Fusion of Stereo Vision, Force-Torque, and Joint Sensors for Estimation of In-Hand Object Location

Paul Hebert, Nicolas Hudson, Jeremy Ma, and Joel Burdick

Abstract—This paper develops a method to fuse stereo vision, force-torque sensor, and joint angle encoder measurements to estimate and track the location of a grasped object within the hand. We pose the problem as a hybrid systems estimation problem, where the continuous states are the object 6D pose, finger contact location, wrist-to-camera transform and the discrete states are the finger contact modes with the object. This paper develops the key measurement equations that govern the fusion process. Experiments with a Barrett Hand, Bumblebee 2 stereo camera, and an ATI omega force-torque sensor validate and demonstrate the method.

I. INTRODUCTION

Manipulation of everyday objects is an important topic for mobile and humanoid robotics. In fine object manipulation tasks, such as part mating or component insertion, accurate localization of the grasped object within the hand may be critical for successful task completion. Moreover, the ability to recognize, locate, and track objects in the hand is a key ingredient of autonomous manipulation.

This paper develops and demonstrates a method to fuse measurements from a stereo vision system, a six-axis force-torque sensor, and the encoders of both hand and arm joints to localize a grasped object within the robotic hand. These sensors would naturally be included in many typical robotic manipulation systems, and therefore such a fusion algorithm has practical applications. Moreover, the fusion of these sensors cannot only provide for more precise object localization, but also provides redundancy when sensors fail, or aids the vision sensor when occlusions occur. To solve this problem, we pose the object localization problem as a hybrid systems estimation problem, which seeks to estimate not only continuous variables such as the grasped object’s pose, but also determine the various possible object-finger contact modes. The continuous estimation component is handled within a standard extended Kalman filter framework, while we use a Bayesian static multiple model (SMM) estimation framework [1] for estimating the discrete contact mode. Our main innovation consists of choosing informative measurement features and developing an effective set of filtering measurement equations.

The paper is organized as follows. After reviewing related work in the next section, Section III describes our overall fusion approach. Section IV presents experimental results on the use of the fusion algorithm.

II. RELATED WORK

The ability to detect, localize, and track an object to be grasped or manipulated has been an important research thrust for some time. Earlier work focused on one primary sensor, namely vision [2]–[7]. Recently, a few works have considered the fusion of sensors which might naturally comprise a robotic grasping system. Most papers that discuss utilizing multiple sensors often use them at different stages of a grasp task. Allen et al [8] used vision and tactile sensing to estimate the finger contact position and applied forces. In further work, Allen et al. [9] extends the sensor suite to include a force/torque sensor, however they do not develop a framework to combine the sensors synchronously. Prats [10] combines vision, force/torque and tactile sensors in a control scheme for manipulation tasks such as sliding a door open. Their work did not seek to accurately localize the object. Control is done via virtual visual servoing method (VVS) in which they track an edge of the door. Fusion is not done in a traditional sense; it is initially done with vision and tactile sensor only and the results are then used as an input in an impedance force controller. Schmid et al. [11] also incorporates a multi-sensor control framework towards opening a door, however, the vision sensor is only used to detect the door handle.

In object localization work, most have not focused on using multiple sensors. Petrovskaya [12] uses only tactile sensors to localize an object using a novel particle filter approach. Similarly, Gadeyne [13] uses only a force controlled robot for object localization using Markov localization techniques. Corcoran [14] expanded the above work by proposing a model to incorporate hand-object tracking and does not incorporate any other sensors to the suite.

A hybrid system formulation is a natural approach in grasping and manipulation tasks. A large portion of the literature involves hybrid control in such tasks [15]–[19]. However, only a smaller portion concerns hybrid estimation. Gadeyne [20] relates most to our contribution in which they incorporate a hybrid probabilistic framework for estimating various contact formations of a grasped object in compliant motion tasks. Our work differs in that we apply a multiple model approach, works well with an extended Kalman filter. We also add stereo vision and joint angle measurements. While we implement a static estimator that is sufficient for our experiments, a more sophisticated hybrid estimator such as a generalized pseudo-Bayesian estimator (GPB)
or an interacting multiple model (IMM) can be used for manipulation and tracking tasks.

III. APPROACH

We assume that the robotic system is endowed with the following:

- A Stereo camera whose transform with respect to the robot's base frame is known - either a rigid transform or a variable transform resulting from a pan-tilt unit.
- A 6-axis force-torque sensor located at the wrist.
- A multi-fingered robotic hand which is mounted on the wrist.
- Joint encoders which measure the state of the joints which articulate robot arm and fingers.

A. Grasping and Estimation Model

We assume that the geometry of the grasped object, \( O \), is a priori known, although our formulation allows the geometry to be learnt adaptively, but such learning is not considered in this paper. While the primary goal of our algorithm is to estimate the grasped object’s pose, estimates of additional states, such as the finger tip contact location on the object, are also produced. These extra states improve upon the localization performance, and also allow for redundancy should some sensors fail or some measurements be corrupted by occlusions. In addition, we model a set of discrete states which correspond to specific contact modes, which describe the possible combinatorial arrangement of fingers and object surfaces. More details of the modeling schemes and the definition of the states are given below. Figure 1 provides a view of key reference frames and variables.

Object Geometric Model. We model the object \( O \) as a rigid body consisting of fixed number of faces/surfaces, \( \{S_s\} (s = 1, \ldots, n_S) \), edges, \( \{E_e\} (e = 1, \ldots, n_E) \), and vertices \( \{V_v\} (v = 1, \ldots, n_v) \). While our experiments use a polyhedral object with flat faces, we only require that each face is a smooth parametrizable surface. Similarly, while the edges of our experimental object are straight lines, our algorithm only requires the edges joining two faces to be smooth curves. Each face is assigned a unique index \( s \in [1, \ldots, n_S] \). A local set of coordinates, \( s^\beta \), is established on each face in order to describe potential finger contact points on that face. A polygonal mesh \( (\mathcal{M}) \) is created using a vertex-face description. Each face is parametrized using two orthogonal parameters for rectangular meshes, or parametrized using barycentric coordinates in the case of triangular meshes. A body object reference frame, \( O \), is chosen at the center of mass. The center of mass location may be unknown, which can be added as a dual Kalman filter parameter estimation problem.

Visual Object Model and Features. For purposes of object detection, localization, and tracking, the rigid body model is augmented by a set of visual features that can be observed by the stereo vision system. Visual features from Ma [21] are implemented, which incorporates SIFT features [22] augmented by their 3D location in the camera frame. Each feature takes the form:

\[
D_i = \{d_i \tilde{d}_i\}, \tag{1}
\]

where \( d_i = [d_{ix} \ d_{iy} \ d_{iz}] \) is the 3-D location of the feature and \( \tilde{d}_i \in \mathbb{R}^{128} \) is the 128-dimensional SIFT feature descriptor. This descriptor captures color, texture, and orientation of the object’s appearance around the feature point. The object model and features are learnt during a training phase where SIFT features of the object are selected from multiple camera viewpoints. The 3D position of these SIFT features is obtained using sparse stereo. A database is built of these 3D SIFT features:

\[
D = \{\{d_0 \tilde{d}_0\} \{d_1 \tilde{d}_1\} \ldots \{d_N \tilde{d}_N\}\}. \tag{2}
\]

As the object is learnt visually, an additional object frame is needed which can be viewed from the various learning viewpoints as opposed to an internal body object frame. Hence, an arbitrary camera object reference frame, \( V \) is chosen. The training phase may be carried out using a fixed camera and a turntable for rotating the object. In the case, where the center of mass is unknown, the transformation between the two object reference frames \( V \) and \( O \) may be added into a dual Kalman filter.

Contact modes. We assume a prehensile grasping model where each finger of the robotic hand contacts the object surface at a single point (on the finger tip) on a unique object face. We allow more than one finger to contact a single face. If the hand has \( n_F \) fingers, then there are \( n_S \cdot n_F \) possible pairings of fingers and object faces. Each pairing is termed a contact mode, where \( C_{n_f} \) denotes the contact mode which pairs the \( f^{th} \) finger with the \( s^{th} \) object face.
does not contact any object surface, it is assigned to the null contact mode, \( C_{0,F} \). The contact modes define a discrete set of binary states (e.g., \( C_{s,f} = 0 \) if the \( f^{th} \) finger does not contact the \( s^{th} \) face, and \( C_{s,f} = 1 \) if the finger does contact the face) which must be estimated along with the continuous model states. Correct estimation of these contact modes constrains and improves the object’s pose estimate.

**Linkage joint measurements.** Measurements of the state of the arm and hand linkages (e.g., rotations or displacements of joints in the hand or robot arm) are used, via the use of forward kinematic equations, to produce pseudo-measurements of fingertip locations or wrist locations. Because the camera may be mounted either rigidly or on a pan-tilt unit with respect to the robot’s base frame, \( B \), the transformation from the camera frame to the wrist frame, \( \hat{w}C = \hat{w}BT_w \in SE(3) \), is a function of the arm joint displacements and possibly the pan-tilt inputs. Using the forward kinematic equations, the arm joint displacement measurements in effect provide a noisy measurement of \( \hat{w}C \), where the noise models both the joint sensor errors as well as errors in the arm’s kinematic calibration.

**System State and Estimation Framework.** The continuous state to be estimated at time \( k \) consists of:

\[
X_k = \{ \hat{c}x_p, \hat{c}x_o, \{ \hat{c}\beta \}_{s=1}^{n_S}, c_T \},
\]

where:
- \( \hat{c}x_p \) and \( \hat{c}x_o \) are the position and orientation of the object in the camera reference frame, respectively
- \( \hat{c}T \) denotes the translation and rotation parameters that form the camera to wrist transform.
- \( \hat{c}\beta \) are the parametrized position of the finger contacts on the object face, \( S_s \).

The discrete state to be estimated consists of the contact modes, \( \{ C_{s,F} \} \) \( (s = 0, \ldots, n_S; f = 1, \ldots, n_F) \).

To estimate the continuous states, we use a standard extended Kalman filter (EKF) framework, which requires the specification of both dynamic and measurement models. The discrete state estimates are managed through a static multiple model (SMM) framework, which is described below.

**B. EKF Dynamic Model**

In the extended Kalman filter framework, any discrete time dynamic model of the form

\[
X_{k+1} = \mathcal{F}(X_k, u_k) + \eta
\]

can be used which models the arm’s kinematics, controller inputs \( u_k \) for the manipulator and possibly pan-tilt unit commands. For simplicity, our experiments assume that the dynamic update model describes a random walk model to account for the manipulator’s dynamics:

\[
X_{k+1} = X_k + \eta,
\]

where \( \eta \) is zero-mean Gaussian noise to account for noise in the system from errors in forward kinematics and joint measurements.

**C. EKF Measurement Models**

We now develop physical models of each sensing process to create measurement models that relate each sensor to the state. The incorporation of all these models within the EKF provides the desired fusion.

**Force-Torque Measurement.** The wrist-mounted six-axis force-torque sensor will be sensitive to the gravitational forces acting on the object and the multi-fingered hand (although, up to the first moment mass of the object, the 3 torque measurements are the only measurements which provide information about the object location). The wrench measured at the wrist, \( W_w \), has the following relationship with the gravitational load on the object and hand mechanism:

\[
W_w = W_w^0 + W_w^H
\]
\[
= [u^T \hat{o} W_w, \hat{\beta} R_w \hat{W}_g] + W_w^H
\]
\[
= [\hat{w} R_w \hat{W}_g, \hat{w} \hat{\rho}_o] \hat{W}_w^H,
\]

where \( \hat{W}_g \) is the gravitational force acting on the object, as measured in the world reference frame, \( \hat{w}R_w \) denotes the orientation of the world frame with respect to the wrist frame, \( \hat{w} \hat{\rho}_o \) is the displacement of the object frame origin with respect to the wrist frame origin, and \( \hat{w} \hat{\rho}_o \) is the \( 3 \times 3 \) skew symmetric matrix such that \( \hat{w} \hat{\rho}_o \hat{\beta} = \hat{w} \hat{\rho}_o \times \hat{\beta} \). \( W_w^H \) denotes the gravity load of the hand and a look-up table of these resting values is computed for various orientations of the hand and fingers.

Note that the torque measured at the wrist sensor, \( T_w \), is the only portion of the force-torque measurement that is dependent on the object pose. Therefore, the measurement equation for the force-torque sensor:

\[
T_w = \hat{w} \hat{p}_o \hat{w} R_w \hat{W}_g + \hat{T}_w^H + \xi_T
\]
\[
\equiv H_T(\hat{c}x_p, c_T) + \hat{T}_w^H + \xi_T,
\]

where \( \hat{T}_w^H \) represents bias torque due to the hand mass, \( \xi_T \) represents measurement noise.

**Hand Joint Measurements.** Measurements from the joints in finger linkages, \( (\theta_f) \), may be used to provide a pseudo-measurement of the finger tip locations by means of forward kinematics.

To include the interaction between the fingers and the object, we include the finger surface location into the measurement model. We assume a simple contact sense which can determine if the \( f^{th} \) finger is in contact. When the finger is in contact we use a closure equation for the \( f^{th} \) finger in the wrist frame, \( w \):

\[
w \hat{p}_f(w) = w \hat{p}_o + w \hat{R}_o \hat{p}_f(M) + \xi_f
\]
\[
\equiv H_f(\hat{c}x_p, \hat{c}x_o, \hat{c}\beta, \hat{c}w^T + \xi_f,
\]

where \( \hat{p}_f(M) \) is the surface location of the finger \( f \) in the object’s reference frame and may be found using a surface parametrization. For example, using a barycentric coordinates as the surface parameters, \( \hat{\beta} = \lambda_{1,2,3} \), and the
three vertices \( v_{1,2,3} \) that are defined by the contact mode \( C_{s,f} \):

\[
\phi_f(M) = \lambda_1 v_1 + \lambda_2 v_2 + \lambda_3 v_3.
\] (9)

**Stereo Vision Measurements.** The vision measurements are the 3D SIFT features \( Z_k = \{ (z_0, \hat{d}_0), \ldots, (z_N, \hat{d}_N) \} \) where \( z_i = [x_i, y_i, z_i] \). Lowe’s [23] Best-Bin-First scheme is used to match the observed SIFT features with those in the database providing the correspondences \( J \in \mathbb{Z}^+ \) so that \( J(i) = j \) links the database features \( d_j \) with observation \( z_i \).

Therefore, the measurement model consists of matching the position components of both the database and observed 3D SIFT features. Assuming at time \( t_k \) there is \( m_k \) matches from the Best-Bin-First algorithm:

\[
\begin{bmatrix}
z_0 \\
z_1 \\
\vdots \\
z_{m_k}
\end{bmatrix} =
\begin{bmatrix}
\tilde{x}_p + c R_0(\bar{x}_o) \mathbf{d}_{J(0)} \\
\tilde{x}_p + c R_0(\bar{x}_o) \mathbf{d}_{J(1)} \\
\vdots \\
\tilde{x}_p + c R_0(\bar{x}_o) \mathbf{d}_{J(m_k)}
\end{bmatrix} + \xi_v
\] (10)

where \( \xi_v \) is a white Gaussian noise associated with the stereo measurements. Similar to Ma [21], improving the computational effort of the vision measurement process is done via limiting the region of interest (ROI) by using the maximum and minimum image coordinates of the matched 3D SIFT features of the previous measurement update.

**Joint Measurement Vector and Sensor Fusion.** Our sensor fusion method is simply based on the use of a joint measurement vector which incorporates all of the measurements outlined above. To formulate the EKF measurement equations, we concatenate all the measurements equations [7, 8, 10] to produce:

\[
H =
\begin{bmatrix}
H_T \\
H_{f=1} \\
\vdots \\
H_{f=F} \\
H_{v}
\end{bmatrix}.
\] (11)

As a natural byproduct of the EKF equations, each of the measurements will contribute to the overall state estimate according to its uncertainty and its sensitivity to specific state variables.

**D. Hybrid State Estimator**

Since we do not assume how the fingers are contacting the object, our filter must be able to estimate which surface is in contact with each finger (i.e., which contact modes are active), in addition to the continuous state variables. Thus, we are faced with a hybrid state estimation problem wherein both the continuous states and discrete contact modes must be estimated. While a variety of hybrid system estimation algorithms and architectures have been proposed, for our experiments we use a static multiple model estimator (SMM) [1]. Each model has an associated measurement model to describe how each finger might contact specific object faces.

The SMM is a bayesian framework, in which it leverages prior probabilities of each model being correct to update and obtain posterior probabilities based on sensor measurements. The SMM assumes that the system will obey only one of the possible models, such that there are no switching during the estimation process (i.e. the fingers are assumed to not change contact modes during the estimation interval). Although the particular mode that is in effect is assumed to stay fixed (static), each model can have its own dynamics making the estimator essentially dynamic.

While the SMM assumes only one model to be in effect during the estimation process, the particular active mode is initially unknown. The prior probability on each of the possible contact modes can be selected by the user and for example, the prior might be provided by a grasp planner. In our experiments we choose a uniform prior:

\[
P(M_m | Z^0) = \mu_m(0) = \mu_0,
\] (12)

where \( Z^0 \) is the available prior information (e.g. from grasp planners) and \( \sum_{m=1}^{N} \mu_m(0) = 1 \). Using Bayes’ formula, the posterior probability of model \( m \) being correct given measurement data upto time \( k \) is given by:

\[
\mu_m(k) = \frac{\mu_m(k-1) \Lambda_m(k) \mu_m(k-1)}{\sum_{i=1}^{N} \Lambda_i(k) \mu_i(k-1)},
\] (13)

where \( \Lambda_m(k) \) is the likelihood function of mode \( m \). Assuming our system uncertainty is locally Gaussian, the likelihood may be computed as:

\[
\Lambda_m(k) \equiv p(\nu_m(k)) = \mathcal{N}(\nu_m(k); 0, S_m(k)),
\] (14)

where \( \nu_m(k) \) is the residual (innovation) caused by the discrepancy between the measured values and the predicted measurement values from mode \( m \) and \( S_m(k) \) is the \( m \)th mode measurement covariance.

The SMM is a modular filter where a separate filter is maintained for each discrete mode. For more complex objects, the complexity of the SMM may be reduced by
considering only local surfaces near each finger. Figure 2 illustrates how the previous state estimate and covariance may be used to compute the next estimate and covariance given multiple modes and their filters. Each mode will run its own filter to produce a mode-conditioned state estimate, $\hat{X}_m^m$, and associated covariance, $P_m^m$, and the mode likelihood, $\Lambda_m$.

Under the linear-Gaussian system assumption, the pdf of the state is a Gaussian mixture:

$$p(X|Z^k) = \sum_{m=1}^{N} \mu_m(k) N(\chi(k); \hat{X}_m^m(k|k), P_m^m(k,k)) .$$  \tag{15}$$

The optimal MMSE of (15) provides the overall state and covariance:

$$\hat{X}(k|k) = \sum_{m=1}^{N} \mu_m(k) \hat{X}_m^m(k|k)$$  \tag{16}$$

$$P(k|k) = \sum_{m=1}^{N} \mu_m(k) \{P_m^m(k|k) + [\hat{X}_m^m(k|k) - \chi(k|k)](\hat{X}_m^m(k|k) - \chi(k|k))^T\} .$$  \tag{17}$$

The overall hybrid estimation process updates each model state estimate after each measurement, as well as the discrete mode estimator.

IV. EXPERIMENTAL RESULTS

To verify and demonstrate the fusion algorithms described above, we implemented them on the following system. Our grasper is a 3-fingered Barrett Hand BH-8 with internal strain gauges in each finger. These sensors allow us to infer if a finger contact is active. Joint encoders are located at the proximal joint. An ATI omega 6-axis force-torque sensor is mounted at the manipulator wrist. Our setup is a static experiment in which the wrist is fixed rigidly to a frame though we can vary the frame to simulate the motion of an arm. A Point-Grey Research Bumblebee 2 is mounted above the working area and is also fixed rigidly. During our experiment, we use an aluminum rectangular block (Length: 15.24cm Width: 5.08cm Height: 5.08cm) so that the center of mass is known accurately. The block is covered with stickers to provide texture for the vision system. The block is placed at an appropriate height to be grasped by the Barrett Hand. At the outset, the vision system is already providing a vision estimate of the object pose. Once grasped, the combined filter begins to incorporate measurements from the fingers and the force-torque sensor. The initial conditions are chosen to be at the camera sensor origin and uncertainty to be identity.

Figure 4 illustrates a typical grasping scenario. The lower right hand corner is a rectified image of the stereo pair and the bounding box is displayed around the object. In the upper left corner depicts the block with learned SIFT features (yellow dots), current detected SIFT features (red dots), and estimated finger locations on the object (large blue circles). Figure 5 shows the state estimate of the object pose in a static grasp. The position ground truth was measured with respect to a known location in the camera’s frame. The mean error in the position estimate was 5.2mm. The orientation ground truth was measured by visually aligning the axis of the block to a known orientation in the hand. The mean error in orientation was $1.51^\circ$ assuming the ZYX angles were $(90, 180, 0)$.

A second experiment was performed to test the basic tracking capability of the system. The block was grasping in a similar position for a few moments, and then it was rotated manually in the grasp by $90^\circ$ about the Z axis. The block is then rotated back to the original position. Figure 6 shows the state estimate of the object position, demonstrating this rotation inside the grasp. The mean error after a few seconds when the estimator steadies is 6-7mm. The mean orientation error is again about $1.58^\circ$. A third and similar grasp is performed except that the block is rotated in increments of $45^\circ$ up until $180^\circ$. Figure 7 shows the orientation estimate with a mean error of $2.21^\circ$. 
the pose of the object within the grasp and also estimate the contact modes associated with fingers contacting sides of the object. We have shown that the fusion of these sensors provide for reasonable object localization and, theoretically, some robustness to occlusion and sensor failure. If a failure were to occur in vision, the position will still be estimated, however the orientation will be limited to solutions provided by the finger measurements as the force-torque sensor cannot detect changes in orientation. Our studies have shown that with our vision method alone the mean error in position estimate is on the order of 1cm, which is roughly twice the mean error using our method. The vision system alone is also prone to mismatched visual features which can provide large sudden errors in object location. These outlier errors are largely eliminated from the fusion of the other stable signals.

We plan to apply this framework to a more dynamic...
situation with the hand and wrist now attached to a manipulator. In addition, we plan to implement a more sophisticated hybrid estimator to attempt to deal with manipulation of objects and changing object grasp configurations. Our next step is to add tactile sensors to the fingers to provide an additional set of measurements which are apt to provide better object localization.

VI. ACKNOWLEDGMENTS

The author gratefully acknowledges the support from the National Science and Engineering Research Council of Canada (NSERC).

REFERENCES