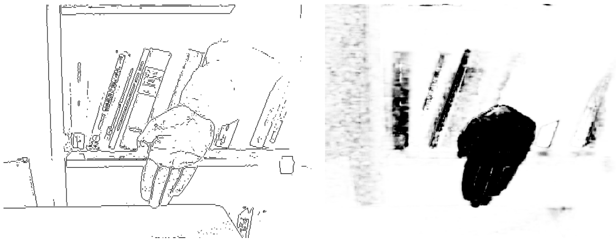

Fig. 4. Image evidence used for tracking. (a) Intensity edges detected by a thresholded gradient operator. (b) Likelihood ratios at each pixel for a color–based skin detector.

corresponding random walks. These dynamics can be justified as the maximum entropy model given observations of the nodes' marginal variances Λ_i.

III. Observation Model

Our hand tracking system is based on a set of efficiently computed edge and color cues. For notational simplicity, we focus on a single video frame for the remainder of this section.

A. Edge Matching Using the Chamfer Distance

As a hand is moved in front of a camera, it obscures the background scene and thus tends to produce intensity edges along the boundaries of its projection in the image plane (see Fig. 4(a)). This edge cue is used by virtually all model–based hand tracking systems [11, 14, 19, 20, 24, 26]. Following [20], we use the Chamfer distance to measure discrepancies between projected model edges and image edges detected by a simple gradient operator. To improve accuracy, we measure distance in terms of both edge position and orientation.

Let $\Pi(x)$ denote the set of edges in the projection of three–dimensional model configuration x, and $\Delta(y)$ the output of an edge detector on the image y. The Chamfer distance $d_E(\Pi(x), \Delta(y))$ is then given by

$$d_E^2(\Pi(x), \Delta(y)) = \sum_{u \in \Pi(x)} \left[\min_{v \in \Delta(y)} g^2(u, v) \right] \tag{5}$$

Here, $g(u, v)$ determines the metric by which errors in edge matches are measured. Letting $u = (u_p, u_\theta)$ denote the position u_p and orientation u_θ of edge u, we define

$$g^2(u, v) = \min \left(\frac{||u_p - v_p||^2}{\sigma^2} + d_\pi^2(u_\theta, v_\theta), g_0 \right) \tag{6}$$

where $d_\pi(u_\theta, v_\theta)$ measures absolute differences in orientation modulo π, and g_0 adds robustness to edge detection failures. Finally, we associate this distance with a likelihood function as follows:

$$p_E(y|x) \propto \exp \left\{ -\lambda_E d_E^2(\Pi(x), \Delta(y)) \right\} \tag{7}$$

For a discussion of the generative model underlying this likelihood function, see [23].

B. Silhouette Matching Using Skin Color Statistics

Skin colored pixels are well known to have predictable statistics [9], and thus provide a powerful cue for hand tracking. We model the color distribution p_{skin} of skin pixels by a single Gaussian in RGB space, with mean and covariance estimated from hand–selected training patches. We assume that non–skin pixels have a uniform color distribution p_{bkgd}.

Let $\Omega(x)$ denote the set of pixels in the silhouette of projected hand model configuration x, and Υ the set of all image pixels. Assuming each pixel is independent, the likelihood of an image y is

$$\begin{aligned} p_C(y|x) &= \prod_{u \in \Omega(x)} p_{\text{skin}}(u) \prod_{v \in \Upsilon \setminus \Omega(x)} p_{\text{bkgd}}(v) \\ &\propto \prod_{u \in \Omega(x)} \frac{p_{\text{skin}}(u)}{p_{\text{bkgd}}(u)} \end{aligned} \tag{8}$$

The second equation follows by neglecting the proportionality constant $\prod_{v \in \Upsilon} p_{\text{bkgd}}(v)$, which is independent of x [4]. Note that we must only evaluate the likelihood ratio over the silhouette region $\Omega(x)$. Figure 4(b) plots these likelihood ratios for a sample hand image.

C. Local Decomposition of Likelihoods

Suppose that the hand model is in a three–dimensional configuration for which there is no self–occlusion. In this case, each hand component will project to a disjoint subset of the image pixels, and the Chamfer distance (eq. (5)) decomposes as

$$d_E^2(\Pi(x), \Delta(y)) = \sum_{i=1}^{16} d_E^2(\Pi(x_i), \Delta(y)) \tag{9}$$

This in turn implies that the edge–based likelihood (eq. (7)) factorizes into a product of terms which provide independent, local evidence for each component:

$$p_E(y|x) \propto \prod_{i=1}^{16} p_E(y|x_i) \tag{10}$$

Similarly, the skin color likelihood (eq. (8)) decomposes as

$$p_C(y|x) \propto \prod_{i=1}^{16} p_C(y|x_i) \tag{11}$$

Note that this statistical decomposition does *not* hold for the original joint angle representation, and is heavily dependent on our choice of a state representation in which the relationship between model parameters and image coordinates is local.

In cases where there is self–occlusion, the local decomposition of eq. (10, 11) will not hold. Nevertheless, we believe that this decomposition will often provide a good approximation. In particular, because occlusion reasoning can only reduce the number of projected model edges, the local decomposition of eq. (10) will always provide an upper bound on the true edge likelihood $p_E(y|x)$.