PERFORMANCE EVALUATION OF POINT MATCHING METHODS IN VIDEO SEQUENCES WITH ABRUPT MOTIONS

Wael Elloumi[*], Sylvie Treuillet, Remy Leconge and Aicha Fonte
Institut Prisme, Polytech'Orléans, 12 rue de Blois, 45067 Orléans cedex2, France

Keywords: Interest Points, Local Descriptors, Matching, Performance Evolution, Video Sequence, Abrupt Motions.

Abstract: In this paper, we compare the performance of matching algorithms in terms of efficiency, robustness, and computation time. Our evaluation uses as criterion, for efficiency and robustness, number of inliers and is carried out for different video sequences with abrupt motions (translation, rotation, combined). We compare SIFT, SURF, cross-correlation with Harris detector, and cross-correlation with SURF detector. Our experiments show that abrupt movements perturb a lot the matching process. They show also that SURF is the most disturbed, by such motions, and which even fails in cases that present a large rotation unlike the rest of descriptors as SIFT and cross-correlation.

1 INTRODUCTION

While the problem of image matching has been studied extensively for various applications, the remaining questions are which feature detector and descriptor performs the best and which one is the most appropriate to match images in real time for camera motion estimation. The field of our interest is the indoor/outdoor navigation assistance for the visually impaired with a low cost body mounted camera. Since vision based localization of mobile robots can rely on the assumption of the smooth movements, human motions are rougher and unpredictable and may cause loss in vision tracking. It will be wise to investigate the limits of the vision based localization especially in video sequences with abrupt motions.

Some previous works propose comparative studies of descriptors. Carneiro and Jepson (Carneiro and Jepson, 2002) introduce a phase-based local feature using Harris corner detector and compare it to the differential invariant features. They use the Receiver Operating Characteristic (ROC) curves as performance criterion and demonstrate that differential invariants are not the best for common illumination changes and 2D rotation. A variant of SIFT algorithm (Lowe, 2004) based PCA performs better on artificial data according to the recall-precision criterion (Ke and Sukthankar, 2004).

Another extension of SIFT outperforms many local descriptors but is more costly in computation time (Mikolajczyk and Schmid, 2005).

All the references cited above compare descriptors on image pairs using data set with artificial or real geometric and photometric transformations but not on video sequences. In this paper, we propose a performance comparison of three popular point matching methods. SIFT (Lowe, 2004), SURF (Bay and al., 2006), and Harris (Harris and Stephen, 1988) with cross-correlation. All of these algorithms are evaluated in efficiency, robustness, and computation time criterion, using the same scenario. The comparison is based on sequences acquired with a camera attached to a robot hand. Efficiency and robustness is evaluated by the number of inliers (correct matches between two images). Next section presents the experimental setup, before results in section 3 and conclusion.

2 EXPERIMENTAL SETUP

Our experimental prototype is composed of USB PC camera (320×240 pixels) fixed on the gripper of a 6dof robot arm and which are connected to a desktop. Intrinsic parameters of the camera are estimated by a prior calibration. The robot arm can be controlled manually by its remote controller or automatically by programming dedicated software, using Cartesian or joint coordinate systems, with an

[*] This study is supported by HERON Technologies SAS and the Conseil Général du LOIRET.

Elloumi W., Treuillet S., Leconge R. and Fonte A. (2010).
PERFORMANCE EVALUATION OF POINT MATCHING METHODS IN VIDEO SEQUENCES WITH ABRUPT MOTIONS.
In *Proceedings of the International Conference on Computer Vision Theory and Applications*, pages 427-430
DOI: 10.5220/0002823304270430

adjustable velocity. This prototype allows us to capture video sequences with rotations, translations and combined motions including zoom effect and abrupt motions.

Experiments data set consists of nine video sequences acquired at 30 fps frame rate in a real scene with brightness changes. We have chosen rotation, translation in the y-direction to create a zoom effect, and combined motion with rotation and translation, to compare the different operators, because these kinds of motions are the most disturbing for matching process. Table 1 give details on the video sequences related to 3 types of motion: number of frame, shift or rotation angle between frames. Several velocities of the robot arm were tested during acquisition: low, medium, or high velocity. Furthermore, to simulate more abrupt motions and considerable transformations, we matched distant key frames. Velocities of motions present in these video sequences are faster than normal motions of a human being. For example, the lowest velocity of translation is 45 cm per second and the lowest velocity of rotation 100 degrees per second.

Based on the state of art presented in the previous section, we have chosen to compare SIFT because it’s the most robust, cross correlation with Harris corner detector because it’s the fastest and SURF descriptor which is considered as a good compromise between computation time and robustness. To have well distributed Harris points, we have divided the images in buckets of size 15×15 pixels. The ZNCC correlation score is applied in 11×11 pixels ROI, with a minimum threshold of 0.8. The cross correlation is used with Harris and also with SURF detector to highlight the influence of the detector on matching process.

For evaluation, we observe the robustness and the computation time. The most popular metrics for robustness are ROC and Recall-Precision curves. Both are based on the number of correct matches and the number of false matches obtained for an image pair. We use the total number of correct matches (inliers) and the percentage of inliers compared to the total number of matched points (inliers and outliers), described by the equation (1).

$$\%inliers = \frac{correctmatches}{correctmatches + falsematches} \qquad (1)$$

The number of correct matches and false matches is determined with Least Median of Square algorithm (Zhang, 1998) by estimating the fundamental matrix in the image pair. The maximum distance from point to epipolar line, beyond which the point is considered an outlier and is not used for computing the final fundamental matrix is equal to 1 pixel. The desirable level of confidence that the matrix is correct is equal to 99%. The only constraint of this method is that we must have at least eight matched features. The computation of the two-view geometry requires that the matches originate from a 3D scene and that the motion is more than a pure rotation. That is respected as the camera is fixed slightly out of the rotation axis on the robot clip.

To develop our comparative study, we perform the following process for each video sequence:

1. Fix the number of frames to skip (frame jump) between images to match.
2. Extract distinctive features in images and match them using the different descriptors.
3. Select inliers from these candidates by estimating the fundamental matrix using LMedS method.

3 RESULTS

In this section, we present in Figure 1 and Table 2 an extract of the results for all carried experiments and discuss the performance of the tested descriptors.

3.1 Image Rotation

Matching is tested between images with a rotation angle between 7 and 120 degrees by varying velocity and image jump. The number of inliers clearly decreases for higher rotation velocity. SIFT descriptor is the most robust to rotation followed by SURF, which fails in fast rotation. Harris based matching is more disturbed than SURF based detector.

3.2 Image Scale Change

Scale change is achieved by a translation up to 1370 mm. All descriptors have a similar robustness, (% of inliers), slightly lower for cross correlation. SURF presents the lowest number of inliers for all velocities. The number of inliers decreases when increasing velocity of robot arm but much less than for the rotation case.

3.3 Combined Motion

Combined motion is performed by simultaneously rotating and translating the robot arm (between 4 and 92 degrees with 1370 mm shift). The performan-

Table 1: Video sequences data set.

Motion	Video sequences	Number of frames	Motion range per frame
Translation (scale change)	Translation_v25	90 frames	15.22 mm
	Translation_v50	49 frames	27.96 mm
	Translation_v100	32 frames	42.81 mm
Rotation	Rotation180_v25	55 frames	3.39 degrees
	Rotation180_v50	24 frames	8.18 degrees
	Rotation180_v100	11 frames	20 degrees
Combined	Combined_v25	47 frames	8.51 mm and 4 degrees
	Combined_v50	27 frames	14.81 mm and 7.2 degrees
	Combined_v100	14 frames	28.57 mm and 15 degrees

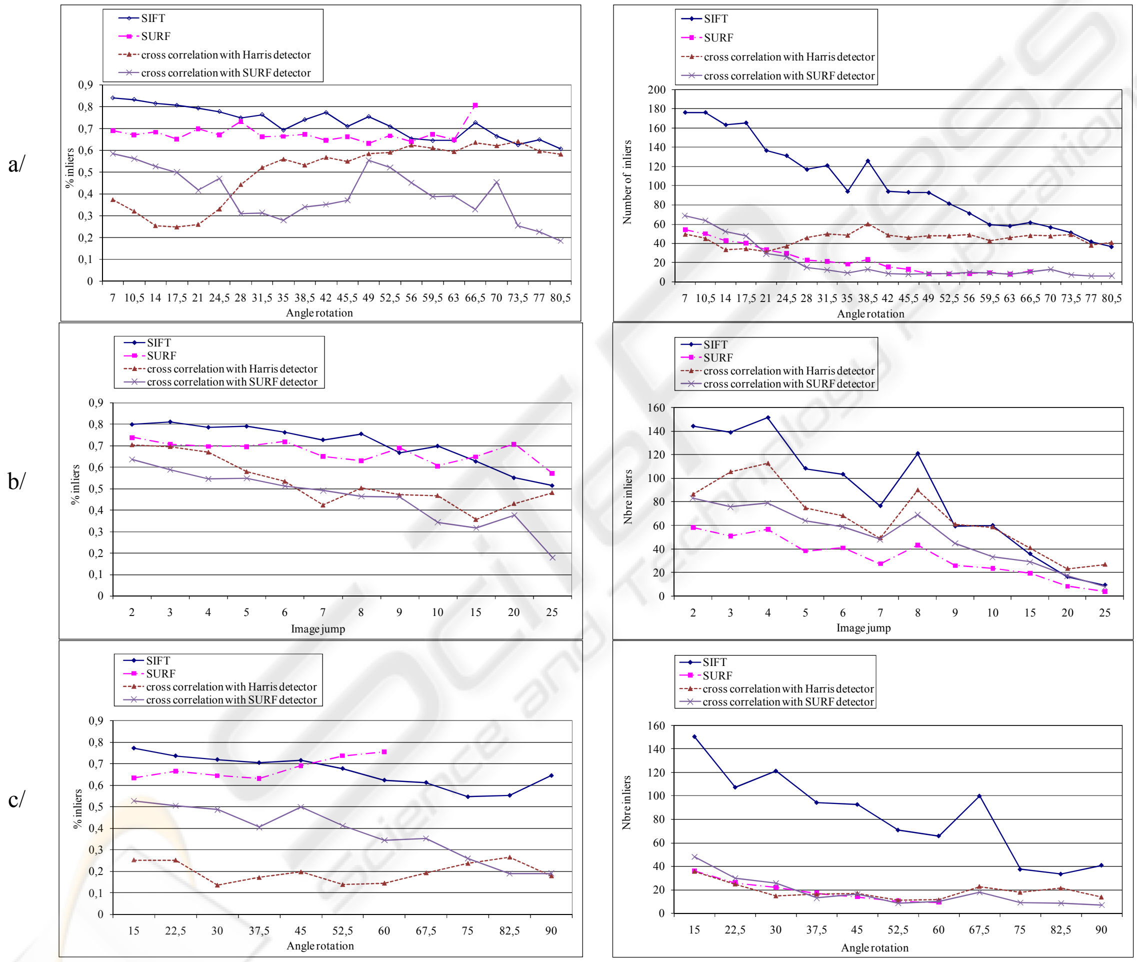

Figure 1: Evaluation of the robustness: % of inliers (left) and number of inliers (right). a/ Rotation180_v25 (first line) b/ Scaling Translation _v100 (second line) c/ Combined _v50 (third line).

Table 2: Computation time for an Intel dual core, 3 GHz, and 2GB memory (in milliseconds).

	SIFT			SURF			Cross correlation with Harris			Cross correlation with SURF		
Datasets	Min	Max	Mean	Min	Max	Mean	Min	Max	Mean	Min	Max	Mean
Trans_v25	1382	2065	1724	778	1411	1064	646	1817	918	718	1818	944
Rot180_v50	1433	2302	1706	610	1503	853	591	1229	888	564	1307	792
Comb_v100	1279	2354	1782	530	1732	886	604	1389	850	545	1327	814

mance of all operators is worse than for simple transformation. SIFT clearly outcomes other descriptors in number of inliers. SURF is better than cross correlation based descriptor only for low velocity (v25) and even fails for large rotation velocity. Rotation is more disturbing than scale changes.

4 CONCLUSIONS

An experimental comparison of the famous matching descriptors is proposed to identify the most appropriate to estimate camera motion. To be as close as possible to our application, we have used several real video sequences with abrupt motions (rotation, scale change, and combined). SIFT performs the best results in terms of number of inliers, but it can not be used for real time applications. SURF and cross correlation are worse than SIFT but can be improved in order to be applied for real time applications. SURF is interesting in the case of scale change. However, its performance becomes similar to the cross correlation in the case of large rotations. In our tests, the matching process is achieved around one second for the best in half VGA images. This matching time remains too high for localizing a person in real time with a body-mounted camera. To overcome this issue, we can use the GPU programming for additional speed up. We also plan to exploit the extra capabilities of the latest smart phones to improve performance. New smart phones contains fast camera which can be combined with accelerometer and a GPS receiver, and future devices will contain magnetic compasses and gyroscopes.

REFERENCES

Bay, H., Tuytelaars, T., Van Gool, L., 2006. SURF: Speeded Up Robust Features, *ECCV*.

Carneiro, G., Jepson, A. D., 2002. Phase-Based Local Features. *ECCV*, 282-296.

Harris, C., Stephens, M., 1988. A combined corner and edge detector. In *Alvey Vision Conf.*, 147–151.

Ke, Y., Sukthankar, R., 2004. PCA-SIFT: A More Distinctive Representation for Local Image Descriptors. *CVPR'04*, vol. 2, 506-513.

Lowe, D., 2004. Distinctive Image Features from Scale-Invariant Keypoints. In *Int. J. of Computer Vision,* vol.2, 91-110.

Mikolajczyk, K., Schmid, C., 2005. A performance evaluation of local descriptors. In *IEEE Trans. on PAMI,* 27(10), 1615–1630.

Zhang, Z., 1998. Determining the Epipolar Geometry and its Uncertainty. A Review in *International Journal of Computer Vision*, volume 27, n° 2, pages 161-198.