Robust features tracking for robotic applications: towards $2\frac{1}{2}$D visual servoing with natural images

François-Xavier Espiau, Ezio Malis and Patrick Rives

Abstract— **This paper concerns the robust tracking of some features extracted from a sequence of images taken with an uncalibrated camera mounted on a mobile robot. Unlike most vision systems, the 3D structure of the observed objects is completely unknown. Thus, position-based visual servoing cannot be used. Similarly, one must be careful when using image-based visual servoing since the depths of the features are unknown. On the other hand, $2\frac{1}{2}$D visual servoing can easily deal with unknown environments since it is only based on projective reconstruction. In order to obtain a good projective reconstruction for a safe vision-based control, we propose in this paper a multi-scale real-time approach to extract robust features. Experiments show that our algorithm can be used in robotics applications when images are noisy and uncontrolled perturbations can break the continuity of the robot motion.**

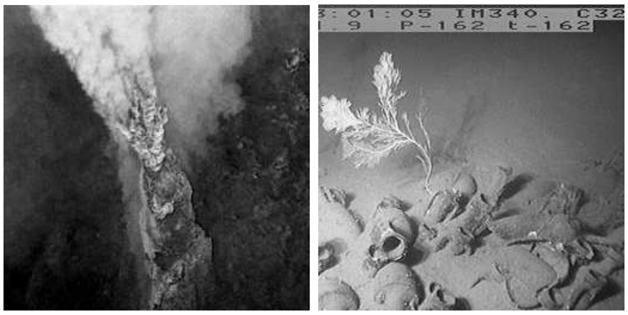
I. Introduction

The choice of a vision-based controller strongly depends on the amount of knowledge about the considered scene. If the 3D structure of the environment is completely unknown, position-based visual servoing [1] can not be used. Similarly, one must be careful when using image-based visual servoing since the depths of the observed objects are unknown and only approximations can be used [2]. In order to deal with completely unknown environments, model-free approaches have been proposed [3] [4]. These methods are based on projective reconstruction of the scene from only features matching. Thus, a key problem when implementing any robust model-free visual servoing is the extraction of good features to use in the projective reconstruction. When considering complex natural images (for example underwater images), we usually cannot observe simple geometric forms such as lines, cylinders or corners but only interest points. Thus, noise in the images can considerably influence the robustness of the extraction. Furthermore, if we use a single uncalibrated camera (instead of a stereo head) computer vision problems (like matching points) are much harder to solve. In robotic applications, the algorithms must be sufficiently robust to insure the safety of the environment and secondly they have to run as close as possible to real-time. The primary objective of this paper is to efficiently solve all these problems. Many approaches have been studied during the last years to deal with natural and often, complex images. In early vision, we usually extract lines or edges with the Canny detector [5], which needs to compute image gradient. Some more recent approaches, using deformable models called "snakes" [6], have been explored and give good results for tracking very complex unknown shapes in video sequences. The 2 1/2 D visual servoing has successfully been used with complex unknown planar contours in [7]. However, these approaches require to properly solve the problem of contour chaining which can be difficult in the case of natural noisy images and for a robotic application can be critical due to the high computational cost. Corners detectors [8] have been successfully used to extract interest points. Some recent works use the resolution of Partial Equation Differential [9] to improve the elimination of noise. Techniques based on Markov fields are also used with complex images. Often, this kind of techniques needs an off line learning step (especially for the textures) to be effective. So the images must have a good "structure" to detect textures quite efficiently. When dealing with video sequences, some authors exploit the motion informations. For example, the optical flow techniques (and all its derivatives like the features tracking [10]) are used to characterize the structure and motion in the scene. Such techniques work well assuming a dominant rigid motion in the scene and a small pixel displacements between two consecutive images in the sequence [11] [12]. In this paper, we consider the case when perturbations in the motion of the mobile robot can provoke sudden large displacements between two consecutive images of the sequence. This is an extremely difficult case which is not solved by standard tracking techniques. The method we propose is able to extract stable features from natural images when a lot of noise is present and when uncontrolled perturbations can break the continuity of the motion of the camera. The approach can be executed in real-time and thus it can be used in a 2 1/2 D visual servoing control scheme for controlling the camera motion.

All authors are with I.N.R.I.A. - I.C.A.R.E. Project - 2004, route des Lucioles - B.P. 93 - 06902 Sophia Antipolis Cedex, France. E-mail: {first name}.{last name}@sophia.inria.fr

II. Extracting robust features

In order to have a robust projective reconstruction for visual servoing, it is necessary to select the good features to extract from the images. Good features should correspond to physical points of the scene and they should be tracked well during the motion of the camera even if light conditions changes. The main problem is the quality of the images (see for example the images in Figure 1): bad gradient, very noisy. Furthermore, the structure of the scene, the camera parameters and the robot motion are unknown. Another problem in robotic applications is the speed of the algorithms. Real-time tracking becomes computationally costly when hundreds of features are tracked in each image. Indeed, it is generally better to track few stable and robust features in order to avoid the time execution problems and the amount of false matches between images. Thus, we obtain a better projective reconstruction. In this paper, we propose to select robust information through a multi-scale approach.



Hydrothermal chimney Amphorae

Fig. 1. Examples of underwater images. Features selection, extraction and real-time tracking are very difficult problems.

A. Multi-scale approach

The multi-scale approach needs a pyramidal structure for each image. The pyramids are built using a standard algorithm: convolution and sub-sampling. Let I_p be the p-th image in the video sequence. We denote $I_{p,0}$ the first level of the pyramid corresponding to the initial image (at the best resolution). So, a new image at level $(k+1)$ in the pyramid is defined by:

$$I_{p,k+1} = f(I_{p,k}) = h \circ (g * (I_{p,k})) \tag{1}$$

where h is a sampling function, g is a smooth function (eg: a Gaussian), $\circ$ and $*$ are respectively the composition and convolution operators. The use of a Gaussian function guarantees an isotropic smoothing in the image which eliminates quite well the noise. The sampling function is simple since we keep one pixel over two. This arbitrary choice doesn't influence the results in practice and is fast to run. The algorithm is stopped when the image size is 64x64 pixels (otherwise the signal is too damaged and becomes unusable). Finally, for each image, we obtain an associated pyramid which contains new smooth images while preserving the major characteristics of the signal.

B. Robust features selection

Due to the lack of simple geometric forms (see Figure 1), we choose interest points instead of lines or "snakes". Interest points are extracted from each image of the pyramid using a modified Harris detector [8] and implemented in an optimal way thanks to the recursively implementation of the Gaussian function proposed by Deriche in [13]. The Harris detector gives good results with natural images and it is well known for the repeatability of the detected points, for its robustness to noise, rotation and light changes [14]. As we already mentioned, from the point of view of robotic applications, a "good interest point" should be: i) *relevant* (i.e, it can be detected in several images corresponding to different poses of the camera and different conditions of illumination), ii) *accurately located* (i.e., which allows a good localization of the corresponding 3D point in the scene). These two requirements can be satisfied thanks to a robust matching of the interest points between the different levels of the pyramid. Indeed, the points detected at the highest level of the pyramid are insensitive to the successive smoothing steps (i.e. they are robust to noise). On the other hand, due to the filtering, they are not precisely localized in the image. If we are able to correctly propagate this set of points through the pyramid to the lower level, we improve their localization in the image (and consequently we improve localization of the the corresponding 3D point). In fact, due to the multiple matchings between the different levels of the pyramid, some cares have to be taken for implementing such an idea. The matching lies on a tree representation (see Figure 2) and the structure of the tree helps us in choosing good interest points.

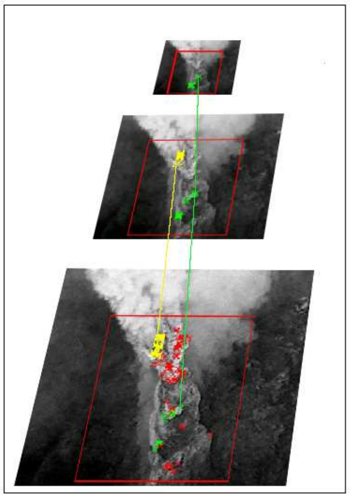


Fig. 2. Example of pyramidal matching

A detailed description of the matching process is given in [15]. The main idea is that a point is relevant and precisely located if it is present in a tree of the maximum length (corresponding to the number of levels in the pyramid) with only one matching at each level. Figure 2 shows the result of the extracted an matched points into the pyramid. The green points correspond to robust points: they are detected in the highest level and the pyramidal matching gives us the best localization in the initial image. Equally, the red points can be classified as noise. Finally, the best points obtained from each image of the sequence can be matched and used for projective reconstruction.

C. Matching features between images

After extracting robust interest points from different images, it is necessary to match them in order to obtain a projective reconstruction which can be used by the 2 1/2 D visual servoing [16]. As we use a single uncalibrated camera, we don't know the internal and external parameters. Even supposing that the scene is rigid, matching points is a very difficult problem. The main difficulty in our case is linked to the quantity of points and the unknown movement between images. If we use only methods based on correlation measures, we will have a high rate of outliers. Moreover, the inliers are not sufficient to compute the projective reconstruction with good precision. For this reason, we use a mixed method: correlation measure and epipolar geometry. To avoid the problem of outliers, we introduce again the use of the pyramid to match points. The main principle is represented on the Figure 3.

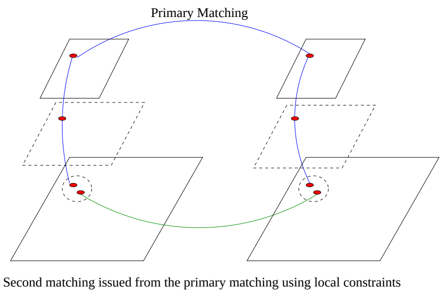


Fig. 3. Matching between two images with the help of the pyramids.

Before matching all points at the levels 0 of the two images, we only match points at the highest levels of the pyramid. Thus, in one hand we have a very good rate of inliers (near 100%) of robust points, and in the other hand, we can easily match points in the intermediate levels because we have labeled the robust ones (to avoid using them twice). Thus, we have an incremental matching within the pyramid and moreover we use local constraints around the neighborhood of the best points to avoid outliers as described on the Figure 3.

III. Projective reconstruction

Once the selected points have been matched between two images, one can obtain a projective reconstruction of the scene. For example one can estimated the fundamental matrix between two images [17] [18] or homographies associated to some planes (eventually virtual [16]). To avoid the problem of the outliers rejection, we can use the good extracted points found with the pyramidal structure. The pyramid has an essential role here: in one hand, we choose only good points and in the other hand, we have few points, and consequently less outliers. So it is easier to match them with a better degree of confidence. The algorithms we use are robust as described in [19]. The main advantage of the multi-scale approach here is the fact that we can match points at several scales of the pyramid and it is not necessary to use the epipolar constraint to insure we obtain good matches.

The aim of the method described in the paper is to provide robust informations in order to control the robot. Vision-based control of a robot can be achieved by recovering the motion of the camera between two views. If the observed structure is non-planar it is possible to compute a fundamental matrix from two views [17]. If the observed structure is planar we must compute a homography matrix since it is not possible to compute the fundamental matrix. The problem is that, in the applications considered in the paper, we don't know if the observed structure is planar or not. One solution consist in maintaining multiple models over time as proposed in [20]. Robust model selection consist to fit the data set first to the model $\mathcal{M}_1$ (the fundamental matrix), then to the model $\mathcal{M}_2$ (the homography matrix) and finally choosing the best model. However, model selection can be very critical and in the case of robotics applications switching between two different models leads to discontinuous control. For these reasons, we prefer to solve the structure from motion problem without model switching using a plane + parallax factorization [16]. A virtual reference plane π is selected by choosing 3 matched points $\mathbf{p}_{1i}$ and $\mathbf{p}_{2i}$ ($\{i = 1, 2, 3, ..., m\}$) in both images. Thus, the image points are related by the following equation:

$$\gamma_i \mathbf{p}_{1i} = \mathbf{H}_{12}\, \mathbf{p}_{2i} + \mu_i \mathbf{e}_{12} \quad \{i = 1, 2, ..., m\} \tag{2}$$

where $\mathbf{H}_{12}$ is the homography matrix linking the points on plane π, $\mathbf{e}_{12}$ is the epipole in the image $\mathcal{I}_2$, γ_i and μ_i are scalar factors. Note that $\mu_i = 0$ if the corresponding 3D point belongs to π. Thus, we have $\mu_1 = \mu_2 = \mu_3 = 0$. The epipolar line $\mathbf{l}_{2i}$ corresponding to the point $\mathbf{p}_{1i}$ in image is:

$$\mathbf{l}_{2i} \propto \mathbf{p}_{1i} \wedge \mathbf{H}_{12}\, \mathbf{p}_{2i} \propto \mathbf{p}_{1i} \wedge \mathbf{e}_{12} \quad \{i = 1, 2, ..., m\} \tag{3}$$

Consider the $3 \times m$ matrix $\mathbf{L}_2$ containing all the epipolar lines:

$$\begin{aligned} \mathbf{L}_2(\mathbf{H}_{12}) &= \left[\begin{array}{cccc} \mathbf{l}_{21} & \mathbf{l}_{22} & \dots & \mathbf{l}_{2m} \end{array} \right] \\ &= \left[\begin{array}{ccc} \mathbf{p}_{11} \wedge \mathbf{H}_{12}\mathbf{p}_{21} & \dots & \mathbf{p}_{1m} \wedge \mathbf{H}_{12}\mathbf{p}_{2m} \end{array} \right] \end{aligned} \tag{4}$$

this matrix must be $rank(\mathbf{L}_2) < 2$ (i.e. $\det(\mathbf{L}_2\mathbf{L}_2^T) = 0$) since all the epipolar lines meet in the epipole (i.e. $\mathbf{e}_{12}^T\mathbf{L}_2 = 0$). Moreover, if the observed scene is planar we have $rank(\mathbf{L}_2) = 0$ (i.e. $\|\mathbf{L}_2\| = 0$). Let $\mathbf{x}$ be a vector containing the entries of matrix $\mathbf{H}_{12}$. Thus, the homography matrix can be estimated by solving the following non-linear minimization problem:

$$\min_{\mathbf{x}} f(\mathbf{x}) = \|\mathbf{L}_2\|$$

$$\text{subject to} \tag{5}$$

$$\min_{\mathbf{x}} g(\mathbf{x}) = \det(\mathbf{L}_2\mathbf{L}_2^T)$$

Function g is minimized when $\partial g(\mathbf{x})/\partial \mathbf{x} = 0$. Thus, problem (5) is equivalent to the following problem:

$$\min_{\mathbf{x}} f(\mathbf{x}) + \boldsymbol{\lambda}^T \frac{\partial g(\mathbf{x})}{\partial \mathbf{x}} \tag{6}$$

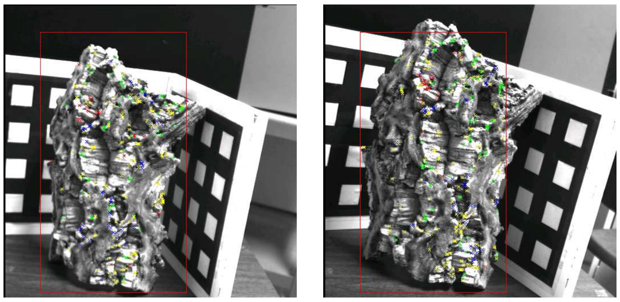
where $\boldsymbol{\lambda}$ is a vector containing the Lagrange multipliers. Finally, the robot can be controlled using the informations extracted from $\mathbf{H}_{12}$ as explained in [3]. In the next section, we show that the homography matrix $\mathbf{H}_{12}$ relative to a virtual plane of the scene can be measured even in cases when fundamental matrices cannot be estimated.

IV. Experimental results

In this section, we present some off-line results we have obtained on a PC Pentium IV 1GHz. The construction of the pyramid can be done at video-rate with a specific hardware. The sampling rate to extract, classify, matching points and compute the epipolar geometry is 50 ms.

A. *Ground truth with an calibrated stereo rig*

In order to validate our approach, a calibrated stereo rig has been used. Thus we have a ground truth which allow to compare our approach with the standard ones. From two images of an unknown object, we extract and match points with and without the multi-scale approach. When using the multi-scale approach a four level pyramid is build for each image (see Figure 4). The red points, extracted from level 3, are the more robust. The blue and the green points are extracted respectively from level 2 and 1. Finally, the yellow points are extracted from level 0 (the original image).



(a) Left Image (b) Right Image

Fig. 4. Points extracted with the multi-scale approach.

The numbers of points extracted in each image and the number of matched points between the two images are given in Table I. Using the multi-scale approach we extract less points but they are classified such that we can keep the more robust for matching. Thus, we have a better ratio of inliers for each image (i.e. correctly matched points / extracted points). Consequently, the error on the projective reconstruction is reduced by a factor of 10. The multi-scale approach has been tested on other images with calibrated stereo cameras and the results we have obtained show our approach gives a better projective reconstruction with a smaller number of robust points.

	N° selected points (image left vs right)	Matches
Mono-scale	304/385	85
Multi-scale	192/186	64

TABLE I

Results for a stereo calibrated rig

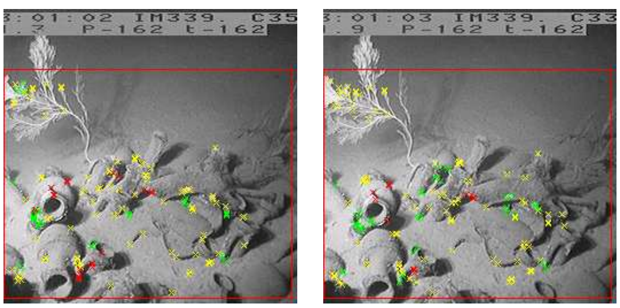
B. *Robust tracking*

The multi-scale approach has been tested with real underwater images of some "amphorae".

B.1 Robust matching

In the first experiment, we show which is the main advantage of choosing few robust points points. Figure 5 shows two images of the sequence. If we consider only the level 0 of the pyramid (i.e. the original image) and we try to match all the points extracted from both images, we only find two matches. Furthermore, matching a large number of points cannot be done in real-time. On the other hand, using the multi-scale approach with only a 3 level pyramid, we find six robust matches. Indeed, Figure 5 shows that 6 red points have been detected and matched at the highest level (level 2) of the pyramid. The green ones have been detected at the level 1 and the yellow ones are classed as noise because they only appear on the first level 0.

Even if linear algorithms for projective reconstruction need at least 8 points [18] [16], non-linear approaches are able to provide a solution with only 6 points.



(a) Left Image (b) Right Image

Fig. 5. Extraction of points with the pyramid for two images

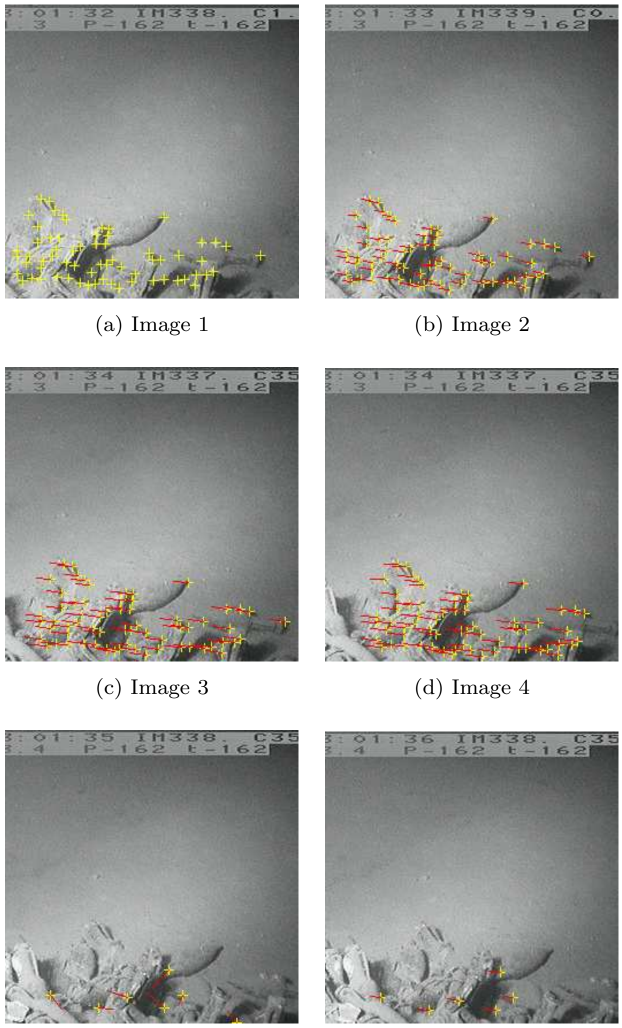
B.2 Estimating small movements

In the case of small movements or when the observed scene can be considered as a plane, the estimation of the fundamental matrix can become very unstable. On the other hand, a homography matrix associated to a virtual plane of the scene can always be estimated as explained in Section III. For example, if we consider the images of Figure 5, the 3D structure of the scene is close to a plane and the displacement is very small. When we compute the fundamental matrix, the epipolar lines are not well estimated. On the other hand, disparate standard tracking methods are able to estimate the same small pure translation. With our robust tracking we use fewer points and we find the same homography matrix:

$$H = \begin{pmatrix} 1.0000 & -0.0000 & 1.0002 \\ -0.0000 & 1.0000 & 2.0010 \\ -0.0000 & 0.0000 & 1.0000 \end{pmatrix}$$

B.3 Tracking with large displacements

When using a mobile robot in a real task, large uncontrolled displacements can occur during the motion of the camera. In such situation, we have compared our method to a robust and standard tracking method based on a Harris points extraction and correlation measures. This method is also fast to compute. Due to the locality aspect of correlation, when the displacements are large and illumination change, we have some problems to track features. Figure 6 shows the result of the standard approach (the red lines are the displacement of the points between image k and image $k-1$). When the displacements are small (around 8 pixels), the algorithm tracks the points. Suddenly, between the image four and five, the illumination changes and the displacement is too big (around 18 pixels). Most of the points are lost and we are not able to compute a robust projective reconstruction. Indeed, the few matched points are wrong. Things are even worse when the displacement between two consecutive images is larger.



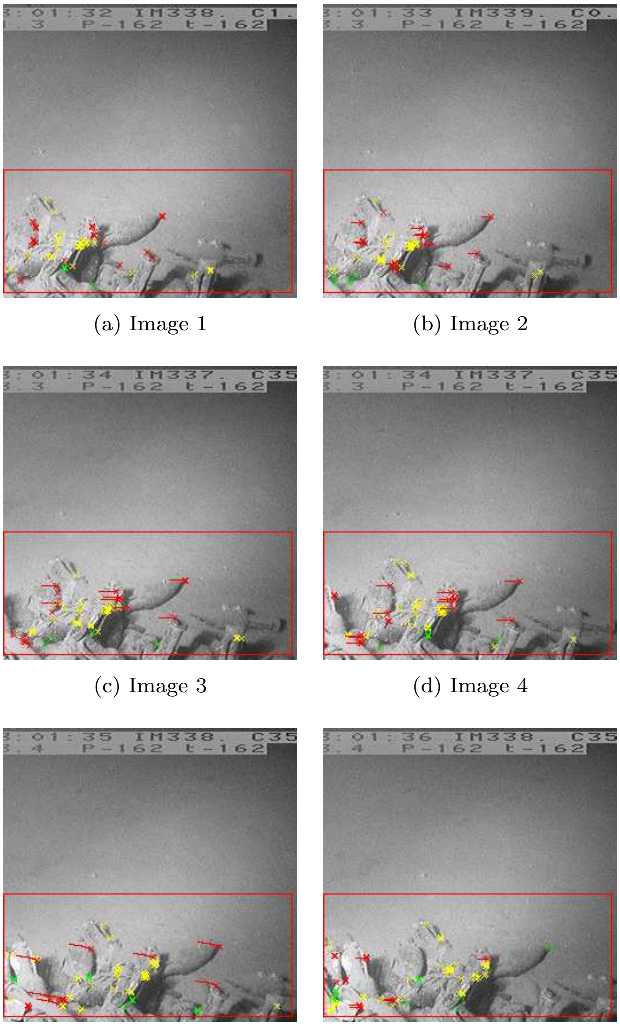
(a) Image 1 (b) Image 2

(c) Image 3 (d) Image 4

(e) Image 5 (f) Image 6

Fig. 6. Standard Tracking

In the figure 7, we present the results obtained with our approach, based on a multi-scale extraction, classification and selection of the points. It is easy to see, that if we can track less features, we can also track them all along the sequence in a robust way. The red points are the best points (i.e. extracted at the highest level of the pyramid), the green ones are less robust (extracted at the intermediate level). Finally, the yellow points (only extracted at the first level) can be matched but with a lower degree of confidence. Note that these points are in fact the same points extracted with the standard method. In this sequence, the main problem is located between images four and five. Even if we loose some points, we can continue to track some robust points and so we can robustly compute the projective reconstruction. This example show that the proposed approach is better than the standard one we have previously used to track features in natural underwater images. Future work will be devoted to compare our method with others standard tracking methods in the presence of larger displacements between images.



(a) Image 1 (b) Image 2

(c) Image 3 (d) Image 4

(e) Image 5 (f) Image 6

Fig. 7. Multi-scale Tracking

V. Conclusion and future work

In this paper, we have presented a method to extract and track robust interest points from noisy images in an unknown, natural and complex environment with an uncalibrated single camera. The method we propose is based on a multi-scale approach and a robust extraction and classification of interest points. Robustness is extremely important in order to obtain a good projective reconstruction. Experiments show that our method gives good results even when standard approaches fail. Future work will concern the real-time control of a robot in an unknown environment using 2 1/2 D visual servoing.

Acknowledgments

This work is supported by Ifremer, the French research institute for exploitation of the sea.

References

[1] W. J. Wilson, C. C. W. Hulls, and G. S. Bell, “Relative end-effector control using Cartesian position-based visual servoing,” *IEEE Trans. on Robotics and Automation*, vol. 12, no. 5, pp. 684–696, October 1996.

[2] B. Espiau, F. Chaumette, and P. Rives, “A new approach to visual servoing in robotics,” *IEEE Trans. on Robotics and Automation*, vol. 8, no. 3, pp. 313–326, June 1992.

[3] E. Malis, F. Chaumette, and S. Boudet, “2 1/2 d visual servoing,” *IEEE Trans. on Robotics and Automation*, vol. 15, no. 2, pp. 234–246, April 1999.

[4] R. Basri, E. Rivlin, and I. Shimshoni, “Visual homing: Surfing on the epipoles,” in *IEEE Int. Conf. on Computer Vision*, Bombay, India, January 1998, pp. 863–869.

[5] J. Canny, “A Computational Approach to Edge Detection,” *IEEE Trans. on Pattern Analysis and Machine Intellig ence*, vol. 8, no. 6, pp. 679–698, november 1986.

[6] M. Kass, A. Witkin, and D. Terzoppoulos, “Snakes: Active Contour Models,” *Int. Journal of Computer Vision*, vol. 1, pp. 321–331, 1988.

[7] G. Chesi, E. Malis, and R. Cipolla, “Automatic segmentation and matching of planar contours for visual servoing,” in *Int. Conf. on Robotics and Automation*, S. Francisco, CA, United States, April 2000, vol. 3, pp. 2753–2758.

[8] C. Harris and M. Stephens, “A Combined Corner and Edge Detector,” in *Proc. of the 4th Alvey Vision Conference*, 1988, pp. 147–151.

[9] P. Kornprobst, R. Deriche, and G. Aubert, “Image restoration via pde,” in *Conference SPIE 2942 : Investigative Image Processing*, Boston (USA), November 1996.

[10] C. Tomasi and T. Kanade, “Detection and Tracking of Point Features,” Tech. Rep. CS-91-132, Carnegie Mellon University, April 1991.

[11] A. Crétual and F. Chaumette, “Positionning a camera parallel to a plane using dynamic visual servoing,” in *IEEE Int. Conf. on Intelligent Robots and Systems*, Grenoble, France, September 1997, vol. 1, pp. 43–48.

[12] A. Crétual and F. Chaumette, “Dynamic stabilization of a pan and tilt camera for sub-marine image visualization.,” *Computer Vision and Image Understanding*, vol. 79, no. 1, pp. 47–65, July 2000.

[13] R. Deriche, “Using Canny’s Criteria to Derive a Recursively Implemented Optimal Edge Detector,” *Int. Journal of Computer Vision*, vol. 1, no. 2, pp. 167–187, may 1987.

[14] P. R. Beaudet, “Rotationally Invariant Image Operators,” *Int. Joint Conf. on Pattern Recognition*, pp. 579-583, 1978.

[15] F.-X. Espiau and P. Rives, “Extracting robust features and 3D reconstruction in underwater images,” in *Proceedings of OCEANS MTS/IEEE*, Honolulu, Hawaii, November 2001.

[16] E. Malis, F. Chaumette, and S. Boudet, “2 1/2 D visual servoing with respect to unknown objects through a new estimation scheme of camera displacement,” *Int. Journal of Computer Vision*, vol. 37, no. 1, pp. 79–97, June 2000.

[17] O. Faugeras, *Three-Dimensional Computer Vision: A Geometric Viewpoint*, MIT Press, 1993.

[18] R. Hartley, “In defense of the 8-point algorithm,” in *Proc. of the 5th ICCV*, 1995, pp. 1064–1070.

[19] Z. Zhang, R. Deriche, Q.-T. Luong, and O. Faugeras, “Robust recovery of the epipolar geometry for an uncalibrated stereo rig,” in *Proc. ECCV 94*, 1994, vol. I, pp. 567–576.

[20] P. Torr, A. W. Fitzgibbon, and A. Zisserman, “Maintaining multiple motion model hypotheses over many views to recover matching and structure,” in *Int. Conf. of Computer Vision*, Bombay, India, January 1998, pp. 485–491.