Micropositioning of a Weakly Calibrated Microassembly System Using Coarse-to-Fine Visual Servoing Strategies

Stephen J. Ralis, Barmeshwar Vikramaditya, and Bradley J. Nelson

Abstract—This paper presents a novel visual servoing framework for micropositioning in three dimensions for assembly and packaging of hybrid microelectromechanical systems (MEMS). The framework incorporates a supervisory logic-based controller that selects feedback from multiple visual sensors in order to execute a microassembly task. The introduction of a visual sensor array allows the motion of microassembly tasks to be controlled globally with a wide angle view at the beginning of the task. Then a high precision view is used for fine motion control at the end of the task. In addition, a depth-from-focus technique is used to visually serve along the optical axis, providing the ability to perform full three-dimensional (3-D) micropositioning under visual control. The supervisory logic-based controller selects the relevant sensor and tracking strategy to be used at a particular stage in the assembly process, allowing the system to take full advantage of the individual sensor’s attributes such as field-of-view, resolution, and depth-of-field. The combination of robust visual tracking and depth estimation within a supervisory control architecture is used to perform high-speed, automatic microinsertions in three dimensions. Experimental results are presented for a micro insertion task performed under this framework in order to demonstrate the feasibility of the approach in high precision assembly of MEMS. Results demonstrate that a relative parts placement repeatable to 2 µm in X, Y and 10 µm in Z is possible without the use of costly packaging equipment and thermal management systems.

Index Terms—Depth-from-focus, MEMS, microassembly, sensor integration, supervisory logic-based control, visual servoing.

I. INTRODUCTION

As microelectromechanical systems (MEMS) devices become more functional and more complex, the need for assembling hybrid MEMS devices, such as miniature drug pumps, actuators [2], [8], sensors [3], [11], optical devices [6], etc. becomes apparent. Packaging these devices in order to protect them from their operating environment while allowing for interfaces to the necessary electrical, mechanical, and fluidic elements is also important. The eventual commercial success of hybrid MEMS technology requires that the handling of these microparts be performed automatically in order to preserve potential economic benefits. In this paper, an approach that achieves micron-level relative parts placement with a weakly calibrated microassembly system is described. A key aspect of the system is the use of continuous vision feedback from an array of visual sensors with different spatial resolutions for controlling the task in three dimensions.

In a macro domain, visually servoed assembly [5], [17] has been shown to effectively compensate for uncertainty in the calibration of camera-lens systems, manipulators, and workspaces. However, manufacturing engineers usually prefer the cost of strongly calibrated parts handling systems to the complexity of vision systems due to issues of cost and reliability. In a micro domain, though, precise calibration is highly dependent on precisely modeled kinematics which are subject to thermal growth errors. Two common techniques for compensating for thermal errors include either the use of expensive cooling systems, or waiting hours for the thermal equilibrium of the device to stabilize. Slocum [24] points out that “thermal growth errors are typically the most difficult to control and compensate.” Vibration must also be compensated. Combined with various other sources of positioning error, the tolerance stack-up problem becomes daunting for microassembly where parts must often be placed relative to one another with micron and even submicron tolerances. Because these types of factors greatly affect the cost and reliability of precision assembly machines, real-time visual feedback can be used effectively and economically as a component of a microassembly system.

This paper describes a supervisory logic-based control architecture that integrates multiple visual sensors. The architecture allows for switching sensing modes based on the task to be performed for optimal speed and repeatability. Experimental results demonstrate the feasibility of this architecture in performing automated micro insertion under visual control. The approach provides highly repeatable relative parts placement without the need for expensive vibration isolation systems and thermal management techniques.

II. MICROASSEMBLY

A. Mechanics of Microassembly

Microassembly tasks differ from their macro counterparts because of the vastly different physics that predominate in the micro domain and that have yet to be completely characterized. Consider assembly in the macro world. The mechanics of manipulation in this domain are predictable and can be modeled accurately. For example, when a gripper opens, forces due to gravity cause the part to drop. This predictability has enabled the
success of many complex sensorless manipulation strategies. In the micro world, forces other than gravity tend to dominate due to scaling effects. For example, a common microassembly scenario is one in which electrostatic forces cause a part to “jump” into a gripper before contact actually occurs. As the gripper opens to place the micropart at its goal, the part may stick to the gripper fingers and not remain at the desired location [4]. If humidity in the room happens to be high, surface tension effects can dominate gravitational forces, and the part would also remain stuck to the gripper. It is estimated that for parts with major dimensions below 100 μm, gravitational forces become less dominant than surface effect forces due to electrostatics, Van der Waals, and surface tension [1]. However, this is only a rough estimate and depends on several factors, such as mass density, surface roughness, humidity, part geometry, electrical grounding, etc. Although particular forces can be defined, their effect on the process can only be roughly estimated. The complex microphysics that must be compensated, combined with the high precision relative parts positioning required by microassembly tasks, provides a challenging packaging issue for hybrid MEMS.

B. Related Work in Microassembly

Currently, microdevices requiring complex manipulation are assembled by hand using an optical microscopes and probes or small tweezers, and is essentially a form of teleoperated micro-manipulation. For example, the authors have assembled many different microdevices by hand using optical microscopes, including miniature fiber optic assemblies, micropumps, and electron columns for miniature scanning electron microscopes. The ultimate goal of this research is to develop robust manipulation strategies for automating these types of assembly tasks.

Many researchers are actively pursuing strategies for manipulating micron sized objects for various applications. For example, researchers have used feedback from a scanning electron microscope (SEM) to teleoperatively guide micromanipulation [22]; systems have been developed for accurate positioning of optical elements [29]; techniques for remote teleoperation of micro/milli sized structures have been developed [10]; vision based methods have been proposed [12], [25], [27]; and microassembly workcells are being built [14], to name a few of the efforts in this area.

III. VISUAL SERVOING SYSTEM MODEL AND CONTROL

A. Overview of Visual Servoing Systems

Visually servoed assembly has been shown to effectively compensate for uncertainty in the calibration of camera-lens systems, manipulators, and workspaces, though all research in this area has concentrated on the macro domain. The first report of an experimental visual servoing system appeared in 1973 [23], though the field was first really defined in 1984 [28]. As the speed of computer processing power has risen while the cost has fallen, high speed (30 Hz) visual servoing frameworks have recently become realizable. A typical visual servoing control loop is shown in Fig. 1. Many differences exist between the various approaches to visual servoing and include A—the space in which reference inputs are provided; B—the dimensionality of the control space and the structure of the controller; C the physical configuration of the system; and D the feature tracking algorithms used by the vision system. Our approach to visual servoing is an image-based one in which controller errors are defined in image coordinates. The advantage of image-based visual servoing is that it eliminates the need to perform an explicit inverse perspective projection mapping. This simplifies the observer dynamics and is generally much easier to implement, and most researchers today prefer an image-based approach.

B. Camera Model

In formulating the visual servoing component of our system, task space coordinates are mapped into sensor space coordinates through a Jacobian mapping. A Jacobian for a camera-lens system of the form

$$\mathbf{J}_v = \mathbf{J}_v(\phi)\mathbf{X}_T$$

(1)

is desired, where $\mathbf{J}_v$ is a velocity vector in sensor space; $\mathbf{J}_v(\phi)$ is the image Jacobian matrix and is a function of the extrinsic and intrinsic parameters of the vision sensor as well as the number of features tracked and their locations on the image plane; and $\mathbf{X}_T$ is a velocity vector in task space. For an eye-in-hand camera mounted on a microscope and allowed to translate and rotate, $\mathbf{J}_v(\phi)$ can be simplified to be the form

$$\mathbf{J}_v = \begin{bmatrix} -\frac{m}{s_x} & 0 & 0 & 0 & -\frac{Zc}{s_x} & \frac{y_0 s_y}{s_y} \\ -\frac{m}{s_y} & 0 & \frac{Zc}{s_y} & 0 & \frac{x_0 s_x}{s_x} & \frac{x_0 s_x}{s_y} \end{bmatrix}$$

(2)

where $m$ is the magnification of the microscope system, $s_x$ and $s_y$ are pixel dimension on the CCD; $x_0$ and $y_0$ are the actual image coordinates of a feature; and $Zc$ is the depth of the feature with respect to the image sensor. A complete derivation of this model can be found in [27] (see Fig. 2).
Generally several features are tracked. Thus, for η feature points the Jacobian is of the form

\[ \mathbf{J}_v = \left[ \mathbf{J}^T_1(t) \cdots \mathbf{J}^T_\eta(t) \right]^T \]  

(3)

where \( \mathbf{J}_i(t) \) is the Jacobian matrix for each feature given by (2).

### C. Optimal Controller

The state equation for the visual servoing system is created by discretizing (1) and rewriting the discretized equation as

\[ \mathbf{x}(k+1) = \mathbf{x}(k) + T \mathbf{J}_v(k) \mathbf{u}(k) \]  

(4)

where \( \mathbf{x}(k) \in \mathbb{R}^{2M} \) (\( M \) is the number of features being tracked); \( T \) is the sampling period of the vision system; and \( \mathbf{u}(k) = [\dot{X}_T \; \dot{Y}_T \; \dot{Z}_T \; \omega_{X_T} \; \omega_{Y_T} \; \omega_{Z_T}]^T \) is the task manipulator's end-effector velocity. The Jacobian is written as \( \mathbf{J}_v(k) \) in order to emphasize its time varying nature due to the changing feature coordinates on the image plane. Because the zoom and focus of the optical system are fixed, the intrinsic parameters of the camera-lens system are constant for the experimental results to be presented.

The control objective of the system is to control end-effector motion in order to place the image plane coordinates of features on the target at some desired position. The desired image plane coordinates could be constant or changing with time. The control strategy used to achieve the control objective is based on the minimization of an objective function that places a cost on errors in feature positions, \( \mathbf{E}(k+1) = (\mathbf{x}(k+1) - \mathbf{x}_D(k+1))^T \mathbf{Q}(\mathbf{x}(k+1) - \mathbf{x}_D(k+1)) \), and a cost on providing control energy or input, \( \mathbf{u}(k) \)  

\[ \mathbf{E}(k+1) = (\mathbf{x}(k+1) - \mathbf{x}_D(k+1))^T \mathbf{Q}(\mathbf{x}(k+1) - \mathbf{x}_D(k+1)) + \mathbf{u}^T(k) \mathbf{L} \mathbf{u}(k). \]  

(5)

This expression is minimized with respect to the current control input \( \mathbf{u}(k) \). The end result yields the following expression for the control input

\[ \mathbf{u}(k) = -\left( T \mathbf{J}_v^T(k) \mathbf{Q}\mathbf{I} \mathbf{J}_v(k) + \mathbf{L} \right)^{-1} T \mathbf{J}_v^T(k) \mathbf{Q}(\mathbf{x}(k) - \mathbf{x}_D(k+1)). \]  

(6)

### D. Supervisory Logic-Based Controller

Sensory feedback from multiple sensors provides rich information about the task space. Multiple vision sensors that have different operating regimes in terms of resolution and field-of-view are used. A logic-based supervisory controller is used to switch between different sources of feedback [7]. The appropriate sensor is determined by the controller and the control system parameters are configured for the active sensor. The architecture is depicted in Fig. 3.

Realizing that at every instant in time only one control input is required, a single controller with adjustable parameters can be formulated in state-space form as

\[ \mathbf{x}(k+1) = \mathbf{x}(k) + T \mathbf{J}_v(k) \mathbf{u}(k) \]

\[ \mathbf{u}(k) = \mathbf{F}_\sigma \left[ (\mathbf{x}(k) - \mathbf{x}_D(k+1)) \right] \]  

(7)

where \( \mathbf{x}(k) \in \mathbb{R}^{2M} \) (\( M \) is the number of features being tracked in both image sensors); \( T \) is the sampling period of the vision system; and \( \mathbf{u}(k) = [\dot{X}_T \; \dot{Y}_T \; \dot{Z}_T]^T \) for micropositioning in \( X, Y \), and \( Z \). The Jacobian for this system is given by (3) where the intrinsic and extrinsic terms in each \( \mathbf{J}_i(t) \) depend on the image sensor used to track feature \( i \). The supervisory signal \( \mathbf{\sigma} \) is transmitted to the controller \( \Sigma \) and selects the controller input to the system based on the visual sensory feedback. The control input \( \mathbf{u} \) changes with respect to the signal output of the supervisor that switches the control gains in \( \mathbf{F}_\sigma \). This provides real time logic-based switching of multiple controllers via feedback of the process from multiple sensors. For a system able to translate in \( X, Y \), and \( Z \) in which a single feature is tracked by each of two image sensors, \( \mathbf{F}_\sigma \) is a \( 3 \times 3 \) matrix of the form (8), shown at the bottom of the page, where \( \mathbf{J}_{1,2} \) are \( 2 \times 3 \) Jacobians for the two image sensors; \( \mathbf{Q}_{1,2} \) (2 \times 2) and \( \mathbf{L}_{1,2} \) (3 \times 3)

\[
\mathbf{F}_\sigma = \begin{bmatrix}
\sigma_1(T \mathbf{J}_1^T(k) \mathbf{Q}_1 T \mathbf{J}_1(k) + \mathbf{L}_1) T \mathbf{J}_1^T(k) \mathbf{Q}_1 & (- \sigma_2(T \mathbf{J}_2^T(k) \mathbf{Q}_2 T \mathbf{J}_2(k) + \mathbf{L}_2) T \mathbf{J}_2^T(k) \mathbf{Q}_2) & 0 \\
0 & 0 & 0 \\
0 & 0 & \sigma_3 \mathbf{G}_z
\end{bmatrix}
\]  

(8)
are controller gain matrices; \( G_z \) is a predetermined constant \( Z \) translation velocity; and \( \sigma_{1,2,3} \) have the value 1 or 0 and select the appropriate control strategy to use at any given moment.

System behavior is determined through a supervisory logic module (see Fig. 3) which specifies \( \sigma_t \) in (8) throughout the micropositioning task. The algorithm used to determine the values for \( \sigma_t \) is based on the feature error vector \( \mathbf{e}(k) - \mathbf{p}(k+1) \) and whether the object to be manipulated is within the depth-of-field of the microscope. The logic is as follows.

1) Initial coarse visual servoing in \( XY \) until the feature error in the global view is eliminated, \( \sigma_1 = 1 \) and \( \sigma_2 = \sigma_3 = 0 \).
2) Constant motion along \( Z \) until the probe is observed within the depth-of-field (see Section IV-B) of the high resolution visual sensor, \( \sigma_1 = \sigma_2 = 0 \) and \( \sigma_3 = 1 \).
3) Fine visual servoing in \( XY \) until the probe is positioned over the hole, i.e. the feature error in the high precision view is eliminated, \( \sigma_1 = \sigma_3 = 0 \) and \( \sigma_2 = 1 \).
4) Final insertion along \( Z \) while simultaneously performing fine visual servoing in \( XY \) to correct for system disturbances during the final insertion, \( \sigma_1 = 0 \) and \( \sigma_2 = \sigma_3 = 1 \).

This logic induces the control strategy described by (7) to exhibit the coarse-to-fine visually servoed micropositioning strategy desired, while using a combination of depth-from-focus and blind moves in \( Z \) to perform the full 3-D microinsertion task.

IV. IMAGE PROCESSING

A. Feature Tracking on the Image Plane

The measurement of the motion of the features on the image plane must be done continuously and quickly. The method used to measure this motion is based on an optical flow technique called sum-of-squared differences (SSD). The inherent assumption is that the intensities around a feature remain constant as that feature moves across the image plane. The displacement of a point \( p_0 = (x_s, y_b) \) at the next time increment to \( p_0 = (x_s + \Delta x, y_b + \Delta y) \), is determined by finding the displacement \( \Delta x = (\Delta x, \Delta y) \) that minimizes the SSD measure \( c(p_0, \Delta x) \). A pyramidal search scheme is used to reduce the search space. A more complete description of the algorithm and its implementation can be found in [17].

B. Motion Along the Optical Axis

1) Depth-Of-Field: Micro assembly tasks require highly magnified views of the task space to provide submicron accuracy. High magnification optical systems usually have a high numerical aperture and thus have a very small depth-of-field.
Fig. 6. Micro insertion workstation and close-up view of the multiple visual sensors.

Fig. 7. Low precision visual servosing: top left: before visual servosing; bottom left: after visual servosing; right: acquisition of desired probe location by visual servosing of X and Y axes.

Fig. 4 shows the insertion probe when it is in focus and when it is out of focus, the depth-of-field being approximately 10 μm. This limited depth-of-field can be exploited to measure depth from the camera using techniques of depth-from-focus/defocus.

Depth-from-focus has been studied extensively as a technique for recovering depth estimates from the limited depth-of-field exhibited by optical lenses [9], [16], [20]. The depth of field formulation for a pinhole camera system is given as

$$\Delta = Da f \left( \frac{1}{af - p(D - f)} - \frac{1}{af + p(D - f)} \right)$$  \hspace{1cm} (9)$$

where Δ is the depth-of-field, D is the focus distance, a is the lens aperture, f is the lens focal length, and p is the minimum of the x and y pixel dimensions on the CCD array [26].

The above formulation is valid for optical systems approximated by a pinhole camera model. However, for high numerical

---

### Table I: Calibration Data

<table>
<thead>
<tr>
<th></th>
<th>Global Visual Sensor</th>
<th>High Precision Visual Sensor</th>
</tr>
</thead>
<tbody>
<tr>
<td>X Pixels</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>X Encoder</td>
<td>54.7</td>
<td>4.3</td>
</tr>
<tr>
<td>X Actual Displacement (μm)</td>
<td>17.8</td>
<td>2.2</td>
</tr>
<tr>
<td>Y Pixels</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>Y Encoder</td>
<td>52.1</td>
<td>4.3</td>
</tr>
<tr>
<td>Y Actual Displacement (μm)</td>
<td>26.8</td>
<td>2.2</td>
</tr>
</tbody>
</table>
where $\lambda_0$ is the wavelength of light in a vacuum, $n$ is the diffractive index of the lenses, $A$ is the numerical aperture of the lens system, and $m$ is the magnification of the optical system [13]. The significance of this equation is that the depth-of-field is on the order of the wavelength of light. This provides the ability to calculate depth of objects in the micro domain with a resolution approaching the wavelength of light.

2) Focus Measure: Generally focused images are characterized by high spatial frequency content, while blurred images have attenuated high frequency content. Fig. 5 shows the histograms corresponding to the focused and defocused probe shown in Fig. 4. The histogram corresponding to the focused probe has intensity variations from approximately 100 to 215 (for an 8-bit CCD). Two notable peaks illustrate the focused characteristic of the feature (left peak) as well as the background (right peak). As a result, a histogram can be used to characterize the level of focus for the feature of interest. A histogram of the region of interest is continuously monitored in order to provide a servoing ability along the optical axis before a final insertion operation is carried out.

The pixels around the edge of the focused feature produce a dip between the histogram peaks. A threshold gray level value is automatically chosen in the trough region that characterizes a reasonable boundary for the focused object from the background [15], [21].

V. EXPERIMENTAL RESULTS

A. Hardware Setup

Experiments were conducted with the micro insertion workcell shown in Fig. 6. The workcell consists of a Daedal positioning platform with independent $X$, $Y$, and $Z$ motion powered by Yaskawa E-series servodrives and servomotors. The multiple visual sensory array consists of a Marshall Electronics V-X0071 video camera on a chip and a Sony XC-75 CCD camera using a microscope zoom lens (Marshall Electronics Inc. V48612MZ). From Fig. 6 one can see that the coarse vision sensor, the video camera on a chip, is not orthogonal to the $XY$ plane in order to avoid occlusion from the fine vision sensor. Although the controller (7) assumes that the depth of the features being tracked is known and remains constant for each sensor throughout the micromanipulation task, the closed-loop visual servoing approach used is easily able to compensate for the limited depth variations of this camera configuration.

Image processing and visual servoing control calculations were performed with a vision system consisting of a digitizer and framegrabber based on Texas Instruments TMS320C40 DSP’s, supplied by Traquair Data Systems. The vision system is able to track up to five $16 \times 16$ feature templates at 30 Hz. Motion control was accomplished utilizing a programmable multi-axis controller (PMAC-PC) servo motion card manufactured by Delta Tau. It should be noted that no special vibration isolation equipment was used to achieve the results, and no attempt to address thermal expansion of devices was necessary.
B. Visual Servoing Results

Micropositioning of the system was categorized into three visual servoing steps: low precision visual servoing utilizing the global view for feedback; high precision depth servoing using the histogram information to visually servo along the optical axis; and final high precision visual servoing in the task space followed by the insertion.

Repetitive micropositioning of a probe to numerous holes on a machined template was accomplished. Resolution of the multiple visual sensory array is listed in Table I. The holes are machined to 254 μm in diameter and are separated by 1 mm in both the X and Y directions. The probe is tapered to 228.6 μm in diameter.

Optimal performance was achieved by tuning the values of the diagonal terms in the control gain matrices $Q_1, Q_2, L_1,$ and


Stephen J. Ralis received the B.M.E. degree from the Catholic University of America, Washington, DC, in 1994 and the M.S. degree in mechanical engineering from the University of Illinois, Chicago, in 1998. He is a Vision Systems Engineer with Electroglas, Inc., Corvallis, OR, engineering new vision technologies to better aid surface/bump inspection for the semiconductor industry. His current research interests lie in design/development of image processing algorithms, computer vision, intelligent sensing systems, nano/micro inspection and manipulation, pattern recognition, and defect classification.

Barmeshwar Vikramaditya received the B.E. degree in mechanical engineering from Delhi College of Engineering, Delhi, India, in 1995, the M.S. degree in mechanical engineering from the University of Illinois, Chicago, in 1997, and is currently pursuing the Ph.D. degree in mechanical engineering from the University of Minnesota, Minneapolis.

His research interests include intelligent sensor based control, mechatronics, computer vision, remote teleoperation, and virtual reality.

Bradley J. Nelson received the B.S. degree (Bronze Tablet) in mechanical engineering from the University of Illinois at Urbana-Champaign in 1984, the M.S. degree in mechanical engineering from the University of Minnesota, Minneapolis, in 1987, and the Ph.D. degree in robotics from the School of Computer Science, Carnegie Mellon University, Pittsburgh, PA, in 1995.

He is the Mayhugh Associate Professor of Mechanical Engineering at the University of Minnesota. He holds a McKnight Land-Grant Professorship.

He has been an Assistant Professor at the University of Illinois at Chicago, has worked as an Engineer for Honeywell, Inc. and Motorola, Inc., and has served as a United States Peace Corps Volunteer in Botswana, Africa. His research interests include automatic assembly, biomedical engineering, controls, computer vision, intelligent systems, manufacturing, mechatronics, MEMS, microassembly, and robotics.

Dr. Nelson received the Office of Naval Research Young Investigator Award and the National Science Foundation Faculty Early Career Development (CA- REER) Award.