Robust Visual Servoing based on Novel View Prediction

A.H. Abdul Hafez[1], Piyush Janawadkar[2], and C.V. Jawahar[2]
[1] CSE Department, College of Engineering, Osmania University, Hyderabad-07, India
hafezsyr@ieee.org
[2] Center for Visual Information Technology, IIIT, Hyderabad-19, India
{piyush_j@students., jawahar@}iiit.ac.in

Abstract

In this paper we propose a novel technique for robust visual servoing in presence of a large proportion of outliers in image measurements. The method employs robust statistical techniques and novel view prediction for improving the performance. We identify a set of points from initial and reference images and compute the essential matrix relating them. The selected points are predicted from the initial and reference images to the current frame using essential matrices. A function of the difference between the observed and predicted image point measurements is used to identify outliers. This technique is validated with many experiments and compared with other robust methods in a simulation framework.

1 Introduction

Visual servoing is the process of positioning the end effector of a robot manipulator with respect to a target object or a set of features. This is achieved by processing the visual feedback and minimizing an appropriate error function. The visual feedback can be image features (2D) or object pose (3D) with respect to the camera frame [1].

Based on the visual information, visual servoing algorithms can be classified to three categories [1]: position-based, image-based, or hybrid (or $2\frac{1}{2}D$) visual servoing. In image based visual servoing, $2D$ visual information is extracted from the images and used directly in the control law to generate the control signal *i.e.*, the screw velocity of the robot end-effector. This velocity is computed by minimizing an error function with the help of an image Jacobian or interaction matrix [2]. The accuracy of the computation depends on the performance of feature detection, matching, tracking, and modeling schemes. If the correspondences between features are noisy, the visual servoing process fails to converge, and the system will reach inaccurate final state or a local minima [3]. Traditionally, approaches like increasing the accuracy of the model of the vision system or improving the local processing methods for tracking and detecting features, take care of these errors [4].

In the visual servoing literature, Kragic and Christensen [5] proposed an algorithm to provide a robust input to the control law. Their algorithm used voting and consensus technique to integrate multiple visual cues. It compensates the effect of outliers in the image processing phase. Andrew *et al.* [6, 7] proposed an M-estimator based statistical approach that utilizes redundancy in image features to detect and reject the outliers. This method is integrated in the visual servoing control law. Redundant visual features are used to keep the full rank of the interaction matrix. The inability to reject outliers in presence of excessive noise is a drawback of this method. This is due to the statistical properties of the median operator. Using RANSAC, the problem of presence of excessive noise can robustly be addressed. RANSAC [8] has a discriminate function with a threshold value to classify the points as outliers or inliers.

In this paper, we propose a new method for robust visual servoing using multiple view geometry. Many recent visual servoing algorithms use results from multiple view geometry to improve the performance of visual servoing algorithms [9, 10, 11]. In contrast to these works, we employ the epipolar constraints to predict a novel image and thereby derive an image-based visual servoing control law which is robust to image noise.

Our method uses both epipolar geometry and statistical techniques for robust visual servoing. Image-based visual servoing needs initial and desired images for calculation of the motion parameters. We improve the robustness of this computations with the help of an additional image with known relationship (say essential matrix that relates them) to the initial frame. From the image acquired by the camera and a predicted image, we identify and suppress outliers for im-

proving the robustness. The measured features with large deviation from the predicted ones are classified as outliers. The threshold value used for the decision making is computed using the residuals of the points with respect to the velocity computed using the predicted one.

2 Background and Previous Work

2.1 Image-based Visual Servoing

The problem of image-based visual servoing is that of positioning the end-effector of a robot arm such that a set of image features S reaches a desired target S^*. The set S may be composed of the coordinates of the points that belong to the object. Other kinds of geometric features like straight line segments, or angles can also be used. Consider the error function

$$e(S) = S - S^*, \tag{1}$$

which is the difference between the current feature vector S and the desired one S^*. By differentiating this error function with respect to time, we get

$$\frac{de}{dt} = \frac{dS}{dt} = (\frac{\partial S}{\partial P})\frac{dP}{dt} = L_S V, \tag{2}$$

where S is a $(2N \times 1)$ features vector obtained by stacking the image coordinates (u_i, v_i) of N interest-points. The velocity $V = (v^T, \omega^T)^T$ is the camera velocity, v is translational velocity and ω is rotational velocity. The pose vector $P = (x, y, z, \alpha, \beta, \gamma)$ is a (6×1) vector, where (x, y, z) represent the 3D coordinates of the camera frame position and the three angles (α, β, γ) represent the camera frame direction with respect to a reference frame. The $(2N \times 6)$ matrix L_S is called the interaction matrix or the image Jacobian. It relates the changes in the image space to the changes in the Cartesian space [2].

Assuming a perspective projection model with unit focal length, the interaction matrix L_{S_i} for each point (u_i, v_i) is given by [2]

$$\begin{bmatrix} -\frac{1}{Z_i} & 0 & \frac{u_i}{Z} & u_i v_i & -(1+{u_i}^2) & v_i \\ 0 & -\frac{1}{Z_i} & \frac{v_i}{Z_i} & 1+{v_i}^2 & -u_i v_i & -u_i \end{bmatrix}, \tag{3}$$

where $i = 1, ..., N$, and Z_i is the depth of the point in the camera coordinate frame. The interaction matrix L_S for the complete set of N points is

$$L_S = \begin{bmatrix} L_{S1} \\ \vdots \\ L_{SN} \end{bmatrix}, \tag{4}$$

where L_{S1} and L_{SN} are the interaction matrices given by Equation(3), and correspond to the N points.

The main objective of the visual servoing process is to minimize the error function $e(S)$. For exponential convergence of the minimization process , we need $\frac{de(S)}{dt} = -\lambda e(S)$ given in Equation (1). By substituting this in Equation (2) and using a simple proportional control law, the required velocity of the camera can be shown [2] to be

$$V = -\lambda L_S^+ e(S). \tag{5}$$

The matrix L_S^+ is the pseudo-inverse of the Jacobian matrix L_S, and λ is a scale factor.

2.2 Robust Image-based Visual Servoing

A robust visual servoing control law based on M-estimator was proposed in [7]. They modified the error function as

$$e(S) = D[S - S^*],$$

where $D = diag(w_1, .., w_i, .., w_{2N})$ is a weighting matrix, and N is the number of points. The weight w_i is zero if the point is an outlier and w_i is one if the point is an inlier. The computation of weights w_i is done using Tukey's robust function [12]. For this objective function, the control law is derived [7] as

$$V = -\lambda [D L_S]^+ D[S - S^*]. \tag{6}$$

One can see, in Equation (6), that the matrix D is being introduced to the error function and the interaction matrix. Entries of the interaction matrix that correspond to the outlier features also will be nullified by the multiplication of zeros. This ensures the complete rejection of outliers.

2.3 Novel View Synthesis

A camera is a mapping from the 3D world to a 2D image. A general projective camera is represented by an arbitrary homogeneous (3×4) matrix of rank 3. The general projective camera M maps world point X to image point x according to $x = MX$. The matrix M includes *internal parameters*, *i.e.* the camera's focal length and the skew angle, and the *external parameters* which specify the camera's position and orientation in the world [13].

Epipolar Geometry describes the relationship between corresponding points in two views. Suppose x and x' are the corresponding points in two views, then the epipolar constraint has the form [13]

$$x^T E x' = 0. \tag{7}$$

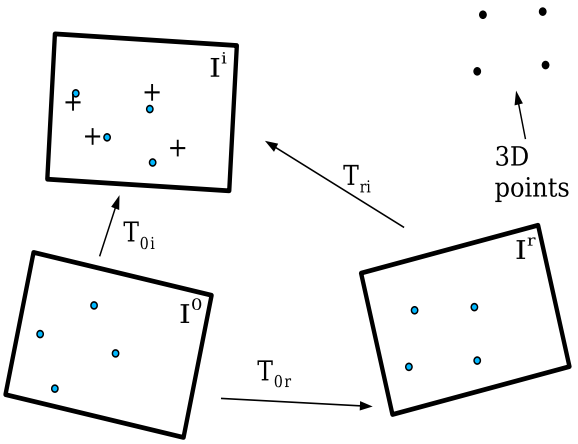


Figure 1: 3D configuration of the proposed method. In the current image I^i, the $\circ$ is the measured point features S^i and the $+$ is the predicted point features $\hat{S}^i$.

The (3×3) matrix E is known as the *essential matrix* and has rank 2. The essential matrix maps a point x' in one view to a line $l = Ex'$ in the other view. This line is called the epipolar line. The essential matrix describes the relative geometry of the two cameras. Suppose the relative transformation between the two cameras is given by $T = \left[\begin{array}{cc} R, & \mathbf{t} \end{array} \right]$ then, it can be shown that

$$E = R^T [\mathbf{t}]_\times, \tag{8}$$

where the matrix $[\mathbf{t}]_\times$ is the antisymmetric matrix associated with vector $\mathbf{t}$. The pairwise epipolar geometry can be used to predict new views [13]. A correspondence between two given images ($x \leftrightarrow x'$) constrains the point in the third image x'' to lie on the lines $E_{31}x$ and $E_{32}x'$. The predicted point in the third view is the intersection of these two epipolar lines, and is given by

$$x'' = E_{31}x \times E_{32}x'. \tag{9}$$

Given a correspondences between two sets of points in two images, a third novel image, which is defined in terms of essential matrices, can be produced by transferring all corresponding pixels from the two given images to this new image using Equation (9). Many algorithms are available for reliable computation of the essential matrix between two images [13]. In our proposed method, we use the Equation (8) to compute the essential matrix.

3 Proposed Method

The robust visual servoing control law presented in Section 2.2 is able to reject a few outliers. Here we propose a solution that works even when the proportion of the noisy points is large. An additional reference image is used to predict a virtual novel image. The transformations between the current image and each of the two initial and reference images are used to predict the novel image. The predicted image is used to detect the outlier points in the current image using a discriminate function. This function uses the residual value of the data points in the current image with respect to the data in the predicted image. During the visual servoing process, a constant value is assigned to the error function which corresponds to the outlier point feature. The error function given in Equation (1) is modified as

$$\hat{e}(S_i) = \begin{cases} S_i - S^* & \text{if } S_i \text{ is inlier.} \\ e_0(S_i) & \text{if } S_i \text{ is outlier.} \end{cases} \tag{10}$$

Here $e_0(S_i)$ is a precomputed constant value of the visual servoing error function. This constant is selected such that it decreases the contribution of the outlier as much as possible, while avoiding the singularity in the control law. By substituting the error function given in Equation (10) in Equation (5), the control law gets modified as

$$\hat{V}_i = -\lambda L_i^+ \hat{e}(S_i), \tag{11}$$

where $\hat{e}(S_i)$ is computed using Equation (10). Note that there are no additional computations when compared to the original image-based visual servoing control law.

The proposed algorithm is divided into an initialization (off-line) and two on-line steps. The first on-line step is the computation of the predicted image, and the second is the identification of outliers using the error function computed from the actual current image and the predicted current image. Figure 1 depicts the geometric configuration of the proposed method. Consider an initial image I^0 of a scene, which consists of a set of 3D points. In addition to the initial image, the camera takes another image (reference image) I^r with a known transformation between these initial and reference camera positions $T_{0r} = [R_{0r}, t_{0r}]$. Select a set of point features S^0 in the initial image I^0 and another corresponding set S^r in the reference image I^r. The novel predicted image at the ith time instance I^i contains the corresponding set $\hat{S}^i$ of these point features.

The novel image computation is done using the velocity measurement of the camera. At each iteration of the visual servoing process, the transformations between the current image and the two (initial and reference) images $T_{0i} = [R_{0i}, t_{0i}]$ and $T_{ri} = [R_{ri}, t_{ri}]$ are

computed. By substituting these two transformations in Equation (8), the essential matrices E_{0i} and E_{ri} are obtained. Using these essential matrices and Equation (9), the current predicted image is computed.

The set of features $\hat{S}^i$ in the current predicted image corresponds to the sets S^0 and S^r. Substituting the features vector $\hat{S}^i$ in the control law given in Equation (5) will give the camera velocity in the ith iteration that is contributed by the features $\hat{S}^i$

$$V = -\lambda L^+_{\hat{S}^i}(\hat{S}^i - S^*). \tag{12}$$

Consider the term $r_{pred} = \lambda(\hat{S}^i - S^*)$ as the state of the current predicted image with respect to the desired one, and the term $r_{actu} = \lambda(S^i - S^*)$ as the state of the actual or measured current image with respect to the desired one. The error required for the discriminant function is defined as $r_i^2 = (r_{pred} - r_{actu})^2$. In literature [7, 13], these are known as the residual values of the points S^i. Using Equations (2), (5) and (12), The residual value for each single feature S_i in the actual current image is given as

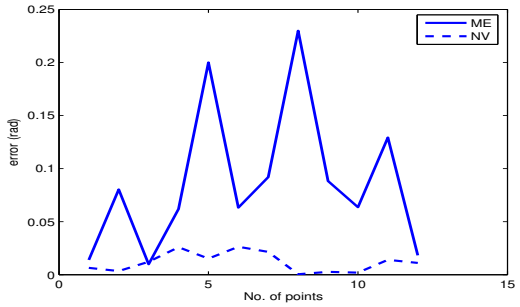
$$r_i^2 = (L_{S_i}V + \lambda(S_i - S_i^*))^2. \tag{13}$$

The feature S_i is considered as an outlier if $r_i \geq t_\sigma$, and is treated as an inlier if $r_i < t_\sigma$, where the threshold value $t_\sigma = \sqrt{5.99}\sigma$ [8]. The term σ is a function of the uncertainty in the velocity measurements of the robot arm in the ith iteration.

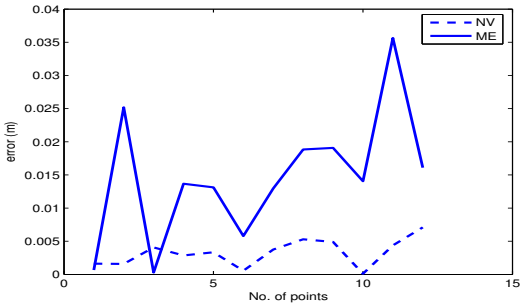
4 Simulation Results

In the simulation experiments we considered a set of 3D points X_i, $i = 1, ..., N$. These points belong to an object in the scene. A positioning task is considered for the study. The robot arm has to move from an initial position to a given desired position. The desired position is specified as a desired image of the object. The image point coordinates are considered as features. Since we have N points, the total number of features is $2N$. In other words, features S_{2i} and S_{2i-1} are from the point x_i. If any of the features S_{2i} or S_{2i-1} is found to be an outlier, other one is also considered as an outlier. The error given in Equation (10) is used for both features.

We conduct simulation experiments to show the behavior of our proposed method. These experiments differ in the number of the points, which are disturbed by noise and the amount of this noise. We conduct our simulation experiments in presence of excessive noise using two robustness methods. The first one is the method described in [7] that uses M-estimator based

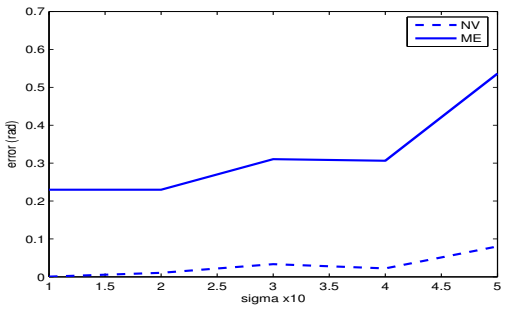


(a)

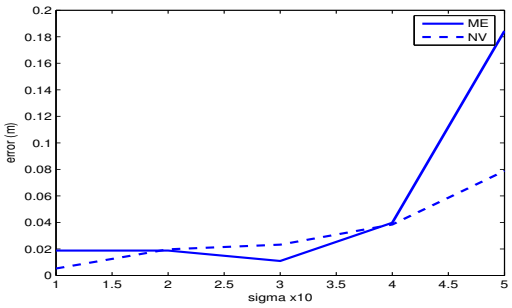


(b)

Figure 2: Rotational error $\|R\|$ in (a), and translational error $\|T\|$ in (b), between final and desired pose versus the number of noisy points.



(a)



(b)

Figure 3: Rotational error $\|R\|$ in (a), and translational error $\|T\|$ in (b), between final and desired pose versus the amount of noise σ added to 8 points out of 12.

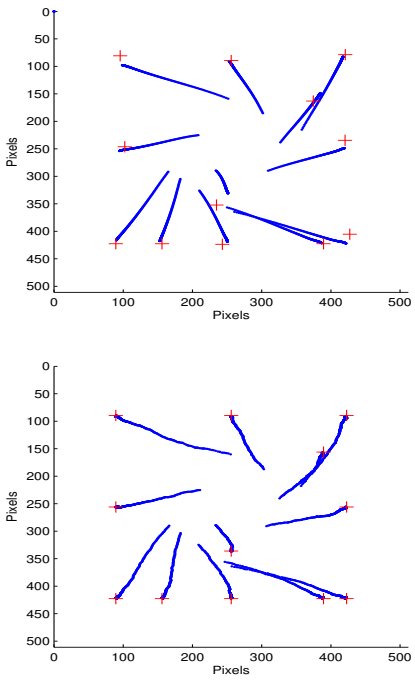
algorithm. We call it here ME method. The second one is the method proposed in this paper. The proposed method is called the novel view based method and will indicate to it as NV method. We introduce the noise in the matching and feature extraction step. We consider the error vector between the final camera pose and the desired one as the convergence error. This error vector is represented by the measurements of the norms of its rotational parts $\|R\|$ and translational one $\|T\|$. The 3D target object consists of $N = 12$ points. To compare the convergence capability of the two methods, we conduct the experiments for the all possible number of noisy points out of the total 12 points.

Figure 2 shows the rotational $\|R\|$ and the translational $\|T\|$ parts of the convergence error versus the number of the noisy points. For a small number of noisy points the performance of the both ME and NV are similar. In contrast, the convergence error in case of the NV method is much less compared to the case of ME method when the number of noisy points is increased.

To prove the results statistically, we repeat the experiments 10 time for a selected number of noisy points. We take the average of the convergence error over the 10 values of the convergence error vector. Table 1 show the average the rotational $\|R\|$ and the translational $\|T\|$ part values of the convergence error versus a selected number of noisy points.

For a fixed number of noisy points (say 8 points out of the total 12), we conduct the experiment for a different amount of noise. Since we use a Gaussian noise, the amount of the noise is represented by the variance value σ^2. We conduct the experiments for the values of σ^2 in the range of $(10, \ldots, 50)$. We can see the notable difference in the rotational part $\|R\|$ of the convergence error between the two ME and NV methods. This is shown in Figure 3. In the same time, there is not much difference between the two methods in the translational part. This can be explained as the error in the image point affects the rotation motion more than the translation one. However, in the both ME and NV methods, the convergence error is increased rapidly after a certain value of noise amount.

To show the properties of the robust visual servoing algorithm using the both robust methods, we consider the case of 8 noisy points and $\sigma^2 = 20$. Figure 4 shows the image trajectory of the point features where the final position is different from the desired one. A local minima of the error function is reached instead of the desired global one. This is depicted in Figure 5. In this figure the norm of the error vector is shown



(a)

(b)

Figure 4: Points trajectory in the image space of image features. The mark + in the image indicates the desired position of the image point.

No. of noisy points	5	8	11
$\|R\|$	0.0054	0.0111	0.0129
$\|T\|$	0.0207	0.0393	0.0312

Table 1: The average of the convergance error over 10 repeated times.

instead of its all components. Figure 6 shows a comparison between the camera trajectories in the Cartesian space in the ideal case where there is no error in the image feature measurements, and each of our proposed method in Figure 6(a) and M-estimator method in Figure 6(b). The difference between the final and desired pose is clear.

We can conclude from the experiments that our method is superior to the previous one regardless to the number of image points disturbed by noise. Results show that it works even in case of all points disturbed by noise. The M-estimator methods are restricted to the case where a little points were disturbed by noise.

5 Conclusion

A novel robust image-based visual servoing method is proposed here. This method classifies the image points to outliers or inliers. The detected outlier is

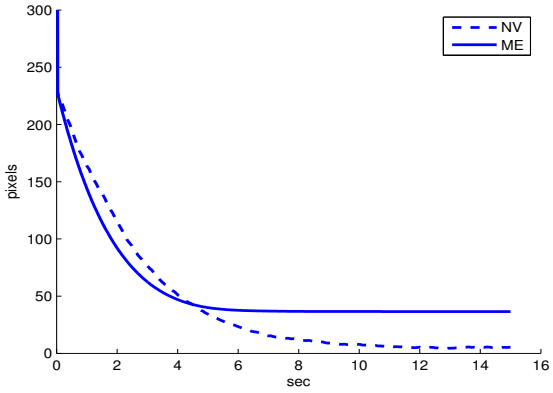


Figure 5: The norm of the error vector of image features versus the time in seconds.

introduced in the control law with a constant error value. The core of this method lies in combining statistical methods with multiple view geometry. As an improvement to the previous work in the robust visual servoing, this method can produce a better convergence with large noisy features proportion. As a future work, this can be extended to visual servoing architectures like 3D and $2\frac{1}{2}D$ visual servoing. Other kind of features may be considered to improve the robustness.

References

[1] E. Malis, F. Chaumette, and S. Boudet, $2\frac{1}{2}D$ visual servoing, *IEEE Trans. on Robotics and Automation* , Vol. 15, pp. 234-246, 1999.

[2] S. Hutchinson, G. Hager, and P. Cork, A Tutorial on visual servo control, *IEEE Trans. on Robotics and Automation* , Vol. 17, pp. 18-27, 1996.

[3] E. Marchand, and F. Chaumette, Feature tracking for visual servoing purposes, *Robotics and Autonomous Systems*, Vol. 52, pp. 53-70, 2005.

[4] P. Li, O. Tahri, and F. Chaumette, A shape tracking algorithm for visual servoing, *IEEE Int. Conf. on Robotics and Automation, ICRA'05* , pp. 2858-2863, Spain, 2005.

[5] D. Kragic, and H. Christensen, Cue integration for visual servoing, *IEEE Trans. on Robotics and Automation* , Vol. 17, pp. 19-26, 2001.

[6] A.I. Comport, M. Pressigout, E. Marchand, F. Chaumette, A Visual Servoing Control Law that is Robust to Image Outliers, *IEEE Int. Conf. on Intelligent Robots and Systems, IROS'03* Vol. 1, pp. 492-497, Nevada, 2003.

[7] E. Marchand, A.I. Comport, and F. Chaumette, Improvements in robust 2D visual servoing, *IEEE Int. Conf. on Robotics and Automation, ICRA'04* , Vol. 1, pp. 745-750, LA, 2004.

[8] P.J. Rousseeuw, A.M. Lero, *Robust Regression and Outlier Detection*, John Wiley and Sons, 1987.

[9] D. Kragic, *Visual Servoing for Manipulation: Robustness and Integration Issues*, Phd thesis, Royal Institute of Technology , 2001.

[10] J. Piazzi, and N.J. Cowan, Multi-view visual seroing using epipoles, *IEEE Int. Conf. on Intelligent Robots and Systems, IROS'04*, Vol. 1, pp. 674-679, 2004.

[11] E. Cervera, F. Berry, and P. Martinet, Is 3D useful in stereo visual servoing?, *IEEE Int. Conf. on Robotics and Automation, ICRA'02*, pp. 1630-1635, USA, 2002.

[12] P. J. Huber, *Robust Statistics*, Wiler. New York, 1981.

[13] Richard Hartley and Andrew Zisserman, *Multiple View Geometry in Computer Vision*, Cambridge University Press, 2003.

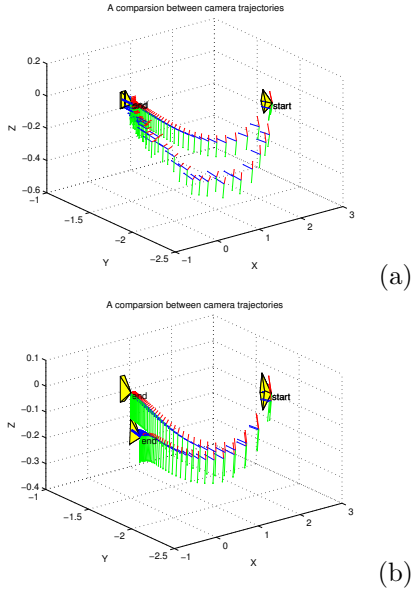


Figure 6: Camera trajectory comparison with the reference case. (a) with the novel view method, (b) with the M-estimator.