USING STEREO VISION AND TACTILE SENSOR FEATURES
For Grasp Planning Control

Madjid Boudaba¹, Nicolas Gorges², Heinz Woern² and Alicia Casals³
¹TES Electronic Solution GmbH, Zettalachring 8, 70567 Stuttgart, Germany
²Institute of Process Control and Robotics, University of Karlsruhe
Engle-Bunte-Ring 8-Gebäude 40.28, 76131 Karlsruhe, Germany
³GRINS: Research Group on Intelligent Robots and Systems
Technical University of Catalonia, Pau Gargallo 5, 08028 Barcelona, Spain

Keywords: Stereo vision, Tactile sensors, Grasp planning, Features matching.

Abstract: Planning the grasp positions either from vision or tactile sensors one can expect various uncertainties. This paper describes a scheme that match visual stereo and tactile data based on stereo vision and tactile sensors. For grasp planning, initially, the grasping positions are generated from stereo features, then the feedback of tactile features is used to match these positions. The result of the matching algorithm is used to control the grasping positions. The grasping process proposed is experimented with an anthropomorphic robotic system.

1 INTRODUCTION

In recent years, considerable research in robotic grasping systems has been published. The proposed system works by using the principle of sensing-planning-action. To place our approach in perspective, we review existing methods for sensor based planning for grasping. The existing literature can be broadly classified into three categories; vision based, tactile based and both vision-tactile based. For all categories, the extracted image features are key factors, they can range from geometric primitives such as edges, lines, vertices and circles to optical flow estimates. The first category uses visual image features to estimate the grasping points and from them define the robot’s motion with respect to the object position and orientation before performing a grasp (Yoshimi and Allen, 1994), (Maekawa et al., 1995), (Smith and Papanikolopoulos, 1996), (Sanz et al., 1998), (Kragic et al., 2001), and (Morales et al., 2002). The second category uses tactile image features to estimate the characteristics of the area in contact with the object (Berger and Khosla, 1991), (Chen et al., 1995), (Perrin et al., 2000), and (Lee and Nicholls, 2000). The last category uses data fusion from both vision and tactile sensors in order to control grasping tasks efficiently (Namiki and Ishikawa, 1999), and (Allen et al., 1999).

This paper is an extension of our previous work (Boudaba and Casals, 2006) and (Boudaba and Casals, 2007) on grasp planning using visual features. In this work, we demonstrate the utility of matching both visual and tactile image features in the context of grasping, or fingers position controlling. In our approach, we avoid using any object model, and instead, we work directly from image features to plan the grasping points. In order to avoid finger positioning errors, matching, by back projecting these tactile features into visual features is required to compute the similarity transformation that relates the grasping region with the sensitive touching area. To achieve a high level of grasping position matching efficiency, two matching schemes are considered in this paper. The first establishes grasp points correspondences between the left and right images of the stereo head. In this scheme, only the grasp positions are back-projected into one side of the stereo image. A second scheme is a region matching where the whole sensitive touching area with the object is used in the back projection into the visual image. All the points belonging to the sensitive area of a tactile sensor are back-projected into the grasp region of visual features. The processing in each match is completely independent and can be performed at its own rate. Our approach based on features matching can play the...
critical role of forcing the fingertips to move towards
the desired positions before the grasping is executed.

2 GRASPING SYSTEM
DESCRIPTION

In robotic grasping tasks, when data from several
sensors are available simultaneously, it is generally
necessary to precisely analyze all of them along the
entire grasping process (see Figure 1). The object
being extracted from a video sequence requires
encoding its contour individually in a layered manner
and provide at the receiver’s side some enhanced
visual information. In the same way, from the data
being extracted from a tactile sensor, the tactile layer
processes and provides the tactile information at
its receiver’s side. The architecture of the whole
grasping system is organized into several modules,
which are embedded in a distributed MCA2 (Mod-
ular Controller Architecture Version 2) software
framework (Scholl et al., 2001). There are mainly
two modules involved in this development; the
stereo vision, tactile sensors, and grasp planning.
In MCA2, every module is structured in a data
vector that allows the module to receive and send
the data from/to other modules, or to take any part
of an output data and permute and copy it to other
modules. In order to control its current functionality,
every module has fully or partially access to the input
data of the other modules depending on the tasks
involved. For instance, the grasp planning module
has full access to the input data of the sensory system
and has partially access to the robot hand. Because
the architecture of the system has a global planning
to access to all the data available to the system, the
grasp planning module can locally plan the grasping
positions without having a global view of the robot’s
environment. For instance, the robot hand needs
some information supplied by the global planning
module such as grasp configurations for the object to
be grasped.

2.1 Visual Layer: Feature Extraction

We consider visual features extraction in the context
of a stereo head (see Figure 2). First, however,
we recall the epipolar geometry technique which is
motivated by considering the search of corresponding
points in the stereo image pair. Since we are dealing
with a stereo head, we need to extract features well
suited for determining the grasp points on the first
image either from the left or right side of a stereo
pair and then computing their correspondences in
the second image. Given the estimated object pose,
placed on the table, the full observability of the
object is then projected into the left and right image
planes. The visual layer takes these images and
calibration data as input (see Figure 1) and provides
as output a set of visual features. Segmentation is
used to separate the object from the background
and other objects in the environment by determining
the coordinates of a closed rectangular bounding
box. After segmenting the region corresponding to
the object, features belonging to the object contour
are extracted. A function is then constructed for
parameters regrouping object features together.

Figure 1: Grasping system description.

Figure 2: Stereo camera system and a Pan-tilt unit.

We denote by $V$ a function regrouping visual pa-
rameters that is defined by

$$V = \{ glist, gparam, com \} \quad (1)$$

where $glist$, $gparam$ and $com$ are the visual features.
During image processing, $V$ is maintained as a dou-
bly linked list of grasping region and intervening their
parameters as $g_{18_{param1}}, \ldots, g_{8_{paramm}}$. A grasping region $g_i$ is defined by its ending points $g_{vi}$ and $g_{vi+1}$, and its orientation $\phi$ with respect to the object’s center of mass $c_{om}$. The resulting parameters of $V$ fully describe the two-dimensional location of features with respect to the image plane. The visual features obtained can be used as input data for both, grasp planning and grasp position matching. For more details about this topic, we refer the reader to our previous work (Boudaba and Casals, 2007).

2.2 Tactile Layer: Feature Extraction

Unlike vision which provides global features of the object, tactile sensor provides local features when the fingertip is in contact with the object. The tactile layer shown in Figure 1 takes as input the data extracted from a set of tactile sensor (or so called Group Of Tactile sensor (GOT)) and the configuration of the robot hand and provides as output a set of tactile features. To simplify the problem, tactile features are treated as visual features using the basic results from different approaches. For the purpose of sensor features matching, extracting edge features are of interest and will be discussed in section 4. Figure 3 illustrates the PCB tactile sensor module with its memory and data control units. The sensor module (from Weiss Robotics, (K.Weiss and Woern, 2005)) consists of a sensitive area organized in matrix of $4 \times 7$ sensor cells with a spatial resolution of 3.8 mm. By using four modules, (two in each gripper finger), the parallel gripper shown in Figure 3 is equipped with a total number of 112 sensor cells.

![Figure 3: Gripper equipped with tactile sensor modules.](image)

The data of the tactile sensor matrix corresponds to a two-dimensional planar image. We analyze this image using moments up to the 2nd order (Hu, 1962). The two-dimensional $(p+q)^{th}$ order moment $m_{pq}$ of an image is defined as the following double sum over all image pixels $(x,y)$ and their values $f(x,y)$:

$$m_{pq} = \sum_{x} \sum_{y} f(x,y)x^py^p, \quad p,q \geq 0 \quad (2)$$

The moment $m_{0,0}$ constitutes the resulting force exerted on the sensor. The center of gravity $cog = (x_c, y_c)^T$ of this force can be computed as follows:

$$x_c = \frac{m_{10}}{m_{00}}, \quad y_c = \frac{m_{01}}{m_{00}} \quad (3)$$

The center of gravity of each tactile sensor matrix determines a contact point of the gripper.

3 GRASP POSITION MATCHING

The Grasping system can be explained in more detail through a set of tasks. In order to complete the grasp matching process, it is necessary to find the relationship between their Cartesian coordinate frames (see Figure 4).

![Figure 4: Sensor frames relationship for grasping.](image)

We define these frames as follows:

- **HCS.** Head Coordinate System has a stand alone configuration of stereo head. The fixation of the head is assured by controlling the pan-tilt angle. The offset to the object coordinate system (OCS) is constant.
GCS. Gripper Coordinate System (also known as the end effector frame) is determined by controlling the angular configuration of the robot arm. The robot arm moves over itself and the measurements given by all joints enable the system to determine the Tool-Center-Point (TCP) relative to the robot coordinate system.

OCS. Object Coordinate System is fixed on the table and does not change its position and orientation during calibration. Once the object pose relative to the stereo head (HCS) is determined, the GCS relative to the object pose is determined by using robot hand - stereo head calibration.

TCS. Tactile Coordinate System. The location of TCS is terminated by controlling the GCS configuration.

WCS. World Coordinate System. The location of the object position and orientation OCS, HCS and the robot base station are determined relative to the WCS.

In the remaining of this work, the kinematics of the robot arm and robot hand are ignored. So far, for grasping using features matching, we have established two mapping relationships between feature frames. The first mapping implies finding the grasp points correspondence between left and right image of the stereo head \((C_{lt}t_oC_r)\). The second mapping implies matching the two apparent features into the tactile and stereo frames \((Tt_oH)\).

### 3.1 Stereo Images Matching: \(C_{lt}t_oC_r\)

We adopt some notations similar, but not identical to the work of (Hartley and Zisserman, 2000) on multiple view geometry (see Figure 4). Considering a point \(P_o\) in contact with the object, its distance to the center of mass, \(d_o\) is measured and its projection into the stereo head and tactile frames is given by \((p_l, p_r)\) and \(n\), respectively. The subscripts, \(o\), \(c\), \(l\), \(r\), and \(t\) are referred to object, contact point, left and right frames, and tactile frame, respectively.

Let \(p_r = (u_r, v_r, 1)^T\) and \(p_l = (u_l, v_l, 1)^T\) denote the projection of \(P_o\) on the right and left images, respectively. The epipolar plane defined by the three points \(P_o\), \(C_l\) and \(C_r\) intersects the two image planes in two epipolar lines \(e_p\) and \(e_p\). The line connecting the two centers of projection \(C_l\) and \(C_r\) intersects the image planes at the conjugate points \(e_r\) and \(e_l\) which are called epipoles. Using the projective coordinates, the epipolar constraints can be written:

\[
    p_r^T F p_l = 0 \quad (4)
\]

where \(F\) is the so-called fundamental matrix which consists of a 3x3 unknown entries and can be expressed as follows:

\[
    F = \begin{bmatrix}
    f_{11} & f_{12} & f_{13} \\
    f_{21} & f_{22} & f_{23} \\
    f_{31} & f_{32} & f_{33}
    \end{bmatrix} \quad (5)
\]

In the calibrated environment, the 9 unknown entries of \(F\) can be captured in an algebraic representation as defined by

\[
    F = C_{lt}^{-T}EC_{lt}^{-1} \quad (6)
\]

where the fundamental matrix \(F\) encapsulates both the intrinsic and the extrinsic parameters of the stereo head, while the essential matrix \(E = [T]_r, R\) which compactly encodes the extrinsic parameters of the stereo head can be composed of the baseline vector \(t = [C_r - C_l] = (t_x, 0, t_z)^T\) and the angular rotation \(\beta\) about the y-axis that renders the left image parallel to the right one, then we have:

\[
    E = \begin{bmatrix}
    0 & -t_z & 0 \\
    t_z & 0 & -t_x \\
    0 & t_x & 0
    \end{bmatrix} \begin{bmatrix}
    \cos(\beta) & 0 & -\sin(\beta) \\
    0 & 1 & 0 \\
    \sin(\beta) & 0 & \cos(\beta)
    \end{bmatrix} \quad (7)
\]

\(C_l\) and \(C_r\) are the intrinsic parameter matrices of the left and right cameras defined by

\[
    C_l = \begin{bmatrix}
    f_{u_l} & 0 & C_{ul} \\
    0 & f_{v_l} & C_{vl} \\
    0 & 0 & 1
    \end{bmatrix}, C_r = \begin{bmatrix}
    f_{u_r} & 0 & C_{ur} \\
    0 & f_{v_r} & C_{vr} \\
    0 & 0 & 1
    \end{bmatrix} \quad (8)
\]

where \(u_{wl}\) and \(v_{wl}\) (resp. \(u_{wr}\) and \(v_{wr}\)) are the coordinates of the principle point (in pixels) of the left (resp. right) camera. \((f_x, f_y)\) are the focal length in x and y direction.

For more details about camera calibration and related topics, we refer to the work of (Faugeras and Toscani, 1986) and (Tsai and Lenz, 1989).

### 3.2 Tactile and Stereo Matching: \(Tt_oH\)

By dealing with the contact constraint, the minimum distance between a fingertip (tactile) and the object can be expressed by a parameter \(d_t\). So keeping a fingertip in touch with the object, the condition \(d_t = 0\) must be maintained, and tactile features are extracted and measured into the tactile frames. Matching these tactile features with visual features implies the computation of a similarity transformation relating the
grasping region to actual sensitive touching area. Tactile and visual features are then related by the following transformation

\[ s_i = T_{ro}Hv_i \]  

(9)

where \( s_i \) and \( v_i \) are points on the tactile and visual image features, respectively. \( T_{ro}H \) is the similarity transformation given by

\[
T_{ro}H = \begin{bmatrix}
    s \cos \alpha & -s \sin \alpha & t_x \\
    s \sin \alpha & s \cos \alpha & t_y \\
    0 & 0 & 1
\end{bmatrix}
\]  

(10)

where \( s \), \( \alpha \), and \([t_x, t_y, 1]\) are scaling, rotation angle and translation vector of the tactile image with respect to the visual image, respectively.

In the calibration cases, the parameters of (10) can be computed directly using homogeneous transformation matrices between frames as shown in Figure 4.

4 IMPLEMENTATION

The implementation of our algorithm for grasping position matching using stereo vision and tactile sensor can be divided into two parts. First is related to the grasp planning using stereo vision. The second part of this implementation is related features matching between stereo vision and tactile sensor.

4.1 Grasp Planning using a Stereo Head

As stated before, the first implementation consists of computing the grasp points correspondence in the stereo vision. More formally, let \( G = \{G_{v1}, G_{v2}, \ldots, G_{vk}\} \) be a set of valid grasps and \( G_{vi} = (g_{vi}, \ldots, g_{vi}, 1)^T \) be its \( i \)-th determined grasp point on the left image, next step is to compute its correspondence on the right image, \( G'_{vi} = (g'_{vi}, g'_{vi}, 1)^T \).

To do this, we first need to establish a mapping relationship between a line and point by exploiting the epipolar constraint defined by (4).

Let

\[
L_{vi} = FG_{vi}^l, \quad L_i = F^T G'_{vi}
\]  

(11)

be the mapping equations where \( F^T \) is the transpose of \( F \) and \( i := 1, 2, \ldots, k \) is the number of grasping points. \( G'_{vi} \) (resp. \( G''_{vi} \)) is the determined grasp point on the left (resp. right) image and \( L_{vi} \) (resp. \( L_i \)) is its corresponding epipolar line on the right (resp. left) image. By exploiting the epipolar constraint (4), the grasping points are constrained to lie along the epipolar lines \( L_{vi} \) and \( L_i \), respectively.

If both grasping points satisfy the relation \( G'_{vi}^l F G'_{vi} = 0 \) then the lines defined by these points are coplanar. This is a necessary condition for the grasp points to correspond.

Given the parameters of a line and a grasp point in one image, the maximum deviation of a point from the line can be computed as follows:

\[
d_i^2 = \text{norm}(L_{vi}, G_{vi}^l), \quad d_i^2 = \text{norm}(L_i, G'_{vi})
\]  

(12)

where \( d_i^2 \) (resp. \( d_i^2 \)) is the maximum deviation of a grasp point on the left (resp. right) image.

We can estimate a cost function with respect to a parameter \( t \) as follows:

\[
C(t) = d_i^2 + d_i^2
\]  

(13)

The minimum threshold \( t_{min} \) corresponds to the \( t \) where the cost function is minimum.

4.2 Features Matching

The second implementation consists of computing the similarity between the stereo and the tactile images features. To compare image features, the Hausdorff metric based on static features matching is used (Huttenlocher et al., 1993).

Given two feature sets: \( S = \{s_1, s_2, \ldots, s_q\} \) and \( V = \{v_1, v_2, \ldots, v_q\} \), the Hausdorff distance from the point set \( S \) to point set \( V \) is defined as

\[
h(S, V) = \max_{s \in S} \min_{v \in V} ||s - v||
\]  

(14)

where \( ||s - v|| \) corresponds to the sum of the pixel difference and indices \( i \) and \( j \) correspond to the size of a searching window.

The matching process is evaluated according to the output of the function (14). The matching that results in the lowest cost is the one that matches the closest grasp planning. Since we want to guide the gripper toward the grasping points previously generated by the grasp planning, the solution consists of reducing the cost function (or so called grasp error) by
USING STEREO VISION AND TACTILE SENSOR FEATURES - For Grasp Planning Control

Table 1: Parameters measure of grasping positions and cost function: obj1, obj2.

<table>
<thead>
<tr>
<th>Object</th>
<th>Left grasp</th>
<th>Right grasp</th>
<th>Threshold: $t = 2$ corresponds to the maximum deviation.</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>$(x_l, y_l)$</td>
<td>$(x_r, y_r)$</td>
<td>$d_l$, $d_r$, $C$</td>
</tr>
<tr>
<td>obj1</td>
<td>358.500 365.500</td>
<td>313.500 377.000</td>
<td>2.934, 2.940, 17.252</td>
</tr>
<tr>
<td></td>
<td>359.000 400.000</td>
<td>312.500 411.500</td>
<td>5.008, 5.018, 50.261</td>
</tr>
<tr>
<td>obj2</td>
<td>355.000 226.000</td>
<td>299.500 237.500</td>
<td>1.726, 1.731, 17.252</td>
</tr>
<tr>
<td></td>
<td>363.000 314.500</td>
<td>309.000 323.500</td>
<td>0.922, 0.926, 1.7028</td>
</tr>
<tr>
<td></td>
<td>309.500 316.500</td>
<td>255.000 321.500</td>
<td>5.251, 5.259, 55.234</td>
</tr>
</tbody>
</table>

Figure 5: Result of two and three-fingered grasp planning algorithms using stereo images.

4.3 Matching Algorithm

- **Input:** images: $im_1$, $im_2$, features: $V_1$, $V_2$. Number of fingers: $k$. Fundamental matrix: $F$. Threshold $t$. Size of window: 7x7 pixel.

- **Output:** Grasping points: $(G_{lv_i}^l, G_{rv_i}^r)$, with $i := 1, ..., k$. Matching: $h(S, V)$

- **Process:**
  1. Perform the features extraction tasks

  - for $i := 1$ to $2$
  - extract features: $V_i := im_i$.

  2. Perform the grasp planning tasks

  - select $V_1$ on which the grasp will be performed.
  - Get valid grasps point $G_v$ from (1)

  3. Perform the grasping point correspondences

  - select $V_2$ on which the grasping correspondences will be performed

  - for $i := 1$ to $k$
  - Compute $G_{lv_i}^l := G_{v_i}$
  - Compute $L_v$ and $G_{rv_i}^r$ using (11)

  4. Perform the matching function

  - for $i, j := 1$ to $7, 7$
  - Compute $h(S, V_i) \leq r$ and $h(S, V_j) \geq r$
  - Compute pixel difference $h(S, V_i)$ from (14)

moving the tactile sensors toward these points. The cost of a solution is expressed as the total sum of contact displacements over the surface of the object from an initial contact configuration. If the result of matching is outside a given margin, then the grasp controller should launch a new measurement via joint angle and position sensors.
5. Perform the cost function
   - Compute $d^2_l$ and $d^2_r$ using (12)
   - Compute $C(t)$ with (13)

6. End.

4.4 Experimental Results

The algorithm was implemented on our experimental system, which consists of a 7 DOF manipulator arm, a robot hand with two fingers, each one equipped with a tactile sensor module mounted directly to the finger tip, and a vision system (see Figure 6). This first prototype of anthropomorphic robot system developed by the German Research Foundation (see Boudaba et al., 2005) is used as platform and demonstrator for a coming generation of service robots. The grasping configuration is based on a stand-alone stereo head (MEGA-D from Videre Design) mounted on a pan-tilt controller unit equipped with a pair of 4.8 mm lenses and a fixed baseline of about 9cm. We have experimented our approach with two different kind of objects placed on a fixed table with a fixed position and orientation (static object). Figure 5 illustrates the results obtained from our matching algorithm using stereo vision. The performance of our results (see Table 1) is validated according to a cost function $C$ defined in the stereo images as the errors between grasping points. The cost that results in the lower value is the one that matches the closest grasp planning. Figures 6(b)-(d) illustrate the feedback of tactile sensor giving the top/bottom position of fingers with respect to the tactile image plane in (b) while (c)-(d) showing the top/bottom sensitive area in touch with the object.

5 CONCLUSIONS

The implementation of our algorithms for grasping points matching using stereo vision and tactile sensor have been detailed. Two schemes for grasping points matching have been included in this work. In the first scheme, stereo vision matching was used to find the grasp points correspondence between left and right images. It is shown that the quality of matching
depends strongly on the precise computation of the intrinsic and extrinsic parameters of the stereo head calibration. The performance of our results is evaluated according to a cost function defined in the stereo images as the errors between a pair of grasping points. In the second scheme, the tactile sensor provides the sensitive area of a fingertip in contact with an object which was used with the grasp region to compute the similarity between both features. Using these two matching schemes, we were able to fuse the visual grasp region with the tactile features and capabilities of reducing or avoiding the grasp positioning errors (or so called controlling the grasp planning) before executing a grasps.

REFERENCES


