin{bmatrix} u(x,y) \\ v(x,y) \end{bmatrix} = \mathbf{XA} \qquad and \qquad \mathbf{X} = \begin{bmatrix} 1 & x & y & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 & x & y \end{bmatrix}, \tag{4}$$

where $\mathbf{A}$ denotes the vector $(a_1, a_2, a_3, a_4, a_5, a_6)^T$, $\dot{\mathbf{X}}$ stands for the 2D velocity. We parameterize a tracked contour $\mathbf{v}_0(s) = (x_0(s), y_0(s))^T$ by $s \in [0,1]$ in the previous frame. Our goal is to deform the contour by an affine transformation from $\mathbf{v}_0(s)$ to the best position $\mathbf{v}(s) = (x(s), y(s))^T$ in the current frame such that the following energy function (substituting for $\mathcal{E}_{int}$ and $\mathcal{E}_{ext}$ in Equation 2) is minimized

$$\mathcal{E}(\mathbf{v}) = \int_0^1 (\alpha |v_s(s)|^2 + \beta |v_{ss}(s)|^2)/2 + P(\mathbf{v}(s)) ds, \tag{5}$$

where $P(x,y)$, defined in section 3.1, is the smoothed confidence field.

4.2 Lagrangian Dynamic System

Terzopoulos[15] constructs a dynamical snake system and allows it to arrive at a minimal energy states as it achieves equilibrium. This dynamic model can be derived by applying the principles of Lagrangian mechanics. A Lagrangian dynamic system is specified by a kinetic energy, a potential energy and a dissipative term. We represent a dynamic contour by introducing a time-varying mapping $\mathbf{v}(s,t)$. The contour is assumed to have constant mass density μ with respect to the curve parameter s. The kinetic energy of the contour is defined as $\mathcal{E}_k(\mathbf{v}) = \frac{\mu}{2} \int_0^1 |\mathbf{v}_t|^2 ds$. The subscript t denotes a time derivative. We combine the kinetic energy and the snake energy function $\mathcal{E}(\mathbf{v})$ (Equation 2) to define the Lagrangian

$$\mathcal{L}(\mathbf{v}) = \frac{1}{2} \int_0^1 \mu |\mathbf{v}_t|^2 ds - \mathcal{E}(\mathbf{v}). \tag{6}$$

If the initial and final positions of the snake are $\mathbf{v}(s,t_0)$ and $\mathbf{v}(s,t_1)$, then the deformable model's motion $\mathbf{v}(s,t)$ from $t = t_0$ to $t = t_1$ is such that the variation of the integral $\int_{t_0}^{t_1} \mathcal{L}(\mathbf{v}) dt$ with respect to $\mathbf{v}$ is zero:

$$\frac{\delta}{\delta \mathbf{v}} (\frac{1}{2} \int_{t_0}^{t_1} \int_0^1 \mu |\mathbf{v}_t|^2 - \alpha |\mathbf{v}_s|^2 - \beta |\mathbf{v}_{ss}|^2 - 2 \cdot P(\mathbf{v}) ds dt) = 0. \tag{7}$$

Once set in motion, a dynamic snake with a mass distribution will move perpetually, unless kinetic energy is dissipated. Given the damping density γ, the Rayleigh dissipation function, $\mathcal{E}_d(\mathbf{v}_t) = -\frac{\gamma}{2}\int_0^1 |\mathbf{v}_t|^2 ds$, is defined in order to dampen the snake so that static equilibrium can be achieved: Evaluating the appropriate variational derivatives of Equation 7 and $\mathcal{E}_d$ in the dissipation functional, the equation of motion for the snake may be written as

$$\mu\mathbf{v}_{tt} + \gamma\mathbf{v}_t - \alpha\mathbf{v}_{ss} + \beta\mathbf{v}_{ssss} = -\nabla P(\mathbf{v}(s,t)), \tag{8}$$

with appropriate initial and boundary conditions, and where the subscripts indicate partial derivatives with respect to t and s, and $\nabla P(\mathbf{v}(s,t))$ is the gradient of the potential. On the left hand side are inertia, damping, stretching, and bending forces. These forces balance the negative gradient of the potential on the right hand side.

4.3 Discretization and Estimation of Affine Parameters

According to Terzopoulos,[15] the discretized version of the Lagrangian dynamics (Equation 8) may in tern be written as a second order differential equation by applying finite difference methods:

$$M\ddot{\mathbf{X}} + C\dot{\mathbf{X}} + K\mathbf{X} = -\nabla P(\mathbf{X}) = f, \tag{9}$$

where the vector $\mathbf{X}$ is the collection of the nodal variables, M is the mass matrix, and C is the damping matrix. Both M and C are diagonal matrices with the diagonal element corresponding to the mass or damping density at a node respectively. K is the stiffness matrix which is a symmetric pentadiagonal matrix constructed from the weights α and β. $\dot{\mathbf{X}}$ denotes the nodal velocities. Recall that for a rigid affine contour, the nodal velocities can be expressed as the linear transformation of the nodal variables at the previous time instant (Equation 4). If we assume that the system is massless (i.e., $M = 0$), substituting $\dot{\mathbf{X}}$ with Equation 4, Equation 9 reduces to the following overdetermined linear equation with only six unknown parameters:

$$C\mathbf{X}\mathbf{A} + K\mathbf{X} = f. \tag{10}$$

Let δA_t denotes an incremental estimate of these parameters at a time step. The final step is to solve the above equation iteratively:

$$\delta A_t = (\mathbf{X}_{t-1}^T\mathbf{X}_{t-1} + \omega\mathbf{I})^{-1}(\omega A_{t-1} + f(\mathbf{X}_{t-1}) - K\mathbf{X}_{t-1}) \tag{11}$$

$$\mathbf{X}_t = \mathbf{X}_{t-1} + \mathbf{X}_{t-1}\delta A_t \tag{12}$$

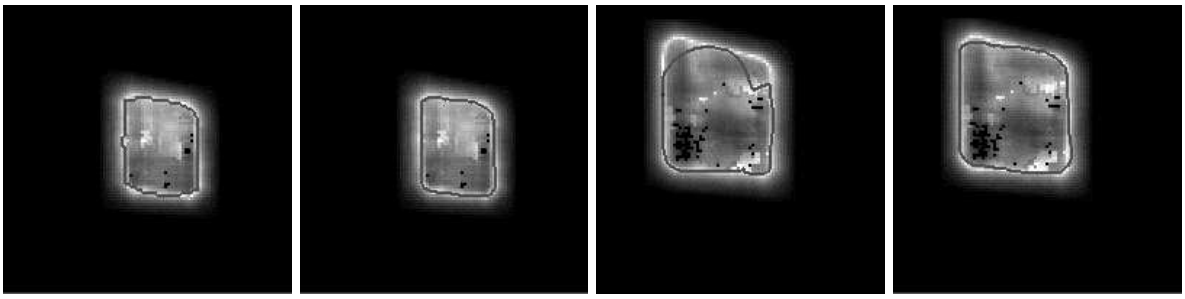
$$A_t = A_{t-1} + \delta A_t^T \begin{bmatrix} A_{t-1}^* & \mathbf{O} \\ \mathbf{O} & A_{t-1}^* \end{bmatrix} \qquad A^* = \begin{bmatrix} 1 & 1 & 1 \\ a_1 & a_2 & a_3 \\ a_4 & a_5 & a_6 \end{bmatrix}, \tag{13}$$

where ω is a step size. Figure 4c and 4d compare the standard snake model with our model when tracking the discontinuity sequence. Since the snake model is influenced by the false confidence value, it is not able to track the target contour. On the other hand, the proposed rigid affine tracking model successfully tracked it without being influenced.

5 ESTIMATION OF TIME-TO-CONTACT

An affine transformation can be decomposed into several independent components which have simple geometric interpretations.[6] Some of the first order differential invariants of the optical flow field are $curl\mathbf{v}$, $div\mathbf{v}$, $def\mathbf{v}$, and μ, which denote 2D rigid rotation, divergence, the magnitude of deformation, and the orientation of the deformation respectively. They can be defined from the partial derivatives of the image velocity

$$\begin{aligned} div\mathbf{v} &= u_x + v_y &&= a_2 + a_6, \\ curl\mathbf{v} &= -u_y + v_x &&= -a_3 + a_5, \\ (def\mathbf{v})\cos 2\mu &= u_x - v_y &&= a_2 - a_6, \\ (def\mathbf{v})\sin 2\mu &= u_y + v_x &&= a_3 + a_5, \end{aligned}$$



(a) (b) (c) (d)

Figure 4: (a) automatic initial position of a closed snake. (b) The snake attracted to the local minimum at the first frame. (c) Tracking using the standard snake model (the last frame in the sequence). (d) Rigid affine contour tracking (last frame).

where $\mathbf{v}(u,v)$ is image velocity at an image point (x,y), and a_i are affine parameters. The optical flow field produced by a rigid body motion is[18]

$$u = (-fV_x + xV_z)/Z + A\frac{xy}{f} - B(\frac{x^2}{f} + f) + Cy, \tag{14}$$

$$v = (-fV_y + yV_z)/Z + A(\frac{y^2}{f} + f) - B\frac{xy}{f} - Cy, \tag{15}$$

where $[u,v]^T$ is the flow vector at $[x,y]^T$, f is the camera focal length, the 3D rigid motion is a rotation $\mathbf{R} = [A,B,C]^T$ plus a translation $\mathbf{V} = [V_x, V_y, V_z]^T$, and $Z = Z(x,y)$ is the scene depth. The time to contact is defined by

$$T_c = \frac{Z}{V_z}. \tag{16}$$

In the case of pure translation along the viewing axis, the divergence of the optical flow field is

$$div\mathbf{v} = u_x + v_y = V_z(\frac{\partial Z(x,y)}{\partial x} + \frac{\partial Z(x,y)}{\partial y} + \frac{2}{Z}). \tag{17}$$

Under a weak perspective assumption, if the distance between the camera and a surface is large with respect to the depth extent of the surface, we can assume that the depth of the object is a constant, i.e., $\frac{\partial Z(x,y)}{\partial x} = \frac{\partial Z(x,y)}{\partial y} = 0$. Then, time-to-contact depends only on the divergence of the image velocity (Equation 17 and 16:)

$$T_c = \frac{2}{div\mathbf{v}} = \frac{2}{a_2 + a_6}. \tag{18}$$

6 EXPERIMENTAL RESULTS

6.1 Synthetic sequence

There are 9 images in the synthetic traffic sign sequence. The camera is undergoing a constant translational motion along the viewing axis. Initially the camera is at 3.1m from the sign, and the initial depth of the sign is 2.85m. The velocity of approach is 0.1m/time unit. The sign is slanted (56 degrees). If the shape of the contour is fixed to be rigid at the first frame and it is tracked actively through the last frame, its final position may not correspond well to the object boundary due to small non-affine shape changes which accumulate over time. We introduce an adaptive component to the rigid contour tracking to reduce the errors. Before the shape of the contour at current position is fixed for the tracking to next frame, the rigid contour is allowed to "relax" a little,

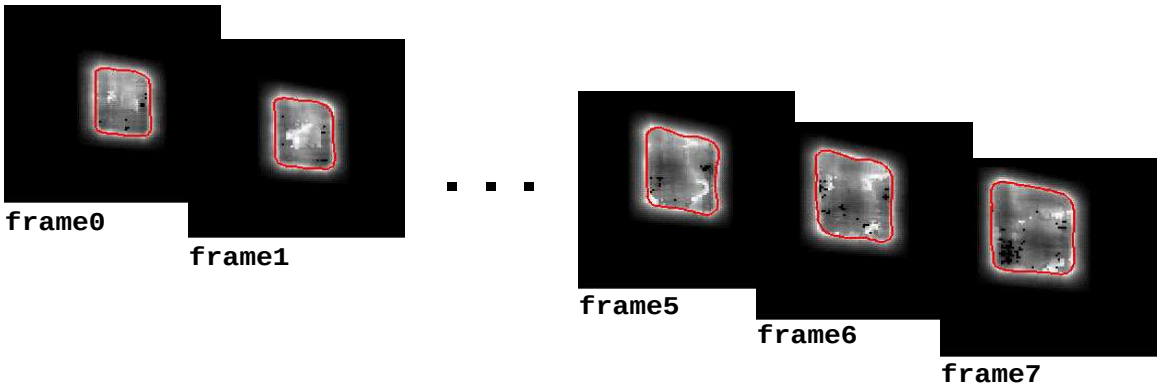


Figure 5: Rigid tracking of the traffic sign boundary.

i.e., a few iterations of the unconstrained snake will allow it to adapt to small shape changes. Figure 5 shows the adaptive rigid contour tracking sequence.

The estimated TTC of the traffic sign is computed with Equation 18. Due to discretization, the smaller the shape changes, the more unreliable the recovered value. On the other hand, when the change of the shape is not small enough, the tracking contour may be attracted by a spurious local maximum. To overcome this tradeoff, we *accumulate* the affine parameters between several frames. Given $i = 0, \cdots, n$, A_i denotes the affine deformation between image i and $i+1$. The accumulated affine parameter $\mathbf{A}$ denotes the affine deformation between image 0 and $n+1$, and the *accumulation time* is n. Figure 7a compares the true TTC and the estimated TTC of each frame from accumulated affine parameters. Initially, the small motion during the first few frames means that the affine parameters are not recovered with high accuracy. The accumulation of results over time improves the accuracy and after five frames the TTC estimate has converged to the correct value. In this experiment the accumulation window is four frames. We wish to note that Kalman filtering may provide a mechanism that integrates new estimate with existing TTC estimates to reduce the uncertainty over time.

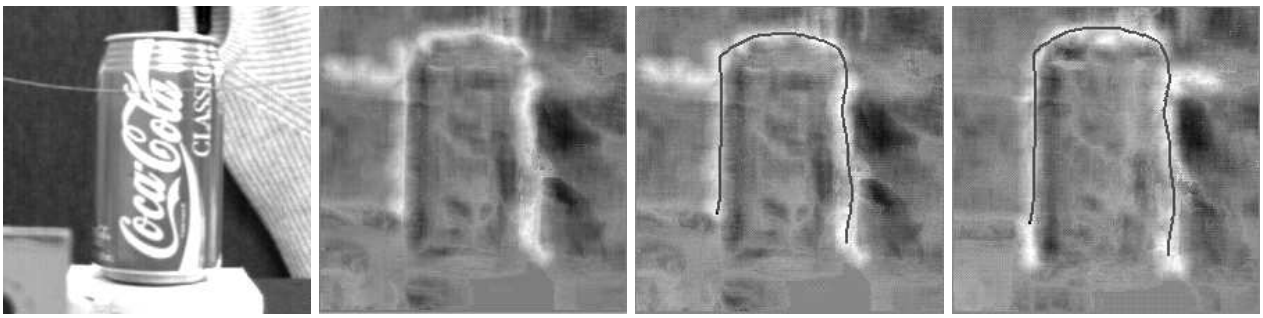
6.2 Natural Image Sequence

The original Coke sequence was collected at NASA Ames Research Center and contains 151 frames. The camera is moving along the viewing axis with the focus of expansion centered the can. We extract a sub-region from the imagery such that its size is 224x224 (Figure 6a), and we use only the last 50 images in the original sequence. Since the difference between the motion of the can and the background is much smaller than a pixel between frames, the discrete correlation method will fail to detect the motion boundary. In order to increase the difference of the motion of the can and the background, we use pairs of images, which are separated by 14 frames, to compute the correlation surface. Both the correlation window and the search window are 15x15. The current approach only detects first-order motion discontinuities and hence the crease where the base of the can comes in contact with the table is not detected.

Since the difference between the motion of the can and the background is about one pixel only, some spurious high confidence points will appear. Therefore, in addition to the combination of the three confidence measures described in Section 3.3, we add an additional test which is similar to the neighborhood test used by Black and Anandan.[4] Given the robust correlation surface, we use the best peak to compute a raw (unsmoothed) flow field. Then, in a small neighborhood around each pixel, we look for two anomalous situations which indicate the presence of a motion boundary: (i) A change in the horizontal or vertical flow of more than 1 pixel. (ii) A change in the horizontal or vertical flow that is inconsistent with forward motion. When either of these is detected, a constant factor is added to the confidence field at that point. Figure 6b shows the discontinuity confidence field computed from a pair of images. The weights used to compute Ψ are: $\lambda_1 = 0.05$, $\lambda_2 = 0.01$, and $\lambda_3 = 0.65$.

There are 11 images in the boundary sequence.

When dealing with real images, due to the presence of open boundaries, multiple objects, and significant noise, the initialization process is not trivial. An open snake was initialized manually with an starting position roughly near the boundary of the Coke can. Figure 6c shows the local maximum found by the snake as dark against the brighter confidence field (frame number is zero). Figure 6d shows the position of the affine contour of last image in the sequence. Figure 7b shows the estimated time-to-contact (TTC) from the accumulated affine parameters with a maximum accumulation time of seven frames. The figure indicates that the TTC is roughly linear, and camera velocity is a constant (no ground truth was available).



(a) (b) (c) (d)

Figure 6: (a) One image from the NASA Coke can sequence. (b) Discontinuity confidence field (see text). (c) The snake attracted to the local minimum. (d) Rigid affine contour tracking.

7 CONCLUSION

There are four main contributions of our work. First, a new framework for automatically detecting and tracking motion boundaries over time was developed. The affine parameters are recovered by the rigid contour tracking and are used to estimate time-to-contact. Second, robust estimation was applied to correlation based approaches. Third, an algorithm to automatically initialize a closed snake based on an attentional mechanism is developed. Finally, we introduce the notion of rigid contour tracking, and show how affine parameters can be recovered from the tracking of the motion boundaries under the assumption of a linear transformation of the object contour.

At present, the robust correlation method cannot distinguish motion boundaries from multiple motions resulting from fragmented occlusion, translucency or reflection. Also, the ratio of two peaks test requires the difference of the multiple motions to be larger than a single pixel. Moreover, the accumulation of deformation between frames can be improved by the introduction of a Kalman filter. Finally, the value of this work will be demonstrated when it is applied to the problems in both motion analysis and active vision. The notions of robust correlation, automatic initialization, and rigid contour tracking introduced in this paper have wider relevance than simply the estimation of time-to-contact.

8 REFERENCES

1. A. A. Amini, S. Tehrani, and T. E. Weymouth. "Using dynamic programming for minimizing the energy of active contours in the presence of hard constraints". *Proc. ICCV*, pages 95–99, Dec. 1988.
2. P. Anandan. "A computational framework and an algorithm for the measurement of visual motion". *International Journal of Computer Vision*, 2:283–310, 1989.

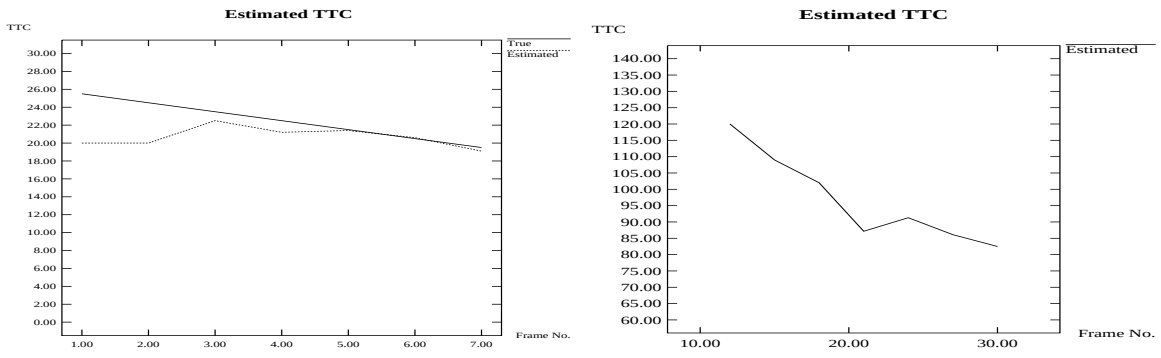


Figure 7: (a) Sign Sequence: estimated TTC (in frames, dotted line) at each frame from the accumulated affine parameters, and true TTC (solid line), x-axis stands for the frame number. (b) Can Sequence: estimated TTC (in frames) at each image from the accumulated affine parameters, x-axis stands for the frame number.

3. M. O. Berger. "Tracking rigid and no polyhedral objects in an image sequence". *Proceedings of The 8th Scandinavian Conference on Image Analysis*, 2:945–952, 1993.
4. M. J. Black and P. Anandan. "Constraints for the early detection of discontinuity from motion". *Proceedings of the National Conference on Artificial Intelligence*, pages 1060–1066, 1990.
5. M. J. Black and P. Anandan. "A framework for the robust estimation of optical flow". *Proc. ICCV*, pages 231–236, May 1993.
6. R. Cipolla and A. Blake. "Surface orientation and time to contact from image divergence and deformation". *Proc.ECCV*, pages 187–202, May 1992.
7. S. M. Culhane and J. K. Tsotsos. "An attentional prototype for early vision". *Proc. ECCV*, pages 551–560, May 1992.
8. R. Curwen and A. Blake. "Dynamic contour: Real-time active spline". In *Active Vision*, pages 39–57. The MIT Press, 1992.
9. F. R. Hampel, E. M. Ronchetti, P. J. Rousseeuw, and W. A. Stahel. *Robust Statistics: the approach based on influence functions.* probability and mathematical statistics. John Wiley & Sons, 1986.
10. J. G. Harris, C. Koch, E. Staats, and J. Luo. "Analog hardware for detecting discontinuities in early vision". *Int. Journal of Comp. Vision*, 4(3):211–223, June 1990.
11. M. Kass, A. Witkin, and D. Terzopoulos. "Snakes: Active contour models". In *Proc. ICCV*, pages 259–268, June 1987.
12. F. Meyer and P. Bouthemy. "Estimation of time-to-collision maps from first order motion models and normal flows". In *Proceedings of the 11th International conference on Pattern Recognition*, Augest 1992.
13. D. W. Murray and B. F. Buxton. "Scene segmentation from visual motion using global optimization". *IEEE Trans. on Pattern Analysis and Machine Intelligence*, PAMI-9(2):220–228, Mar. 1987.
14. D. Shulman and J. Hervé. "Regularization of discontinuous flow fields". In *Proc. Workshop on Visual Motion*, pages 81–85, Irvine, CA, Mar. 1989. IEEE Computer Society Press.
15. D. Terzopoulos and R. Szeliski. "Tracking with kalman snakes". In *Active Vision*, pages 3–20. The MIT Press, 1992.
16. W. B. Thompson, K. M. Mutch, and V. Berzins. "Edge detection in optical flow fields". In *Proc. of the Second National Conference on Artificial Intelligence*, pages 26–29, Aug. 1982.
17. N. Ueda and K. Mase. "Tracking moving contours using energy-minimizing elastic contour models". *Proc. ECCV*, pages 453–457, May 1992.
18. S. Xu and P.-E. Danielsson. "On computing time-to-collision". In *Proceedings of the 8th Scandinavian Conference on Image Analysis*, pages 1349–1356, 1993.
19. A. L. Yuille, D. S. Cohen, and P. W. Hallinan. "Feature extraction from faces using deformable templates". *Proc. CVPR*, pages 104–109, June 1989.