Object exploration using vision and active touch

Chuanyu Yang, Nathan F. Lepora, *Member, IEEE*

***Abstract*— Achieving object exploration with passive vision and active touch has been under investigation for thirty years. We build upon recent progress in biomimetic active touch that combines perception via Bayesian evidence accumulation with controlling the tactile sensor using perceived stimulus location. Here, passive vision is combined with active touch by providing a visual prior for each perceptual decision, with the precision of this prior setting the relative contribution of each modality. The performance is examined on an edge following task using a tactile fingertip (the TacTip) mounted on a robot arm. We find that the quality of exploration is a U-shaped function of the relative contribution of vision and touch; moreover, multi-modal performance is more robust, completing the contour when touch alone fails. The overall system has several parallels with biological theories of perception, and thus plausibly represents a robot model of visuo-tactile exploration in humans.**

I. INTRODUCTION

Thirty years after seminal work on integrating vision and touch for object recognition [1]–[3], how much progress has been made? Work from 1988 used passive stereo-vision and a tactile probe (with 128 taxels) to refine sparse 3D visual contours by actively exploring the surface with touch for objects such as disks and cups [2]. Considering a snapshot of progress from a recent 2015 workshop on 'See and Touch' [4], the focus has maybe shifted to interacting physically with objects, yet the general problem of how to combine 3D vision and touch still remains unsolved.

So why has progress been slow on using vision and touch to explore and recognize objects? It seems unlikely that limitations in computer vision and AI is the cause, as both fields have expanded enormously; similarly, the development of tactile and 3D vision sensors has received an enormous amount of attention. In our view, the likely cause is that active touch (*i.e.* combining tactile sensing and sensor control) has been hard to implement in practice, as rather presciently said back in 1988: 'Active touch sensing provides accurate and robust shape information, but it exacts its price for this information by demanding powerful control of the medium, which makes it difficult to use' [2, Sec. 4].

The aim of this paper is to investigate how passive 3D vision can combine with and benefit active tactile exploration of an object contour. We build upon recent progress in tactile exploration [5], [6] based on biomimetic active touch [7]–[9] that combines perception via Bayesian evidence accumulation with controlling the tactile sensor via perceived stimulus location. Robust exploratory behaviour then emerges from a control policy that maintains the sensor on the contour while moving along it [5], [6]. Here, passive vision is combined with active touch by giving a visual prior for each perceptual decision that is updated with evidence from tactile sensing. The precision of the prior then sets the relative contribution of vision and touch to the overall exploration of the object.

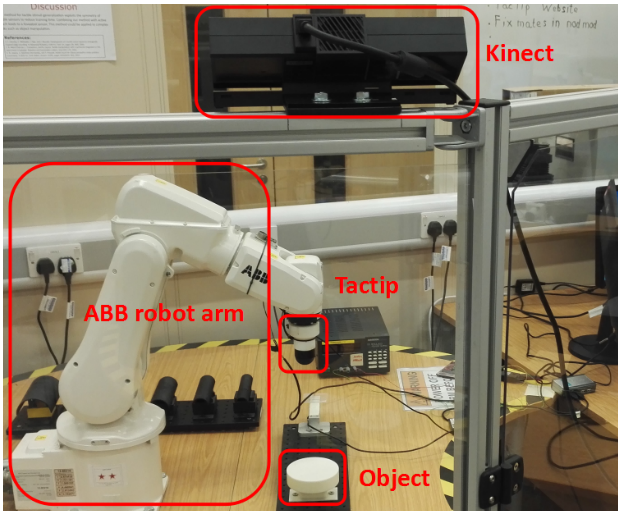

Fig. 1. Tactile robotic system, comprising a tactile fingertip (the TacTip) mounted as an end effector on a 6-dof ABB robot arm, and a 3D-vision system comprising a Kinect V2 camera. The aim is to explore the object.

For validation, we consider contour following around a circular disk with a biomimetic tactile fingertip mounted on a robot arm, using a Kinect V2 camera for 3D vision (Fig. 1). Being situated away from the arm (on a safety barrier), the camera images the object at low resolution and an oblique angle, resulting in an inaccurate contour. However, we show it provides a good visual prior for tactile exploration, since: (i) the quality of contour following is a U-shaped function of the relative contribution of vision and touch; and (ii) the robustness is better with touch and vision, completing the entire contour when touch alone fails due to becoming lost.

II. BACKGROUND AND RELATED WORK

As covered in the introduction, seminal work in the 1980s first combined vision and robot touch for 3D object exploration and recognition [1], [2]. The area has since diversified, which we survey briefly by separating into various tasks:

1) Perception/classification: Some surface features are difficult to classify using vision or touch alone, and so both modalities must be combined. Visuo-tactile methods have attained attributes such as elasticity, mass and relational

NL was supported in part by a grant from the Engineering and Physical Sciences Research Council (EPSRC) on 'Tactile superresolution sensing' (EP/M02993X/1) and a Leadership Award from the Leverhulme Trust on 'A biomimetic forebrain for robot touch' (RL-2016-39).

CY and NL are with the Department of Engineering Mathematics and Bristol Robotics Laboratory, University of Bristol, Bristol, UK (Email: n.lepora@bristol.ac.uk).

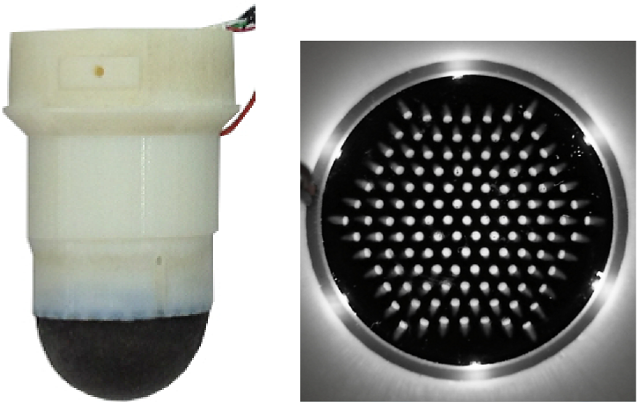

Fig. 2. Left: The tactile fingertip (TacTip) comprising a compliant sensitive tip and a housing for the electronics and internal camera. Right: image from internal camera showing the array of sensing elements.

constraints [10] and object pose [11], [12]. Shape has been determined with methods such as tactile glances at discrete points on the object [13], visual and tactile feedback from grasping [14], combining visual and tactile exploratory procedures [15] and visuo-tactile fusion [16].

2) Exploration/mapping: Another application of combined vision and and touch is exploring the environment to build a surface map or model. Early work in the field focussed on using passive stereo vision and active touch to explore object contours and features [1]–[3]. Following work has combined touch and 3D vision to rapidly label surface features [17], map object surfaces [18], match tactile features to visual maps [19] and nest visual and tactile control loops to improve surface exploration [20], [21].

3) Grasping/manipulation: An important problem is for a robot to grasp and/or manipulate an object using fingertip tactile sensing and a camera. Early work focussed on the complementarity of vision and touch for grasping, either using vision to estimate the large-scale shape and touch for small-scale geometric and force information [22], or using touch to solve the occlusion problem for estimating contact location when calculating grasp forces from joint encoders [23]. Recent research has focussed more on in-hand object manipulation, with some of the above-mentioned work using pose-estimation [12] or visuo-tactile control [20], [21].

In this paper, we consider how to combine vision and touch to perform exploratory contour following around edges of unknown objects. This work is based on previous studies of active touch for tactile exploration [5], [6], here extended to introducing visual sensing into the robotic system.

III. METHODS

A. Robotic system

The robotic system comprises a tactile fingertip (TacTip) mounted on a robot arm, with a Kinect V2 vision sensor mounted nearby to image the task space (Fig. 1). Individual components are described below.

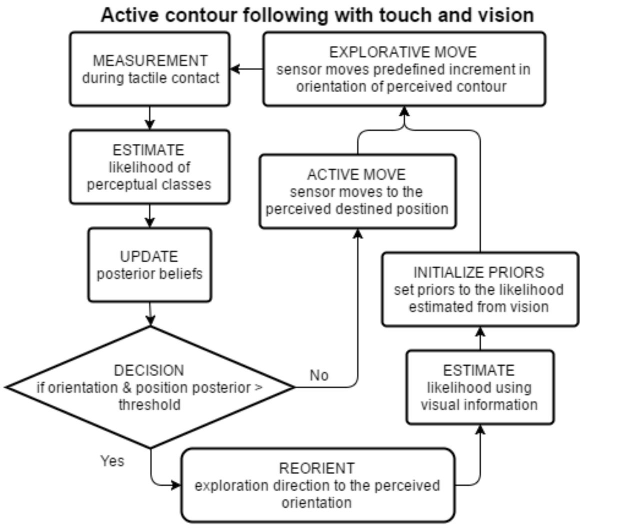
Active contour following with touch and vision
MEASUREMENT
during tactile contact
EXPLORATIVE MOVE
sensor moves predefined increment in orientation of perceived contour
ESTIMATE
likelihood of perceptual classes
ACTIVE MOVE
sensor moves to the perceived destined position
UPDATE
posterior beliefs
INITIALIZE PRIORS
set priors to the likelihood estimated from vision
DECISION
if orientation & position posterior > threshold
No
ESTIMATE
likelihood using visual information
Yes
REORIENT
exploration direction to the perceived orientation

Fig. 3. Active exploration algorithm using vision and touch. During each perceptual decision, sensory data from discrete tactile contacts feeds into a likelihood model that updates evidence for radial displacement and edge angle, which is used to move the sensor radially to maintain edge contact. After the evidence crosses a threshold, the exploration direction is reoriented to the perceived edge angle and the evidence initialized to the visual priors.

1) The Tactile fingerTip (TacTip): In this study, we use a soft biomimetic optical tactile sensor known as the TacTip (Fig. 2, left). It is a biologically-inspired device based upon the deformation of the epidermal layers of the human glabrous skin [24]. The TacTip consists of several components. The tip comprises a black flexible outer skin (tango black) containing a clear gelatinous polymer held with a 3 mm thick transparent acrylic window; on the inside of the skin are 127 pins tipped by white markers, which transduce deformation of the membrane into visible movement of pins. The 3D-printed body of the sensor holds internal LED lighting and a Microsoft Cinema HD webcam (resolution 640×480 pixels, sampled at $\sim$20 fps).

The particular design of the TacTip used here has a 40 mm diameter hemispherical sensing pad with 127 tactile pins arranged in a hexagonal lattice with pin-to-pin spacing $\sim$3 mm (Fig. 2, right). Deformation of the sensing pad is transduced into pin movements, which are tracked optically using the webcam (details below).

2) Robot arm: The TacTip is mounted as an end-effector on an IRB 120 robotic arm (ABB Robotics). It is a compact and relatively lightweight (25 kg), 6 degree-of-freedom robot arm, with maximum horizontal reach 580 mm and maximum payload 3 kg. The robot can precisely and repeatedly position its end effector with an absolute repeatability of 0.01 mm.

3) 3D-vision system: The workspace is imaged with a Kinect V2 (Microsoft) RGB-D sensor. This Kinect includes a 1080p resolution video camera and a 512x424 pixel monovision infrared camera (plus emitter) with maximum detection range 8 m, allowing it to capture both a colored 2D image and a depth image of the scene at 30 fps.

4) Software architecture: A modular software framework is deployed whereby the main control and perception algorithms are implemented in MATLAB. The framework runs on a standard Windows 8 or 10 PC, for compatibility with the Kinect V2 SDK released by Microsoft.

The PC sends control commands to the robot arm via TCP/IP ports and receives the TacTip data via USB. We used an IronPython client to convert MATLAB outputs into variables to interface with the robot controller (a RAPID API) that commands the arm movements. Simultaneously, a python server on the PC implements image capture and preprocessing to quantify surface deformation of the TacTip by tracking the internal pins with opencv (http://opencv.org/). Similar methods are used in other recent papers [6], [7].

B. Algorithmic methods: Biomimetic active touch

Biomimetic active touch is defined by three principles based on biological perception (Fig 3): (i) an underlying evidence accumulation part for decision making; (ii) an action selection part enacted during the decision making; and (iii) sensory encoding of how percepts relate to stimuli. A summary is given here; for more details we refer to ref. [7].

In the following we will use measurement model of the tactile data to give likelihoods for discrete angle θ_i and radial displacement r_l classes from the contact data, with $i \in [1, N_{\rm id}]$ and $l \in [1, N_{\rm loc}]$. The model inputs the tactile sensor values $s_k(j)$ for data dimension $k \in [1, N_{\rm dims}]$ and time-sample $j \in [1, N_{\rm samples}]$. Typically, we use $N_{\rm dims} = 127 \times 2$ data dimensions (127 pins with x- and y-components), and $N_{\rm samples}$ ranging from 20-25 for a single tap z_t (about 1 sec of data). We also use $N_{\rm id} = 18$ angle classes spanning 360 degs and $N_{\rm loc} = 20$ radial displacement classes spanning 20 mm from free-space to completely on the object. These values were chosen to give a reasonable amount of data for training the classifier, and are consistent with related studies.

1) Perceptual evidence accumulation: Bayes' rule is applied recursively after each contact z_t ($t = 1, 2, 3...$) to update the posterior beliefs for each perceptual class, using the likelihoods $P(z_t|r_l, \theta_i)$ of the contact data

$$P(r_l, \theta_i|z_{1:t}) = \frac{P(z_t|r_l, \theta_i)P(r_l, \theta_i|z_{1:t-1})}{P(z_t|z_{1:t-1})}, \qquad (1)$$

with the normalization term from the marginal probabilities $P(z_t|z_{t-1})$ of the current contact given prior contacts

$$P(z_t|z_{1:t-1}) = \sum_{l=1}^{N_r} \sum_{i=1}^{N_\theta} P(z_t|r_l, \theta_i)P(r_l, \theta_i|z_{1:t-1}). \qquad (2)$$

A key aspect of the formalism here is that the evidence accumulation from touch can begin from a prior $P(r_l, \theta_i|z_0)$ set from the passive vision (defined below in Sec. III-C.3).

The perception is complete when a marginal belief for edge orientation reaches a decision threshold $p_{\rm dec}$, when the *maximal a posteriori* estimate of the angular class is taken:

$$\text{if any } P(\theta_i|z_{1:t_{\rm dec}}) = \sum_{l=1}^{N_r} P(r_l, \theta_i|z_{1:t_{\rm dec}}) > p_{\rm dec}$$
$$\text{then } \theta_{\rm dec} = \arg\max_{\theta_i} P(\theta_i|z_{1:t_{\rm dec}}). \qquad (3)$$

The decision threshold is a free parameter that trades off the number of contacts $t_{\rm dec}$ to make a decision against decision accuracy (here set at $p_{\rm dec} = 2/N_\theta$ to give $t_{\rm dec} \sim 3$ taps).

During the perception, an intermediate estimate of the radial displacement will be used for active perception

$$r_{\rm est}(t) = \arg\max_{r_l} \sum_{i=1}^{N_\theta} P(r_l, \theta_i|z_{1:t}), \qquad (4)$$

which feeds into the action selection described below.

2) Action selection: These actions are selected with a control policy that inputs the last perceived angle and current estimates of radial displacement, with output action components comprising a tangential exploratory move and a radial corrective move.

The exploratory component of the action moves the sensor tangentially along the edge by a fixed amount Δe (here set to a gain g times 3 mm) along the last perceived angle $\theta_{\rm dec}$.

The corrective component of the action moves the sensor radially towards a pre-set displacement $r_{\rm fix}$ from the edge. Its direction of movement $\theta_{\rm dec} + 90°$ is orthogonal to the last complete angle decision and its magnitude is proportional to the currently estimated radial displacement

$$\Delta r(t) = \pi_r \left[P(r_l, \theta_n|z_{1:t})\right] = \left[\, g\,(r_{\rm fix} - r_{\rm est}(t))\,\right]_l . \qquad (5)$$

Here the sensor is fixated to the middle of the perceptual range $r_{\rm fix} = 0$ mm, which is aligned to centre on the edge. The notation $[\cdot]_l$ represents rounding down to the nearest class r_l. The same gain g applies both action components.

Following previous work on biomimetic active touch [7], after every action for active perception (radially along the normal), a compensatory transformation of the perceptual beliefs is made to maintain an allocentric belief frame

$$P(r_l, \theta_i|z_{1:t}) \leftarrow P([r - \Delta r(t)]_l, \theta_i|z_{1:t}). \qquad (6)$$

For simplicity, the (undetermined) beliefs shifted from outside the location range are assumed uniformly distributed.

3) Sensory encoding: Given a test tap z with samples $s_k(j)$ discretized into bins $b_k(j)$, the measurement model is built from the mean likelihood:

$$\log(P(z|r_l, \theta_i)) = \sum_{j=1}^{N_{\rm samples}} \sum_{k=1}^{N_{\rm dims}} \frac{\log(P(b_k(j)|r_l, \theta_i))}{N_{\rm samples}N_{\rm dims}}, \quad (7)$$

assuming statistical independence between all data dimensions k and time samples j. The sensor values s_k for data dimension k, which are then binned into 100 equal intervals I_b with sampling distribution given by the normalized histogram counts over all training data for each class.

C. Algorithmic methods: Visual perception

Vision here provides supplementary information to active touch. The imaged contour is encoded as probability distribution over the radial displacement r and edge angle θ that can be then fused with the likelihoods for tactile perception.

While the focus of this paper is on the fusion of vision with touch, the extraction of the contour is a non-trivial vision

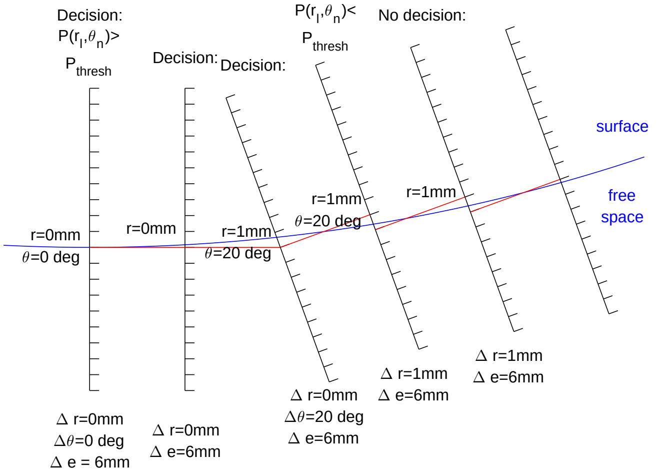

Fig. 4. Exploratory tactile servoing, shown over a few steps of the control loop from Fig. 3 (with unit gain for the exploration step size in this example).

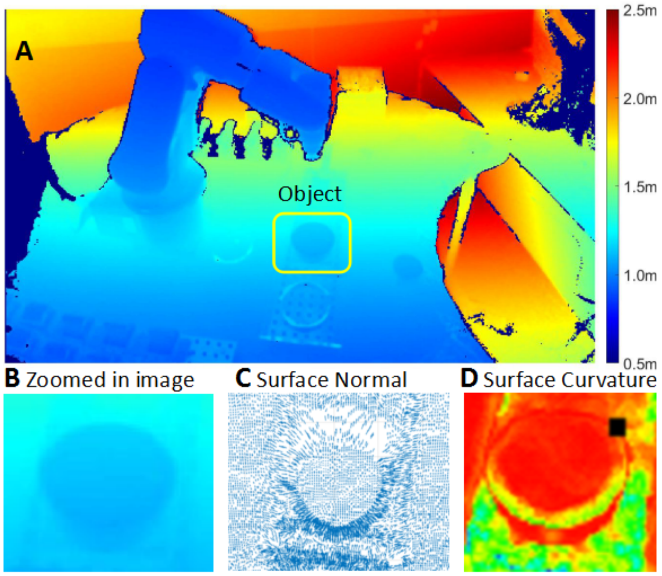

Fig. 5. A. Truncated depth image of the workspace obtained from the Kinect. B. Zoomed in depth image of the target object. C. Calculated surface normals of the object. D. Calculated surface curvatures of the object.

problem and our approach is detailed here. Other vision-base methods for contour extraction exist, but are outside the scope of this study which focuses on visuo-tactile fusion.

1) Surface normal and curvature: Following [25], [26], the surface normals and curvatures are found by applying PlanePCA to the Kinect depth image (Fig. 5). The method considers a query point $\vec{q}$, and then applies Principal Component Analysis (PCA) to the covariance matrix C from the k nearest neighboring points $\vec{p}_i$ in the point cloud,

$$C(\vec{q}) = \sum_{i=1}^{k} (\vec{p}_i - \vec{p}_0)^T (\vec{p}_i - \vec{p}_0), \quad C \cdot \vec{v}_j = \lambda_j \cdot \vec{v}_j, \tag{8}$$

where $\vec{p}_0$ represents the centroid of the neighborhood, and λ_j and $\vec{v}_j$ are the three eigenvalues and eigenvectors of the matrix C. These eigenvalues are used to estimate the principle surface curvatures, with the least significant eigenvector (minimum eigenvalue) the estimated surface normal.

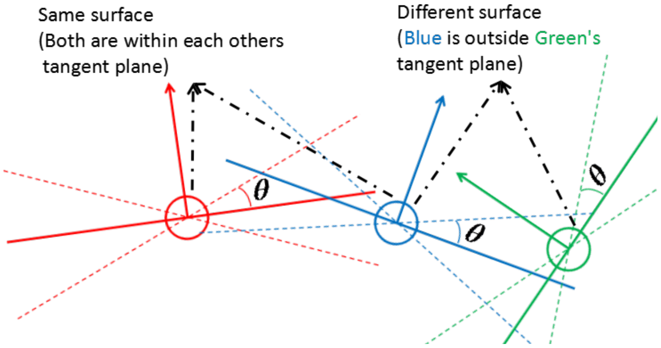

Fig. 6. Representation of 3 data points and their estimated tangent planes. The red and blue points lie within each others' tangent planes, and are considered connected. The green point is not connected to either point.

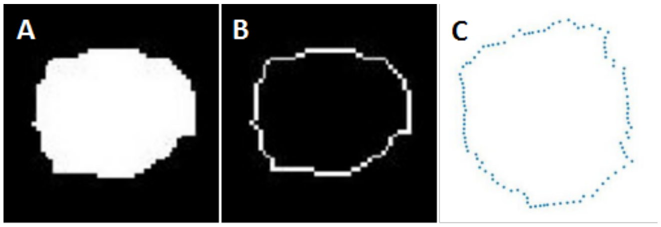

Fig. 7. A. Segmented Region. B. The perimeter of the segmented region. C. Top view of the perceived contour points of the object in 3D space

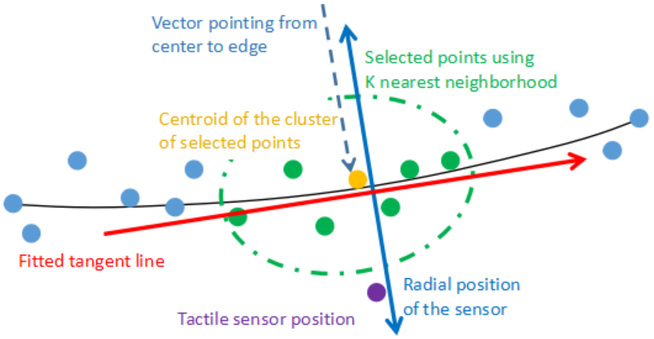

Fig. 8. Determination of visual prior. The edge orientation is estimated from the fitted tangent (red line) and the radial position from the centroid, from the k nearest neighbour points to the tactile sensor location.

Note that as the signs of the normals computed via PCA is ambiguous to $\pm\vec{n}_i$, to ensure consistent orientation over the entire point cloud they are reoriented $\vec{n}_i \cdot (\vec{p}_{\text{view}} - \vec{p}_i) > 0$ to point towards a single viewpoint $\vec{p}_{\text{view}}$.

2) Contour estimation: The point cloud, normals and curvatures found above are then 'organized' so that each point corresponds to one pixel in the 2D color image. This organized point cloud allows estimation of the surface bounded by the contour via region-growing segmentation [27].

First, a set of seed points are chosen from which to grow regions that encompass adjacent points if they satisfy the region membership criteria given below (Algorithm 1). These added points serve as new seeds for the next region-growing process, iterating until no more new neighboring points can

Algorithm 1 Surface Segmentation Algorithm

```
input: point cloud $P$, normals $N$, curvatures $C$
variable: available point list $A$, available seed list $S$
$A \leftarrow P$, $R \leftarrow \varnothing$, $S \leftarrow \varnothing$
Select an initial seed point $p_{\text{seed}}$
$A \leftarrow A - p_{\text{seed}}$, $R \leftarrow R \cup p_{\text{seed}}$, $S \leftarrow S \cup p_{\text{seed}}$
for $i = 1$ to size($S$) do
    Select a new seed point $p_i$ from the seed list $S$
    for $j = 1$ to size($A$) do
        Select point $p_j \in A$ from neighborhood of point $p_i$
        if $||\vec{p}_i - \vec{p}_j|| > d_0$ & $\lambda(\vec{p}_i) > \lambda_0$ & $\vartheta(\vec{n}_i, \vec{n}_j) > \vartheta_0$
        then
            break
        end if
        $A \leftarrow A - p_j$, $R \leftarrow R \cup p_j$, $S \leftarrow S \cup p_j$
    end for
end for
return Region $R$
```

be found. The three criteria for region membership are:

a) Distance-based criterion: The Euclidean distance between the seed point $\vec{p}_i$ and a selected neighbour point $\vec{p}_i$ must be below an allowed maximum $||\vec{p}_i - \vec{p}_j|| > d$.

b) Curvature-based criterion: Only points with low curvature are included $\lambda(\vec{p}_i) < \lambda_0$, to terminate on the surface edge.

c) Normal-based criterion: Neighbouring surface normals define tangent planes that must lie within an angle range of each other $\vartheta(\vec{n}_i, \vec{n}_j) < \vartheta_0$ (Fig. 6) [28].

Here we used values $d_0 = 0.5$, $\lambda_0 = 0.1$ and $\vartheta_0 = 8°$, which capture most points by the curvature and normal criteria, with distance-based criterion for outliers.

The algorithm output is the segmented region of interest R (Fig. 7). The perimeter of the 2D region is extracted and mapped to the 3D point cloud to obtain the 3D spatial coordinates of the surface contour.

3) Sensory encoding: The visual data is then re-encoded as a prior to be fused with the likelihood model of the tactile data, to perceive edge radial displacement and angle (r, θ).

We take the k-nearest neighbor points on the estimated contour from the tactile sensor position $\vec{p}$ (Fig. 8). A tangent line with unit direction $\vec{d}$ is fitted to these $k = 5$ points using linear regression (with sign $\pm\vec{d}$ constrained to $\vec{p}_{\text{center}} \cdot \vec{d} > 0$). Then the edge angle is given by $\vec{d} = (\cos\theta, \sin\theta, 0)$ and radial distance taken from the sensor to the tangent line.

The visual prior over positions and orientations of the edge is estimated from a Gaussian model over the 20 location classes r_l and the 18 orientation classes θ_i

$$P(r_l, \theta_i | z_0) = \frac{1}{2\pi\sigma^2} \exp\left[-\frac{(r_l - r)^2}{2\sigma^2\kappa_r^2} - \frac{(\theta_i - \theta)^2}{2\sigma^2\kappa_\theta^2} \right], \quad (9)$$

with standard deviation σ adjusted to change the precision of the visual prior and hence its influence when fused with the tactile sensory input. Parameters $\kappa_r = 20$ mm and $\kappa_\theta = 360°$ are normalization factors given by their ranges.

TABLE I

AVERAGE ANGULAR AND POSITIONAL ERROR AROUND A CONTOUR

Perceptual error	uni-modal vision	multi-modal touch and vision	uni-modal touch
Angular, $\bar{e}_\theta$	16.1°	5.8° ($\sigma = 0.30$)	13.4°
Positional, $\bar{e}_r$	6.9 mm	2.1 mm ($\sigma = 0.18$)	3.4 mm

IV. RESULTS

A. Edge following with uni-modal vision or touch

First, we verify that the tactile perception method (Sec. III-B) and the visual perception method (Sec. III-C) are individually able to follow the edge of the circular test object.

Using vision alone, the perceived edge using the Kinect camera gives a contour that roughly approximates the circular edge (Fig. 9A). This contour could be used to guide the sensor around the object edge with vision alone. Calculating the mean square error of perceived edge orientation relative to that of the actual edge around the entire loop gives an error $e_\theta = 16.1°$ (shown to left of Fig. 10).

Using touch alone, the tactile sensor is able to successfully complete an entire circuit of the object edge (Fig. 9E; tactile gain $g = 1$), as observed previously in other work [6]. The mean square error of perceived edge orientation relative to that of the actual edge around the entire loop is $e_\theta = 13.4°$ (shown to right of Fig. 10).

Therefore both vision and touch alone are able to successfully follow the contour, but each is not very accurate (angle errors $> 10°$). Also touch can be unreliable in some situations, in that sensor noise can cause the edge following to fail, as revealed by doubling the tactile gain to $g = 2$: after about 10 contacts, the sensor loses contact with the edge and becomes lost (Fig. 11D).

Our main thesis is that combined vision and touch will give more reliable task performance, which we examine in the next section. Interestingly, the visual contour is biased inside the object (Fig. 9A) whereas the tactile contour is biased outside (Fig. 9E), giving further support that combining these two modalities will improve edge-following performance.

B. Edge following with multi-modal vision and touch

The next experiments consider using touch and vision together to follow the circular edge. The control gain is kept at $g = 1$ (the same as for uni-modal touch in Sec. IV-A), but the precision σ that weights the visual and tactile information is varied from 10^{-2} (mainly vision) to 10^2 (mainly touch).

The principal effect of combining touch and vision is that the traced contours (Fig. 9B-D; red curves) both improve upon and lie intermediate to the contours from uni-modal vision (Fig. 9A; green curve) and uni-modal touch (Fig.9E; red curve). These observations are supported quantitatively from considering the angular and positional perceptual errors averaged around the contour (Figs 10A,B), which both show a continual improvement for multi-modal perception over uni-modal vision ($\sigma = 0$) and uni-modal touch ($\sigma = \infty$)

Overall, the best contour (Fig. 9C) is at an intermediate weighting of touch and vision ($\sigma = 0.30$), when the angular

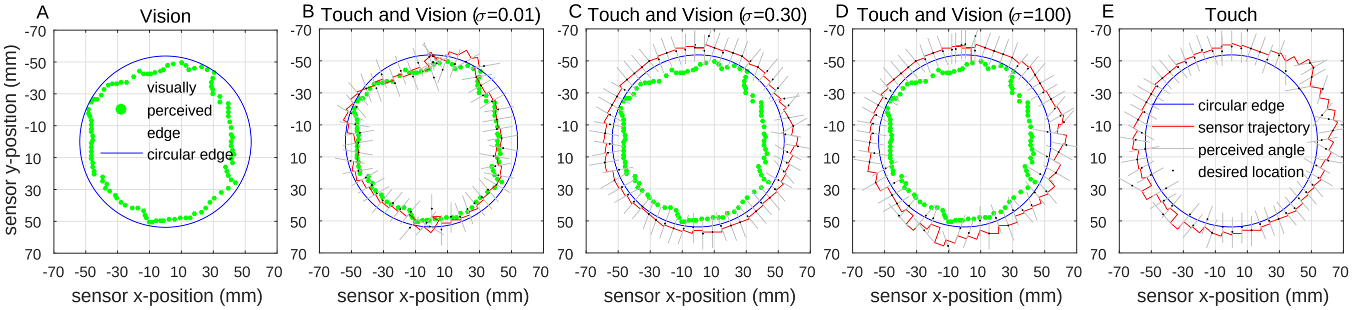

Fig. 9. Exploration trajectory under different combinations of vision and touch. The visually perceived edge (A-E, green markers) is inside the actual edge (blue circle). The trajectory from touch alone (E, red curve) lies outside this edge. Combining vision and touch gives intermediate trajectories (B-D) that have lower angular error than the uni-modal trajectories (A,E). The optimal trajectory occurs at an intermediate weighting (C) of vision and touch. Note that the overshoot in panels C-E is because the fixation point that the robot tries to follow is displaced from the edge.

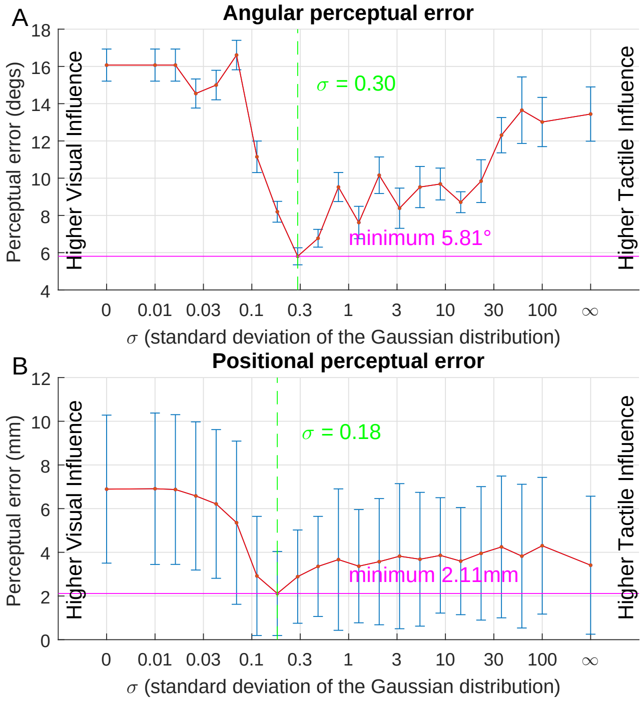

Fig. 10. Angular perceptual error and positional perceptual error averaged around the exploration path, plotted against the precision σ of the visual prior. Both error plots reach a minimum (0.30,0.18) in the central range.

perception error averaged around the contour is a minimum (Fig. 10A). At that combination of vision and touch, the average angular error $\bar{e}_\theta = 5.8°$ is much lower than that for uni-modal vision ($\bar{e}_\theta = 16.1°$) and uni-modal touch ($\bar{e}_\theta = 13.4°$). Similarly, the radial perceptual error $\bar{e}_r$ has a minimum at a nearby intermediate weighting of touch and vision ($\sigma = 0.18$). These values are summarized in Table I.

Curiously, the effect of including touch in uni-modal vision is different from including vision in uni-modal touch. For the edge-following task considered here, uni-modal vision gives poorer contour tracing than uni-modal touch (Table I). When vision is dominant, increasing the influence of touch over uni-modal vision ($10^{-2} \leq \sigma \leq 0.3$), causes the perceptual errors to initially change very little, but then improve precipitously to the optimal edge-following behaviour (Fig. 10, left-side). Meanwhile, when touch is dominant, increasing the influence of vision over uni-modal touch ($0.3 \leq \sigma \leq 2$), causes the perceptual errors to improve gradually to that same optimum (Fig. 10, right-side). This is likely because when σ is set too low (i.e. strong visual prior), the belief threshold is reached too early and the decision process terminates without gathering enough evidence.

C. *Robustness of multi-modal vision and touch*

Our final experiments examine if multi-modal perception benefits the robustness of task performance. Due to the exploratory nature of active touch, a wrong move may result in a failed task where the robot loses the edge and becomes ‘lost’; conversely, vision always gives a complete (but possibly inaccurate) contour. Therefore, we compare the performance of uni-modal touch with that of multi-modal touch and vision to discover whether introducing visual information into active touch improves system robustness.

Robustness is probed by varying the gain g that scales the magnitude of the tangential exploratory moves around the edge and the radial corrective moves during active touch. For uni-modal touch, increasing the gain from $g = 0.5, 1, 1.5, 2$ results in progressively poorer edge-following performance until task failure at $g = 2$ (Figs 11A-D, red curves). This observation is supported qualitatively with an increase in angular perception error with increased gain (Table II).

Multi-modal touch and vision improves the robustness of task performance, in not failing to complete the task at high gains (Figs 11E-H, red curves). In particular, for gain $g = 2$, multi-modal vision and touch are able to successfully perform a complete circuit of the circular edge (Fig. 11H), whereas uni-modal touch then fails the task (Fig. 11D). For these experiments the visuo-tactile weighting was set at $\sigma = 0.3$, corresponding to the optimum of the angular perceptual errors at gain $g = 1$ (Fig. 10A).

The improved robustness of touch and vision together is related to the quality of task performance, in that the multi-modal angular and positional perceptual errors are also smaller (Table II) and the sensor trajectory of the visual-tactile system is smoother than the tactile system (Fig. 11). This behaviour is expected because improved performance

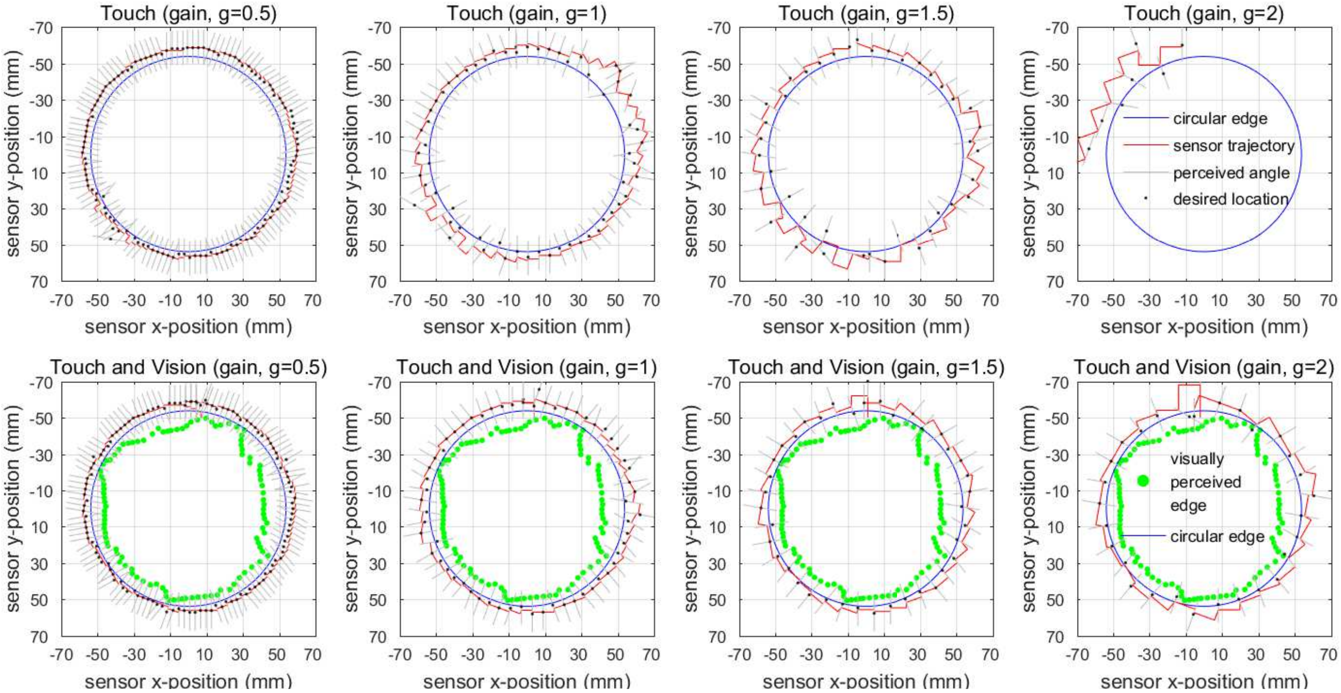

Fig. 11. Exploration trajectory for uni-modal touch (top row) and multi-modal vision and touch (bottom row). Trajectories (red curves) are considered at different values of the gain g that scales both the exploration step ($3g$ mm) and active perception step. While larger gains produce more variation in the trajectory, this can be partially compensated with vision. In consequence, touch and vision together completes the contour when touch alone fails.

TABLE II
AVERAGE ANGULAR ERROR, AGAINST VALUES OF THE CONTROL GAIN g THAT SCALES THE EXPLORATION AND ACTIVE PERCEPTION STEPS.

Method	Gain			
	0.5	1.0	1.5	2.0
Touch and Vision ($\sigma = 0.30$)	3.35	4.09	3.27	3.22
Touch	4.08	5.55	4.47	N/A

is accompanied by smaller corrective moves, giving greater stability in keeping the sensor on the object.

V. Discussion

This study investigated how to explore an object with passive 3D vision and active touch, using recent progress in tactile exploration [5], [6] and biomimetic active touch [7]–[9]. The active touch combines a Bayesian evidence accumulation model of perceptual decision making with a control loop for regulating how the tactile sensor contacts a stimulus. The exploration comprises a series of decision episodes, each terminating when the evidence is sufficiently strong to decide edge angle [5], [6], which gives the exploration direction for the next episode. Each decision can begin from prior evidence that biases the forthcoming perception [5] that we take from a passive visual estimate of contour angle and radial position relative to the tactile sensor.

A key aspect of our approach for visuo-tactile sensory fusion is that the prior has a peak at a favoured angle-displacement class and has a precision for how peaked it is around that class. The computer vision approach used here (segmentation using curvature and surface normal) estimates edge angle and radial displacement. In our approach we assume the precision σ is a free parameter that took on the role of weighting vision and touch in the perceptual decision making.

The first main result is that the quality of the exploration is a U-shaped function of the relative contribution of vision and touch, with an optimum at intermediate weighting ($\sigma = 0.3$, from range 0.1-10 shown on Figs 9,10). Uni-modal vision and touch both gave inaccurate edge following, which was improved by combining the two modalities. In general, we expect the weighting between touch and vision will be task specific, depending on the quality of perception from the two modalities. Interestingly, psychophysical experiments reveal that humans fuse tactile (haptic) and visual sensory information in an analogous manner [29].

The second main result is that task performance is more robust when vision is combined with touch, completing the contour when touch alone would fail (Fig. 11; Table II). The tactile edge following becomes more inaccurate with increasing control gain (equivalently exploration step size) until the task fails when the sensor loses contact with the edge. When touch is combined with vision, the edge following is stabilized to not become lost. We attribute this improved robustness partly to improved perceptual performance, but also that vision gives an approximate representation of the entire contour which is lacking from touch alone.

Generally, one would expect the reliability of visual perception would determine the precision σ, depending for example on the light conditions or the object colour and material. We showed empirically that a particular weighting gave optimal exploration, but this leaves open the question of how that weighting should be determined in practise. One possibility is that the variance is directly related to the signal-to-noise ratio of the sensor modality; another is that the precision σ could be learnt from trying to improve the perception during task performance.

An related question is how does our approach relate to human visuo-tactile perception and object exploration? Because the underlying framework of biomimetic active touch is based on Bayesian evidence accumulation, there are parallels with leading models from perceptual neuroscience. Also, object exploration via edge following represents a task of psychophysical importance in humans as a fundamental exploratory procedure for characterizing objects [30]. While it is known that humans fuse simple tactile and visual information in a Bayesian optimal fashion [29], less is known about how the senses combine to explore objects. A robot embodiment of this task represents a putative model for how visuo-tactile exploration is enacted in humans.

Acknowledgements: We thank Ben Ward-Cherrier, Luke Cramphorn, Nick Pestell, Kirsty Aquilina and John Lloyd. The data used can be accessed in the repositories at http://lepora.com/publications.htm and at the University of Bristol data repository, data.bris, at http://doi.org/b92s.

References

[1] P. Allen and R. Bajcsy. *Object recognition using vision and touch.* (Doctoral dissertation, Columbia University), 1985.

[2] P. Allen. Integrating vision and touch for object recognition tasks. *The International Journal of Robotics Research*, 7(6):15–33, 1988.

[3] S. Stansfield. A robotic perceptual system utilizing passive vision and active touch. *The International journal of robotics research*, 7(6):138–161, 1988.

[4] A. Cherubini, Y. Mezouar, D. Navarro-Alarcon, M. Prats, and J. Corrales Ramon. See and Touch: 1st Workshop on multimodal sensor-based robot control for HRI and soft manipulation.

[5] U. Martinez-Hernandez, T. Dodd, M. Evans, T. Prescott, and N. Lepora. Active sensorimotor control for tactile exploration. *Robotics and Autonomous Systems*, 87:15–27, 2017.

[6] N. Lepora, K. Aquilina, and L Cramphorn. Exploratory tactile servoing with active touch. *IEEE Robotics and Automation Letters*, 2017.

[7] N. Lepora. Biomimetic active touch with fingertips and whiskers. *IEEE Transactions on Haptics*, 9(2):170–183, 2016.

[8] N. Lepora, U. Martinez-Hernandez, and T. Prescott. Active touch for robust perception under position uncertainty. In *Proc. IEEE Int. Conf. Robot. Autom. (ICRA)*, pages 3005–3010, 2013.

[9] N. Lepora, U. Martinez-Hernandez, and T. Prescott. Active bayesian perception for simultaneous object localization and identification. In *Robotics: Science and Systems*, 2013.

[10] H. Tanaka, K. Kushihama, N. Ueda, and S. Hirai. A vision-based haptic exploration. In *Proc. IEEE Int. Conf. Robot. Autom. (ICRA)*, pages 3441–3448, 2003.

[11] K. Honda, T. Hasegawa, T. Kiriki, and T. Matsuoka. Real-time pose estimation of an object manipulated by multi-fingered hand using 3D stereo vision and tactile sensing. In *Proc. IEEE/RSJ Int. Conf. Intell. Robots and Syst. (IROS)*, pages 1814–1819, 1998.

[12] J. Bimbo, L. Seneviratne, K. Althoefer, and H. Liu. Combining touch and vision for the estimation of an object's pose during manipulation. In *Proc. IEEE/RSJ Int. Conf. Intell. Robots and Syst. (IROS)*, pages 4021–4026, 2013.

[13] M. Bjorkman, Y. Bekiroglu, V. Hogman, and D. Kragic. Enhancing visual perception of shape through tactile glances. In *Proc. IEEE/RSJ Int. Conf. Intell. Robots and Syst. (IROS)*, pages 3180–3186, 2013.

[14] P. Güler, Y. Bekiroglu, X. Gratal, K. Pauwels, and D. Kragic. What's in the container? classifying object contents from vision and touch. In *Proc. IEEE/RSJ Int. Conf. Intell. Robots and Syst. (IROS)*, pages 3961–3968, 2014.

[15] B. Higy, C. Ciliberto, L. Rosasco, and L. Natale. Combining sensory modalities and exploratory procedures to improve haptic object recognition in robotics. In *Proc. IEEE-RAS Int. Conf. on Humanoid Robots (Humanoids)*, pages 117–124, 2016.

[16] H. Liu, Y. Yu, F. Sun, and J. Gu. Visual-tactile fusion for object recognition. *IEEE Transactions on Automation Science and Engineering*, 2016.

[17] T. Bhattacharjee, A. Shenoi, D. Park, J. Rehg, and C. Kemp. Combining tactile sensing and vision for rapid haptic mapping. In *Proc. IEEE/RSJ Int. Conf. Intell. Robots and Syst. (IROS)*, pages 1200–1207, 2015.

[18] N. Jamali, C. Ciliberto, L. Rosasco, and L. Natale. Active perception: Building objects' models using tactile exploration. In *Proc. IEEE-RAS Int. Conf. on Humanoid Robots (Humanoids)*, pages 179–185, 2016.

[19] S. Luo, W. Mou, K. Althoefer, and H. Liu. Localizing the object contact through matching tactile features with visual map. In *Proc. IEEE Int. Conf. Robot. Autom. (ICRA)*, pages 3903–3908, 2015.

[20] Q. Li, C. Elbrechter, R. Haschke, and H. Ritter. Integrating vision, haptics and proprioception into a feedback controller for in-hand manipulation of unknown objects. In *Proc. IEEE/RSJ Int. Conf. Intell. Robots and Syst. (IROS)*, pages 2466–2471, 2013.

[21] Q. Li, R. Haschke, and H. Ritter. A visuo-tactile control framework for manipulation and exploration of unknown objects. In *Proc. IEEE-RAS Int. Conf. on Humanoid Robots (Humanoids)*, pages 610–615, 2015.

[22] J. Son, R. Howe, J. Wang, and G. Hager. Preliminary results on grasping with vision and touch. In *Proc. IEEE/RSJ Int. Conf. Intell. Robots and Syst. (IROS)*, pages 1068–1075, 1996.

[23] P. Allen, A. Miller, P. Oh, and B. Leibowitz. Using tactile and visual sensing with a robotic hand. In *Proc. IEEE Int. Conf. Robot. Autom. (ICRA)*, volume 1, pages 676–681, 1997.

[24] C. Chorley, C. Melhuish, T. Pipe, and J. Rossiter. Development of a tactile sensor based on biologically inspired edge encoding. In *International Conference Advanced Robotics (ICAR)*, pages 1–6, 2009.

[25] K Klasing, D Althoff, D. Wollherr, and M. Buss. Comparison of surface normal estimation methods for range sensing applications. In *Proc. IEEE Int. Conf. Robot. Autom. (ICRA)*, pages 3206–3211, 2009.

[26] K. Jordan and P. Mordohai. A quantitative evaluation of surface normal estimation in point clouds. In *Proc. IEEE/RSJ Int. Conf. Intell. Robots and Syst. (IROS)*, pages 4220–4226, 2014.

[27] A. Trevor, S. Gedikli, R. Rusu, and H. Christensen. Efficient organized point cloud segmentation with connected components. *Semantic Perception Mapping and Exploration (SPME)*, 2013.

[28] E. Castillo, J. Liang, and H. Zhao. Point cloud segmentation and denoising via constrained nonlinear least squares normal estimates. In *Innovations for Shape Analysis*, pages 283–299. Springer, 2013.

[29] M. Ernst and M. Banks. Humans integrate visual and haptic information in a statistically optimal fashion. *Nature*, 415(6870):429–433, 2002.

[30] S. Lederman and R. Klatzky. Hand movements: A window into haptic object recognition. *Cognitive psychology*, 19(3):342–368, 1987.