Performance Evaluation of Point Feature Detectors for Eye-in-Hand Visual Servoing

C. Lazar, Member, IEEE and A. Burlacu, Student Member, IEEE

Abstract—This paper presents a new approach to evaluate the performances of point feature detectors for eye-in-hand visual servoing systems. The performances are analyzed in terms of stability and robustness criteria defined for a sequence of images. The first image represents the start position and the last one contains the reaching of the desired position for grasping the target object. The performances are back-analyzed, every image being compared with the last one which contains the desired features. Real-time experiments with a set-up consisting of a six d.o.f ABB robot with an eye-in-hand configuration were used to evaluate the performance. Point features are extracted with Harris and SIFT detectors and experiments show better results for the last descriptor.

I. INTRODUCTION

Visual servoing allows by processing a visual feedback the positioning of the end-effector of a robot manipulator with respect to a target object or a set of features [1], [2]. Visual information can be object pose or image features related to the camera frame. Thus, visual servoing systems are classified in position-based and image-based structures. Visual servoing systems typically use one of two camera configurations: end-effector mounted or fixed in the workspace. The first, often called an eye-in-hand configuration, has the camera mounted on the robot’s end-effector.

Image-based visual servoing algorithm is a technique whereby features, derived from the image plane information, are servo controlled to a desired goal configuration. Tasks in Cartesian space are re-posed as a servo-control task directly in the image feature space. The image-based visual servoing approach is based on controlling the image feature kinematics. The accuracy of the positioning of the end-effector depends on the performance of the feature detection, matching tracking and modeling schemes [3].

Performance evaluation has become more and more important for computer vision, in the last years being reported evaluations of point feature detectors in the context of matching, recognition or texture classification [4]-[6]. Work has also been done on the evaluation of feature detectors for visual servoing application [7], [8]. It is more difficult to evaluate the performance of point feature detectors for eye-in-hand visual servoing due to the permanent varying of the camera position which implies viewpoint change, rotation, scale transform and illumination fluctuate.

In this paper, the performances of Harris and SIFT detectors are evaluated, related to visual servoing systems using a new approach. The performances are defined in terms of stability and robustness criteria computed for a sequence of images. For the last criterion, a new descriptor based on image moments is developed. The first image represents the start position and the last one contains the reaching of the desired position for grasping the target object. The performances are back-analyzed, every point features from each image being compared with those from the last one, which contains the desired features. To compare the performances of the point feature detectors, real time experiments were done with a set-up consisting of a six d.o.f ABB robot with an eye-in-hand configuration, having the camera mounted on the robot’s end-effector, a rectangular object and the obtained results are presented.

In the remainder of this paper, Section II presents how point features are used in image-based visual servoing structure and their performance evaluation using stability and robustness criteria. Section III presents an overview of Harris and SIFT detectors. In Section IV, experimental results are demonstrated and, finally, concluding remarks are given in Section V.

II. PERFORMANCE EVALUATION

Vision-based control is used to bring the robot manipulator to a grasping configuration that may be defined in terms of image features. The image-based visual servo control uses an error measured in the image to directly control the motion of the robot end-effector. The key relationship employed by all image-based approaches which uses point features is given by [9]:

\[ \dot{f} = L(f, z)\xi, \]

where \( f = [x_1, y_1, \ldots, x_n, y_n]^T \in R^{2n} \) is a vector with the image coordinates \([x_i, y_i]\) of each point feature, \( \xi \in R^n \) is the camera velocity and \( L \) is the interaction matrix depending...
on $f$ and $z$, a vector of depth values. Image-based control does not explicitly take camera motion into account, but determines a desired trajectory in the image features space and maps this trajectory, using the interaction matrix $L$, to a camera velocity.

In image-based control, the image error is defined by:

$$e(t) = f(t) - f_d$$  \hspace{1cm} (2)

in which $f_d$ is the desired feature employed as set-point in control algorithm. For the eye-in-hand visual servoing structure, the desired feature $f_d$ represents the point features obtained at the final position of the end-effector after the motion done for reaching an object target from a grasping configuration.

Considering the candidate Lyapunov function:

$$V(t) = \frac{1}{2} \| e(t) \|^2 = \frac{1}{2} e^T e$$  \hspace{1cm} (3)

the following control algorithm is derived [9]:

$$\xi = -\lambda L^{-1}(f, z)e(t), \quad \lambda > 0$$  \hspace{1cm} (4)

when the interaction matrix is square and nonsingular. For $L \in \mathbb{R}^{2nxm}$ and $2n > m$, $L^{-1}$ from (4) is replaced with the pseudoinverse $L^+$:

$$L^+ = (L^T L)^{-1} L^T.$$  \hspace{1cm} (5)

From the above equations, it results that the image-based control law accuracy depends on the quality of error measured in the image by means of point features. During the end-effector motion to the object target, the visual sensor takes images at every sampling period, called frames, from which the point features are derived.

The performance of these point features is very important for the quality of the image-based control systems. The point features must be stable in all frames of the image sequence obtained during the end-effector motion to the object target. In order to evaluate this performance, the following controller is developed. The point features $f_d$ from the final frame are considered and the ratio $r$ is computed using:

$$r = \frac{n_d}{n_k} k = \frac{1}{p}$$  \hspace{1cm} (6)

where $n_d$ is a number of point features from the final frame, $n_k$ the number of point features from every frame of the image sequence and $p$ the number of frames of image sequence. The stability performance shows the quality of point feature detectors regarding the repeatability of the point features detection in an image sequence and a good behavior means to obtain for $r$ a value as close as possible to the unit.

Visual information and, thus, point features are influenced by the camera motion by viewpoint change, rotation, scale transform and illumination fluctuate. To evaluate the robustness of image features with respect to camera motion, a descriptor invariant at translation, rotation and scale is developed in this paper. For this, it is considered the Euclidean distance between the centroid and all $n$ point features from an image:

$$d(i) = d ((x_i, y_i), (x_e, y_e)), \quad i = 1, n,$$  \hspace{1cm} (7)

where the co-ordinates of feature points centroid are given by:

$$x_e = \frac{1}{n} \sum x_i, \quad y_e = \frac{1}{n} \sum y_i.$$  \hspace{1cm} (8)

Based on the distance (7), translation, rotation, and scale invariant one-dimensional point features sequence moments can be estimated as the following moments [10] are defined:

$$m_i = \frac{1}{n} \sum_{i=1}^n d(i)$$  \hspace{1cm} (9)

and the centered moment:

$$\mu_i = \frac{1}{n} \sum_{i=1}^n \left[ d(i) - m_i \right].$$  \hspace{1cm} (10)

Using $\mu_i$ and $m_i$, less noise-sensitive results can be obtained from the following point features descriptor [11]:

$$F = \frac{\mu_i^{1/2}}{m_i}.$$  \hspace{1cm} (11)

First, the descriptor $F_d$ for the feature points from the final frame is computed. Then, the descriptors $F_k$ for the other frames of image sequence are computed and the robustness of the feature points can be evaluated using the robustness error:

$$e_r = F_d - F_k.$$  \hspace{1cm} (12)

For a good robustness performance, the error must be less than a threshold $\beta$.

By means of the two criteria presented in this section, it can be established, as a necessary condition, if the visual features extracted with a certain algorithm can be used to derive the image based control law.
Many different techniques for detecting point features have been developed. Among these, in this paper, Harris and SIFT detectors are used to evaluate the performances due to their success in visual servoing systems.

A. Harris detector

Harris corner detector is an algorithm based on an underlying assumption that corners are associated with maxima of the local autocorrelation function [12]. The autocorrelation function measures the local changes of the signal. This measure is obtained by correlating a patch with its neighboring patches that is with the ones shifted by a small amount in different directions. In case of a corner, the autocorrelation function is high for all shifts directions. It is less sensitive to noise in the image than most other algorithms, because the computations are based entirely on first derivatives. The algorithm has proved popular due to its high reliability in finding L junctions and its good temporal stability making it an attractive corner detector for tracking.

The local change of image intensity \( I \) is defined by [12]:

\[
E(\Delta x, \Delta y) = \sum_{x,y} G(x, y) \left[ I(x + \Delta x, y + \Delta y) - I(x, y) \right]^2 ,
\]

where \( G \) is a smooth circular window, e.g. the Gaussian:

\[
G(x, y) = \frac{1}{2\pi\sigma^2} e^{-\frac{(x^2+y^2)}{2\sigma^2}} .
\]

Approximating the first gradients by:

\[
I_x = I \ast [-1 \quad 0 \quad 1] \quad , \quad I_y = I \ast [-1 \quad 0 \quad 1]^T ,
\]

for small shifts \( E \) can be written:

\[
E(\Delta x, \Delta y) = \left( \Delta x, \Delta y \right) A(x, y) \left( \Delta x, \Delta y \right)^T ,
\]

where:

\[
A(x, y) = \sum_{x,y} G(x, y) \begin{bmatrix} I_{x}^2 & I_{x}I_{y} \\ I_{x}I_{y} & I_{y}^2 \end{bmatrix} = \begin{bmatrix} \langle I_{x}^2 \rangle & \langle I_{x}I_{y} \rangle \\ \langle I_{x}I_{y} \rangle & \langle I_{y}^2 \rangle \end{bmatrix} .
\]

Thus, the change \( E \) for small shifts is closely related to the local auto correlation function. Matrix \( A \) captures the structure of the neighborhood \( G \). A corner is detected if the matrix is of rank two and both eigenvalues are large. For each corner pixel is computed the cornerness value:

\[
cv = \frac{\langle I_{x}^2 \rangle + \langle I_{y}^2 \rangle}{\langle I_{x}^2 \rangle \langle I_{y}^2 \rangle - \langle I_{x}I_{y} \rangle^2}.
\]

A good corner is defined as having a small value \( cv \) below a threshold \( \alpha \). The number of surrounding pixels required to calculate \( cv \) is determined by the size of the Gaussian smoothing kernel. A 3x3 pixel smoothing kernel gives a 5x5 pixel computation area, a 5x5 pixel smoothing kernel gives a 7x7 pixel computation area, etc.

B. SIFT detector

SIFT detector is based on the algorithm developed by Lowe [13] for detection of scale-invariant image features. The algorithm is structured in four stages: two for invariant keypoints detection and other two for invariant keypoints tracking. In this paper, only the first two stages dedicated on detection of scale-space extremes and accurate keypoint localization are used for feature extraction.

In order to generate the scale space, an image denoted \( I(x,y) \) must be part of a convolution with the Gaussian kernel (14). The successive image filtering implementation for creating the scale – space is generated by:

\[
L(x, y, \sigma) = G(x, y, \sigma) \ast I(x, y) .
\]

If the \( \sigma \) parameter from (19) is considered to be the application:

\[
\sigma : [\sigma_{\min}, \sigma_{\max}] \times [0, S-1] \rightarrow \mathbb{R}
\]

with:

\[
\sigma(o,s) = \sigma_0 \cdot \frac{o+s}{S} ,
\]

then computation for the scale – space of \( I \) can be implemented recursively. In (21), \( o \) represents an octave of the \( \sigma \) axis from the scale space, \( S \) the number of levels for each octave and \( s \) the index of a level from the \( o \) octave.

Having the scale – space, the difference-of-Gaussian (DoG) space is developed using:

\[
D(x, y, \sigma) = L(x, y, \sigma) - L(x, y, 2\sigma) .
\]

In the DoG space, that can be interpreted as a derivation of the scale – space, there are considered positions \( \tau = [x \quad y \quad \sigma] \) where the value \( D(\tau) \) is larger than all of its 26 neighbors from the 3x3 neighborhood cube. The ensemble of these positions is denoted with \( A \). Using different criteria, the stability of each element from \( A \) can be established.
The first step is to find the extremes of DoG in the neighborhood of each element from \( \Lambda \). Lowe's approach was to express the \( D \) function using its Taylor expansion:

\[
D(\tau) = D(0) + \frac{\partial D}{\partial \tau} \tau + \frac{\partial^2 D}{\partial \tau^2} \tau^2.
\]  

(23)

The resulting extremes of function \( D \) are obtained by derivative calculus of (23) and in the scale – space have the location:

\[
\hat{\tau} = -\frac{\partial D}{\partial \tau} \frac{\partial D}{\partial ^2 \tau^2}.
\]  

(24)

For each candidate, keypoint interpolation of nearby data is used to accurately determine its position. Keypoints with low contrast are removed being considered as responses along edges.

IV. EXPERIMENTAL RESULTS

To compare the performances of the point feature detectors presented in Section III, experiments were performed with a set-up consisting of a six d.o.f ABB robot with an eye-in-hand configuration, having the camera mounted on the robot’s end-effector, and a rectangular object. The set-up structure is given in Fig. 1.

With the visual servoing system, 8 images using a sample period of 1 sec and having 640x480 resolution were acquired. The position to reach is defined by the 8th image of the sequence.

![Fig. 1 Visual servoing set-up](image)

It is considered the first image, which represents the start position and, in 8 steps, the desired position to grasp the target object is reached.

The trajectory of the end-effector consists in 4 steps (images from 1 to 4), where the closeness to the object and imposed depth are achieved and other more 4 steps (images from 5 to 8) which make up a rotation motion in order to obtain the correct grasping position of the object. The performances are back-analyzed, every image being compared with the last one which contains the desired features \( f_d \). The first 4 images are employed to evaluate the performances for the viewpoint change of the camera and the next 4 images for the image rotation. Due to the camera motion, illumination changes appear by shifting the camera position with respect to the light source.

![Fig. 2 Image sequence and point features obtained with Harris detector](image)
Point features are extracted with Harris and SIFT detectors from each image of the sequence. The results obtained with Harris detector are given in Fig. 2 and with SIFT detector in Fig. 3.

Applying Harris detector in the final image, the desired point features are obtained. A threshold $\alpha = 29$ was considered for the cornerness value. From Fig. 2, the 8th image, it results that the point features are the corner of the object. Doing a back-analysis, the last 4 images show us the influence of image rotation and the first 4 images the effect of image rotation. Through viewpoint change of the camera, the illumination level is modified resulting more corners in the first images as in the last image. Due to camera rotation, illumination is also changed leading to losing of the right-down corner in the 6th image.

Using SIFT detector, the consequence is that the desired point features are changed, being located in the holes of the object (Fig. 3). At the same time, this detector is not so vulnerable at viewpoint change, image rotation and illumination change as Harris detector.

In order to evaluate the performance of the two detectors, the criteria presented in Section II are employed. The repeatability of the point features detection in the image sequence was tested using the stability performance (6) and in Fig. 4, the obtained results with the two detectors are given.

The stability performance is very good for SIFT detector, being 1 for the last 7 images and not so good for Harris detector, when $r$ is far to 1 for viewpoint change and around 1 for image rotation.

The robustness performance is evaluated using the robustness error (12). To find out this performance, the feature descriptors $F_k$, $k = 1..8$, were computed for each detector and the obtained results are depicted in Fig. 5. The feature descriptors of Harris detector have a greater variation in comparison with SIFT detector.

Computing the robustness error using (12), larger values for Harris detector resulted, as it is shown in Fig. 6. SIFT detector has a good robustness for viewpoint change and an outstanding one for image rotation, whereas Harris detector
has a poor robustness for the first 4 images and a good one for the last 4 images.

Harris and SIFT detectors are also compared using the image error (2). From the start image to the desired one, the error is computed employing the positions of point features. The image error for Harris detector is given in Fig. 7, taking into account the evolution in all images of the sequence of the 4 point features from the desired image.

Using the 6 point features obtained in the last image with SIFT detector, acting in the same manner as in the case of Harris detector, the evolution of the image error from Fig. 8 is obtained.

All the tests made for the performance evaluation of the two point feature detectors show that with SIFT descriptor much better results are obtained.

V. CONCLUSIONS

In this paper, a new approach to evaluate the performances of point feature detectors for eye-in-hand visual servoing systems has been presented. The performances were defined in terms of stability and robustness criteria computed for a sequence of images. For the robustness performance, a new descriptor based on image moments was developed. The performances were back-analyzed, every point features from each image being compared with those from the last one, which contained the desired features. Point features were extracted with Harris and SIFT descriptors. To compare the performances of the point feature detectors, real time experiments were done with a set-up consisting of a six d.o.f ABB robot with an eye-in-hand configuration, having the camera mounted on the robot’s end-effector, and a rectangular object. The results obtained showed better performances for SIFT descriptor.

REFERENCES