34th Annual International Conference of the IEEE EMBS
San Diego, California USA, 28 August - 1 September, 2012

Time-to-contact maps for navigation with a low resolution visual prosthesis

Chris McCarthy and Nick Barnes

Abstract— **The perception of independently moving objects in the scene is an important capability for prosthetic vision, but is impeded by the limited resolution and dynamic range of current and near-term retinal prostheses. We propose a novel, biologically-inspired visual representation for prosthetic vision based on the recovery of time-to-contact (τ) with surfaces in the scene. The representation directly encodes the extent of motion towards the observer, placing greatest emphasis on objects posing an imminent threat of collision. Our results suggest the proposed τ-based representation may facilitate earlier perception of incoming objects, and provide clearer distinction between moving objects and the static structure of the scene compared with intensity and depth-based scene representations.**

I. INTRODUCTION

The last decade has seen significant progress in the development of retinal visual prostheses. In most cases, scene imagery is obtained via high resolution digital images captured from one or more head mounted cameras [1], [2]. Vision processing is then employed to translate the image data to some condensed encoding of the scene, suitable for transfer via eletrical stimulation of retinal ganglion cells, and onto the visual cortex. The result is a so called *phosphene image* [3]: an array of phospherous light spots, each loosely corresponding to one stimulating electrode (though interactions between electrodes almost certainly occur). The brightness and size of each phosphene varies with the amount of current delivered, allowing images of the scene to be rendered.

Enabling safe mobility in dynamic environments is a challenging and important problem for prosthetic vision. Current and near-term retinal prostheses are severely limited in both resolution and dynamic range. This has motivated researchers to consider efficient visual representations of the scene that convey as much of the scene structure as possible, and maximise functional outcomes for implantees. In the context of mobility, prosthetic and simulated prosthetic vision (SPV, see [4], [5] for a review)) studies have considered basic navigation tasks using intensity-based scene representations (*i.e.*, phosphene brightness conveys scene luminance), typically in high contrast environments [6], [7], [8], [9], [10], [11]. More recently, investigations using a depth-based representation of the scene (*i.e.*, phosphene brightness conveys surface proximity) [12] have also been conducted, showing potential advantages in the presence of obstacles. These studies have primarily focussed on mobility in static environments.

Most real-world environments (*e.g.*, urban and office environments) are dynamic. To achieve safe and efficient mobility in such environments, prosthetic vision must facilitate fast and accurate perception of imminent collisions, while also conveying the static scene structure. While depth-based representations offer potential advantages for achieving the latter (compared with intensity-based), the perception of relative motion between observer and objects in the scene remains difficult due to a lack of visual scene features from which to infer motion. Temporal changes of depth provide an obvious cue for motion, but require sufficient dynamic range to observe the change in time to take appropriate action. The cognitive demands associated with perceiving depth changes induced by multiple free moving objects are also likely to impede functional outcomes.

An alternative visual representation is to encode the proximity of surfaces with respect to their *time-to-contact*. That is, the ratio of an object's relative velocity towards the observer, and its distance from the observer. The key difference is that 'depth' becomes a temporal measure rather than a spatial one. Thus, an approaching object is considered to be *closer* than another at the same absolute distance from the observer if its velocity towards the observer is greater.

There is strong evidence of the use of time-to-contact for motor control across a wide range of animal species [13], [14], [15]. In most cases this is attributed to looming sensitive neural mechanisms that measure the apparent expansion of intensity patterns on the retina. Such visuo-motor schemes have also been successfully applied in robotic systems (*e.g.*, [16], [17], [18]).

The use of time-to-contact offers several advantages for prosthetic vision:

1) it provides a direct and immediate encoding of potential contact with surfaces in the scene;
2) it relates surface proximity to observer motion (rather than gaze direction), and thus places greatest emphasis on objects posing a direct threat of collision with respect to this motion;
3) it has a biological basis, providing a cue known to be utilised for numerous animal visuo-motor control tasks [19].

In this paper we introduce time-to-contact as an important visual cue for mobility with prosthetic vision. We propose a

All authors are with NICTA Canberra Research Laboratory, Canberra ACT, Australia and College of Engineering and Computer Science, Australian National University, Canberra ACT, Australia.

The authors acknowledge the contributions of Paulette Lieby and Adele Scott, authors of the simulated prosthetic vision software used in this paper. This research was supported by the Australian Research Council (ARC) through its Special Research Initiative (SRI) in Bionic Vision Science and Technology grant to Bionic Vision Australia (BVA). NICTA is funded by the Australian Government as represented by the Department of Broadband, Communications and the Digital Economy and the Australian Research Council through the ICT Centre of Excellence program.

978-1-4577-1787-1/12/$26.00 ©2012 IEEE

novel visual representation for prosthetic vision to facilitate safe mobility in dynamic environments. We present quantitative and qualitative results demonstrating the effectiveness of the proposed time-to-contact representation (τ-based) for emphasising free-moving incoming objects, using simulated prosthetic vision.

II. Time-to-contact

Time-to-contact is defined as the ratio of the surface distance and its component of velocity toward the observer such that for a viewing direction $\hat{p} \in \mathbb{R}^3$:

$$\tau(\hat{p}) = \frac{R(\hat{p})}{(\hat{p} \cdot \vec{t})}, \tag{1}$$

where $R(\hat{p})$ is the radial depth of the surface along $\hat{p}$, and $\vec{t} \in \mathbb{R}^3$ is the translational motion of the surface with respect to the observer.

There are numerous ways to compute time-to-contact. Most commonly it is computed from the divergence of the optical flow field [16], [17], [18]. This, however, is complicated by the confounding of surface gradient and translational motion in the deformation component of the measured divergence [16], [20]. Thus, in general, only a bound on time-to-contact can be computed from optical flow. However, given a dense depth image of the scene (*e.g.*, stereo disparity from binocular images, kinect, *etc.*), time-to-contact may be computed precisely from temporal changes of depth.

Let $Z_t(x, y)$ and $Z_{t-1}(x, y) \in \mathbb{R}$ be dense depth images, providing for each image point the depth of the point projecting to that location (we assume a pinhole camera model). Let $I_t(x, y)$ and $I_{t-1}(x, y)$ be the corresponding intensity images for each depth image, assumed to be aligned with each depth image. Let $F_t(x, y) \in \mathbb{R}^2$ be the set of point correspondences (expressed as a 2-vector) for all points in I_t such that each point is mapped to its origin in the previous frame. In our results, we compute these correspondences using a pyramidal implementation of Lucas and Kanade's gradient-based technique as described by Bouguet [21] and provided in the OpenCV developers library[1]

From F_t, we compute the relative velocity, Δ_t, between observer and the scene along each viewing direction (*i.e.*, $\hat{p} \cdot \vec{t}$) such that:

$$\Delta_t(x, y) = Z_{t-1}(F(x, y)) - Z_t(x, y), \tag{2}$$

thereby giving the component of velocity parallel to the optical axis of the camera. To minimise the effects of noisy depth estimates and erroneous point correspondences, median filtering is applied to the resulting Δ_t image.

Assuming that the observer's translational motion is approximately aligned with the optical axis[2], the final time-to-contact map, $\tau_t(x, y) \in \mathbb{R}$, is then computed as:

$$\tau_t(x, y) = \frac{Z_t(x, y)}{\Delta_t(x, y) \cos(\theta_\text{p})}, \tag{3}$$

where θ_p is the viewing angle of the image point (x, y) with respect to the image origin (computed from known intrinsic camera parameters).

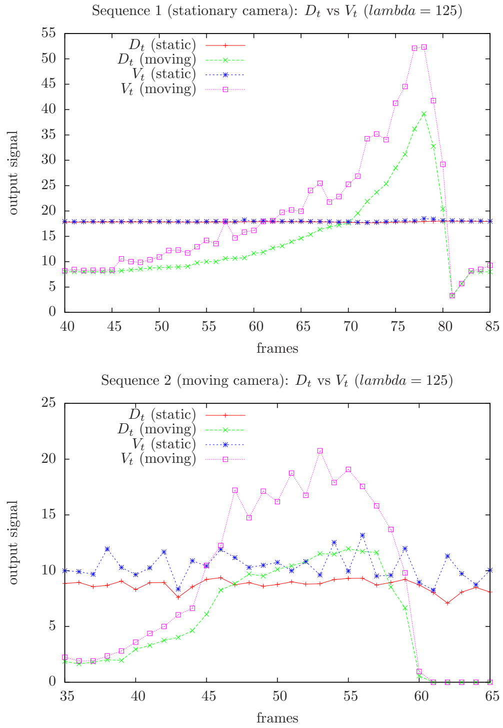

Fig. 1. Results comparing output response from D_t (*i.e.*, depth-based) and V_t (τ-based) from both sequences. Figure plots output signals (from which phosphene brightness is determined) using mean response from static and moving object measurement windows (see Figure 2 for their locations).

A. A τ-based representation for prosthetic vision

We now describe our proposed use of time-to-contact as a novel τ-based representation for navigation and mobility in dynamic environments. Time-to-contact is inherently defined with respect to motion in the scene, and thus provides no structural information under motionless conditions. We therefore propose a unified visual representation to handle all conditions, placing τ_t^{-1} in linear combination with the current disaprity image, $D_t \in \mathbb{R}$ of the scene such that:

$$V_t(x, y) = \max\Big(v_\text{min}, \min\Big(D_t(x, y) + \frac{\lambda}{\tau_t(x, y)}, v_\text{max}\Big)\Big), \tag{4}$$

where λ is a scale factor determining the extent of influence of τ_t on the representation ($\lambda = 125$ in all results presented), and v_min and v_max are lower and upper bounds in the resulting response. Note that the reciprocol, τ^{-1}, ensures faster approaching objects induce a larger and exponentially increasing response as an object approaches. This is akin to visual looming, from which time-to-contact is typically inferred. Conversly, the entire term vanishes when no motion is present (*i.e.*, $\tau = \infty$), thereby defaulting to a standard depth-based scene representation.

[1] http://opencv.willowgarage.com.

[2] note that robust and efficient methods for computing camera egomotion exist, and may also be applied if required.

III. Results

Two image sequences were constructed using a kinect sensor to validate and compare the τ-based representation. *Sequence 1* shows a corridor scene containing an overhanging and motionless black box, and a person walking from the back of the corridor (and behind the box), towards the sensor. *Sequence 2* depicts the sensor *moving* along an office corridor at approximately constant velocity while a person enters the corridor and walks past. Sample frames from both sequences are given in the left-most column of Figure 2[3].

A. Assessment of output response

Figure 1 shows plots of the output response of, D_t, and the τ-based representation, V_t, over the course of both sequences. Measurements for both were taken as the average response from two 50×50 measurement windows, positioned on a: 1. static-object, and 2. moving-object region in the scene. (measurement regions for both sequences are shown in Figure 2). Measurement locations remained constant throughout each sequence.

Figure 1(a) shows both D_t and V_t remain constant and the same within the static-object window. However, a clear distinction is apparent in output responses from the moving-object window, with a clear increase in response resulting from the measured time-to-contact of the moving object. At the point where the walker and box are equidistant from the camera (frame 70), the output response from V_t is 1.41 times greater than D_t in the moving-object region.

Figure 1(b) shows plots obtained for Sequence 2. In this case the camera is in motion, and thus a non-zero τ contribution is apparent across the scene. However, a clear and stable increase in response is observed from the moving-object measurement window, indicating that V_t is correctly emphasising the approaching walker. Plots from the static-object window (positioned on the corridor floor) indicate reasonable constancy of the response.

Figure 2 visualises these results, showing from left-to-right: the original image, the disparity image, D_t, the recipricol time-to-contact image, τ_t^{-1}, and the proposed representation, V_t. . The approaching walker in both sequences appears brighter in V_t images (right-most column) compared with D_t (second column). τ^{-1} images (column 3) confirm it is only approaching surfaces being emphasised in V_t.

B. Simulated prosthetic image comparison

Figure 3 shows example phosphene images from both sequences, comparing visualisations obtained using: intensity-based, depth-based and the proposed τ-based representation. The phosphene images are rendered using the system described in [12], with 98 phosphenes and 8 (3-bit) brightness levels[4]. In both cases, the τ-based representation appears to provide the clearest visualisation of the approaching walker while still providing the static structural cues apparent in the depth-based representation.

IV. Conclusion

Safe mobility in dynamic environments is an important capability for prosthetic vision. We have proposed a novel, biologically-inspired visual scene representation that encodes the time-to-contact of surfaces in the scene, emphasising those objects posing an imminent threat of collision. Results demonstrate how the proposed τ-based representation may be used to provide earlier perception of incoming objects (via increased phosphene brightness) than depth alone. Visual comparisons with intenstity- and depth-based representations in simulated prosthetic vision suggest potential advantages for navigation with current and near-term visual prostheses.

References

[1] L. Hallum, G. Suaning, and N. Lovell, "Contribution to the theory of prosthetic vision," *ASAIO journal*, vol. 50, no. 4, p. 392, 2004.

[2] W. Dobelle, "Artificial vision for the blind by connecting a television camera to the visual cortex," *ASAIO journal*, vol. 46, no. 1, p. 3, 2000.

[3] G. Brindley and W. Lewin, "The sensations produced by electrical stimulation of the visual cortex," *The Journal of Physiology*, vol. 196, no. 2, p. 479, 1968.

[4] S. C. Chen, G. J. Suaning, J. W. Morley, and N. H. Lovell, "Rehabilitation regimes based upon psychophysical studies of prosthetic vision," *Journal of Neural Engineering*, vol. 6, no. 3, 2009, to appear.

[5] ——, "Simulating prosthetic vision: II. measuring functional capacity," *Vision Research*, vol. 49, no. 19, pp. 2329 – 2343, 2009.

[6] K. Cha, K. W. Horch, and R. A. Normann, "Mobility performance with a pixelized vision system," *Vision Research*, vol. 32, no. 7, pp. 1367 – 1372, 1992.

[7] G. Dagnelie, P. Keane, V. Narla, L. Yang, J. Weiland, and M. Humayun, "Real and virtual mobility performance in simulated prosthetic vision," *Journal of Neural Engineering*, vol. 4, pp. S92–S101, 2007.

[8] D. J. A. and A. J. Maeder, "Mobility enhancement and assessment for a visual prosthesis," in *SPIE Medical Imaging 2004: Physiology, Function, and Structure from Medical Images*. International Society for Optical Engineering, 2004.

[9] J. Dowling, W. Boles, and A. Maeder, "Mobility assessment using simulated artificial human vision," in *Proceedings of the 2005 Workshop on Computer Vision Applications for the Visually Impaired (CVAVI)*, june 2005.

[10] M. S. H. N. Parikh and J. D. Weiland, "Mobility experiments with simulated vision and peripheral cues," in *Proceedings of the Association for Research in Vision and Ophthalmology (ARVO)*, 2010.

[11] M. S. Humayun, L. da Cruz, G. Dagnelie, S. Mohand-Said, P. Stanga, R. N. Agrawal, and R. J. Greenberg, "Interim performance results from the second sight Argus II retinal prosthesis study," in *Proceedings of the Association for Research in Vision and Ophthalmology (ARVO)*, 2010.

[12] P. Lieby, N. Barnes, C. McCarthy, N. Liu, H. Dennett, J. Walker, V. Botea, and A. Scott, "Substituting depth for intensity and real-time phosphene rendering: Visual navigation under low vision conditions," in *Proceedings of IEEE EMBC 2011*. IEEE, 2011, pp. 8017–8020.

[13] F. C. Rind, "Collision avoidance: from the locust eye to a seeing machine," in *From Living Eyes to Seeing Machines*, M. V. Srinivasan and S. Venkatesh, Eds., 1997, pp. 105–125.

[14] R. M. Robertson and A. G. Johnson, "Collision avoidance of flying locusts: steering torques and behaviour," *Journal of Experimental Biology*, vol. 183, pp. 35–60, 1993.

[15] M. V. Srinivasan, S. W. Zhang, J. S. Chahl, E. Barth, and S. Venkatesh, "How honeybees make grazing landings on flat surfaces," *Biological Cybernetics*, vol. 83, pp. 171–83, 2000.

[16] R. C. Nelson and J. Y. Alloimonos, "Obstacle avoidance using flow field divergence," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 11, no. 10, pp. 1102–6, 1989.

[17] C. McCarthy and G. Metta, "Biologically-inspired time and location of impact prediction from optical flow," in *Proceedings of the IEEE International Conference on Robotics and Automation (ICRA)*. IEEE, 2011, pp. 6199–6204.

[3] Videos showing full image sequence are available at `http://cecs.anu.edu.au/~cdmcc/taumaps`

[4] There are 98 electrodes on Bionic Vision Australia's wide view implant. Around 8 levels of brightness is approximately consistent with current results from human trials of retinal prosthetic vision.

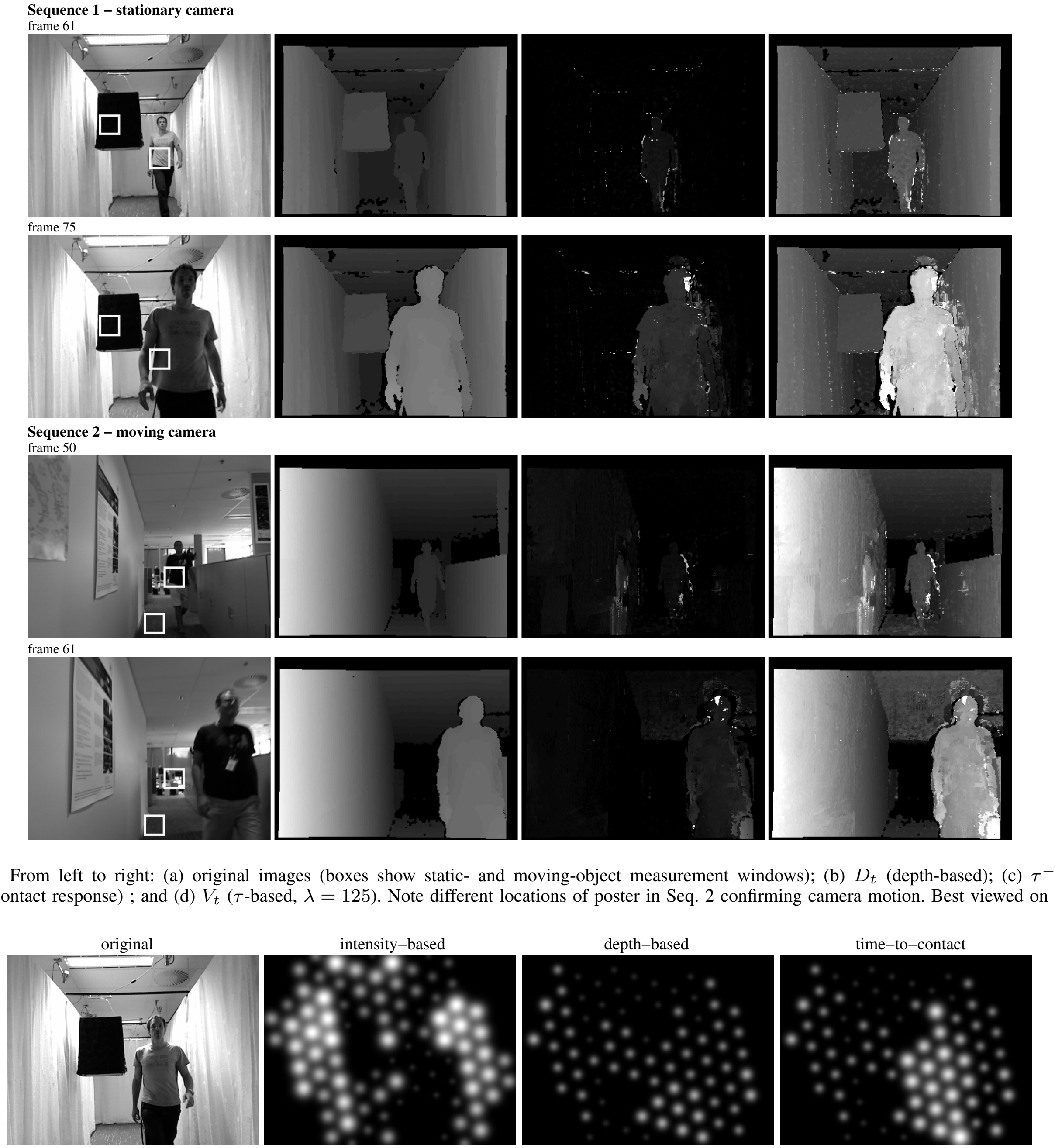

Fig. 2. From left to right: (a) original images (boxes show static- and moving-object measurement windows); (b) D_t (depth-based); (c) τ^{-1} (*i.e.*, time-to-contact response) ; and (d) V_t (τ-based, $\lambda = 125$). Note different locations of poster in Seq. 2 confirming camera motion. Best viewed on screen.

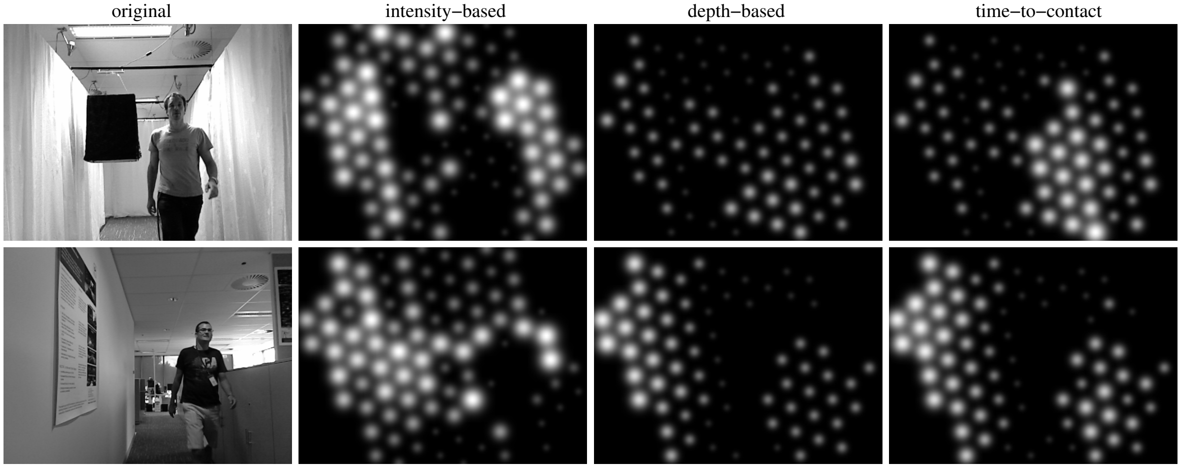

Fig. 3. Qualitative comparison of visual scene representations using simulated prosthetic vision (98 phosphenes, 8 brightness levels) showing from left to right: (a) the original image, (b) intensity-based, (c) depth-based and (d) τ-based representations.

[18] C. McCarthy and N. Barnes, “A unified strategy for landing and docking using spherical flow divergence,” *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2012, in press.

[19] M. Lappe, “Building blocks for time-to-contact estimation by the brain,” in *Time-to-contact*, ser. Advances in Psychology, H. Hecht and G. Savelsbergh, Eds. Amsterdam, The Netherlands: Elsevier, 2004.

[20] M. Subbarao, “Bounds on time-to-collision and rotational component from first-order derivatives of image flow,” in *Computer Vision, Graphics, and Image Processing*, vol. 50, 1990, pp. 329 – 41.

[21] J.-Y. Bouguet, “Pyramidal implementation of the Lucas Kanade feature tracker description of the algorithm,” in *OpenCV Documentation*. Intel Corporation, 2000.