Reliable Features Matching for Humanoid Robots

Alberto Pretto, Emanuele Menegatti, Enrico Pagello†
Dept. of Information Engineering
The University of Padua, ITALY
†*also with* ISIB-CNR, Padua, ITALY
Email: alberto.pretto@dei.unipd.it

***Abstract*— This paper describes a visual feature detector and descriptor scheme designed to address the specific problems of humanoid robots in the tasks of visual odometry, localization, and SLAM (Simultaneous Localization And Mapping). During walking, turning, and squatting movements, the camera of a humanoid robot moves in jerky and sometimes unpredictable way. This causes an undesired motion blur in the images grabbed by the robot camera, that negatively affects the performance of the image processing algorithms. Indeed, the classical features detector and descriptor filtering techniques, that proved to work so well for wheeled robots, do not perform so reliably in humanoid robots. This paper presents a method to detect image interest points (invariant to scale transformation and rotations) robust to motion-blur introduced by the camera motion. Our approach is based on a preprocessing step to estimate the *Point Spread Function* (PSF) of the motion blur. The PSF is used to deconvolve the image reducing the blur. Then, we apply a feature detector inspired by SURF approach and the feature descriptor from SIFT. Experiments performed on standard datasets corrupted with motion blur and on images taken by a camera mounted on a small humanoid robot show the effectiveness of the proposed technique. Our approach presents higher performances and higher reliability in matching features in the different images of a sequence affected by motion-blur.**

I. Introduction

Local invariant features descriptors are widely used in the last years in many robotics tasks. In particular, several authors proposed localization techniques based on vision systems which exploit the process of matching invariant features along image sequences. Just to give some examples: in [11] interest points extracted from a perspective camera are used to build a topological map of the environment; in [1] a Monte-Carlo localization scheme based on feature matching of panoramic images is presented; in [22] the kidnapped robot problem is solved matching distinctive landmarks using a stereo camera; in [8] an invariant feature-based SLAM approach is presented using a perspective camera and EKFs (Extended Kalman Filter); in [3] a stereo-camera is used in a FastSLAM framework, in [2] a location graph-based Visual SLAM using an omnidirectional camerais presented; in [20] a visual odometry system based on feature detection and matching able to estimate the motion of a stereo camera or a perspective camera is presented. With respect to humanoid robot research, Bennewitz *et al.* [6] present a Monte-Carlo localization approach using invariant features tested on a humanoid robot equipped with a single perspective camera, Stasse *et al.* [23] present a 3D SLAM application for humanoid robots based on a standard EKF framework. In [26] invariant features are used to identify and localize objects in concrete humanoid scenarios. Despite that, image processing in humanoid robot is a very complex task: design and motion of the robot introduce a lot of blurring phenomena in the images grabbed by the camera. While in [26] is used the big and very expansive HRP-2 humanoid that perform a very stable slow walk, in [**?**] (where it is used a Sony QRIO robot) the motion-blur problem is taken into account and their approach attempts to minimize this problem. Bennewitxz *et al.* highlight in [6] that due to unstable motion of the humanoid platforms, missing odometers, severe body vibrations, and shaking of the camera, standard localization techniques are less robust on a humanoid robot compared to a wheeled robot. As noted by Berthoz in his plenary talk at ICRA 2007 the gait of a humanoid robot should be designed in order to stabilize the head and so to simplify perception as it is done by animals and humans. However, this is not simple for the current humanoid technology. Moreover, we are interested in working with small humanoid robots. Indeed, the final aim of our work is the development of a robust visual odometry for small and cheap humanoid robots, that usually have no odometry information from its servo-motors. In addition, small humanoid robots oscillate and vibrate a lot while moving, and usually are equipped with low cost cameras, which are high sensitive to motion blur phenomena.

This work proposes an invariant features detector-descriptor scheme robust to motion-blur effect. The basic idea is to introduce a blind-deconvolution step before starting the invariant feature detection and description process. During this step, the parameters of the unknown blurring function (PSF, *Point Spread Function*) are estimated using a direct method. If a motion blur is detected (the PSF magnitude is over a preset threshold), the image is deblurred according to the estimate PSF using an efficient Wiener filter. The invariant features detection and description is then performed in the restored image.

A. Overview

In Section II are summarized previous works on detection and description of invariant features and some known techniques in image motion-blur estimation and deconvolution. Section III introduces our interest point detector and descriptor scheme, based on the state-of-the art previous methods. In Section IV is presented the technique we use for the blind-deconvolution of the image before searching for interest points. Experiments and comparisons with other techniques on stan-

dard dataset and on images grabbed by our small humanoid robot are presented in Section V.

II. Related Work

A. Local features detector and descriptors

The best-known and widely used feature detector and descriptor scheme it was introduced by Lowe [14] and called SIFT (Scale-invariant feature transform). SIFT features are invariant to image scale and rotation, and are quite robust in matching across affine transformations and changing of viewpoint. For scale space analysis, images are convolved with a Gaussian kernel (the only possible smoothing kernels for scale space representation, as showed by Lindeberg [13]). Interest points are efficiently detected using a Difference-of-Gaussian (DoG) filter and taking the maxima over spatial and scale dimensions. At each extracted point, it is assigned an orientation using orientation gradients of neighbour sample points, computed on smoothed image correspondent on characteristic scale of the extracted point. According to the orientation, it is assigned a 128-dimensional vector, that represent the local oriented gradients in a region around the point. As shown in [17], SIFT features outperformed previous features detectors-descriptor schemes (e.g. shape context [5], steerable filters [9], differential invariants [10]). Ke *et al.* proposed a variation of the SIFT features, called PCA-SIFT: applying PCA in the gradient images, the descriptor is reduced to a 36-dimensional vector, matching step is so faster. PCA-SIFT are robust to focus-blur noise, but are less distinctive compared with SIFT [17]. Mikolajczyk *et al.* [18] proposed a novel approach for detecting interest points invariant to scale and affine transformation: interest points are chosen by detecting the local maxima of the Harris function of the image over location, and the local maxima of the Laplacian-of-Gaussian (LoG) over the scale. The affine shape of a point neighborhood is then estimated based on second moment matrix. In [17] is presented a novel descriptor called GLOH (Gradient location-orientation histogram), an extension of the SIFT descriptor designed to increase its robustness and distinctiveness: it also uses PCA to reduce the dimension of the descriptor. GLOH obtains little better performance than SIFT but at cost of less computational efficiency. Recently, Bay *et al.* [4] present a novel and computationally efficient scale and invariant feature detector-descriptor called SURF (Speeded Up Robust Features): to detect interest points, the maxima over location and scale of the determinant of the Hessian are selected. The Hessian is computed over scaled images using an efficient approximation based on the integral images technique. Descriptors are obtained calculating Haar Wavelet responses in the regions around interest points using integral images. Repeatability and distinctiveness performance are similar to previous proposed schemes, but SURF features can be computed much faster.

B. Blind deconvolution

Motion blur is the effect of the relative movements between the camera and the objects of the observed scene during the exposure time (i.e., the integration time of the grabbed image). Using deconvolution techniques, the image can be partially deblurred. Richardson-Lucy algorithm [15] and Expectation-Maximization [12] method are well known iterative deconvolution procedure, Wiener filter [21] is a non-iterative image restoration technique that tries to build an "optimal" estimate of the unblurred image. In order to restore the image with deconvolution techniques, the motion blur parameters (direction and extent) must be known: if they are unknown, the image restoration process is called *blind deconvolution*. In [25] is presented the whitening method: this is a non-iterative method that identify the PSF by high-pass filtering the blurred image. The filtered image is characterized mostly by the correlation property of the blur function. In [24] the PSF is obtained searching for image moments that are invariant with respect to the motion blur. In [16] blur direction is determined by an inertia-like tensor, while the extent is determined finding zeros of the blur slice of the power spectrum or bispectrum in this direction. In [19] PSF is estimated by using Radon transform to find direction and fuzzy set concepts to find its extend.

III. Interest Point Detector and Descriptor Scheme

The proposed invariant features scheme takes advantage of two successfully approaches: we use a detector method similar to SURF features [4] and the descriptor proposed in SIFT features scheme[14].

A. Interest Point Detector

The first step is to select a set of interest point that are invariant to scale transformation. This is performed searching for features in a scale space representation of the images [13], obtained convolving the original images $I(x, y)$ with Gaussian smoothing filters $G(x, y, \sigma)$ and increasing standard deviation values σ (normally referred as the *scale* of the smoothed image):

$$L(x, y, \sigma) = G(x, y, \sigma) * I(x, y) \tag{1}$$

The scale space is divided in *octaves* (i.e. the last smoothed image of the octave has twice the scale of the first). As in [14], each octaves is divided into an integer number s of intervals, with scales $\sigma_i = \sigma_{i-1} * 2^{\frac{1}{s}}$, where σ_0 is the initial scale chosen to be 1.6. We choose $s = 3$, so we compute eq. (1) at scales $1.6, 2.0159, 2.5398, 3.2, 4.0317$. The latest scale is computed to detect local scale space maxima at the higher scale of the octave, i.e. 3.2. Once an octave is completed, the image is resampled to half its original size: this image has obviously twice the scale of the original image. A new octave is then processed on the resampled (smaller) image, using the same σ_i values. Normally the number of the octaves is 4, it can be reduced to obtain much faster computation of the detector. In order to detect interest points, the scaled images $L(x, y, \sigma)$ are convolved with filters that response mainly to invariant local features of the image. Harris and Hessian based detectors response to corners and highly textured points,

whereas Difference-of-Gaussian (DoG) (used in [14]) and Laplacian-of-Gaussian (LoG) based detectors response mainly to blobs: the latters descriptors are less stable due to the possibility to detect points closed to contours of straight edges [18]. As in [4], we use the determinant of Hessian of the scaled image for selecting both location and characteristic scale of the interest points: the trace of Hessian is the LoG, taking the determinant points in witch the second derivative change in only one direction are penalized (e.g. straight edges):

$$det(\sigma^2 H(x,y,\sigma)) = det\begin{bmatrix} \sigma^2 L_{xx}(x,y,\sigma) & \sigma^2 L_{xy}(x,y,\sigma) \\ \sigma^2 L_{xy}(x,y,\sigma) & \sigma^2 L_{yy}(x,y,\sigma) \end{bmatrix} \quad (2)$$

where in eq. (2) L_{xx}, L_{yy}, L_{xy} are the second derivatives of the scaled images $L(x,y,\sigma)$. The second derivatives are multiply with the square of the scale σ: this is due to the fact that the amplitude of spatial derivatives decreases with scale, so normalization is required for true scale invariance [13]). The implementation strategy is to convolve the initial image with Gaussian smoothing filter at different scales: at this scope, we use an efficient Gaussian smoothing algorithm provided with OpenCV library[1]. First and second derivatives are then computed in scaled images: in eq. (3) the first and second Gaussian derivatives in x direction are computed.

$$\begin{aligned} L_x(x,y,\sigma) &= L(x+1,y,\sigma) - L(x-1,y,\sigma) \\ L_{xx}(x,y,\sigma) &= L_x(x+1,y,\sigma) - L_x(x-1,y,\sigma) \end{aligned} \quad (3)$$

First derivatives are stored in memory for efficient computation of the descriptors (see Section III-B), second derivatives are used to compute the determinant of Hessian (eq. 2). In Figure 1 are showed the resulting derivative based filters using in our approach. To improve computation speed of the detector, it

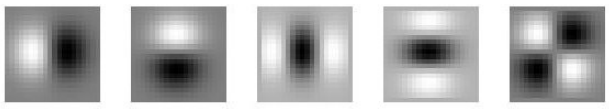


Fig. 1. Gaussian derivatives up to second order ($\sigma = 3.6$).

can be used more approximated Gaussian derivatives: in [4] discretised and cropped filter are used, computed using the *integral images* technique. We show only a slight degradation of the descriptor reliability using this fast method. Once computed the determinant of Hessian for each location of the multi-scaled image, interest point are detected searching for local maxima over scale and location space in a $3 \times 3 \times 3$ neighbourhood of each point: only local maxima with determinant of Hessian greater than a threshold are selected as interest points. Finally the location and the scale (called *characteristic scale*) of the extracted points are interpolated [7] by fitting a 3D quadratic to the scale-space determinant of Hessian and taking the maxima of this quadratic. This step is useful to obtain a more accurate characteristic scale of the point (negatively affected by the discrete nature of the scale space) and to reduce the localization errors.

[1]http://sourceforge.net/projects/opencvlibrary/

B. Interest Point Descriptor

In our experience we note that the SIFT descriptor is slightly more stable than SURF descriptor. We decided to implement SIFT descriptor, tuning the parameters of the algorithm to improve reliability.

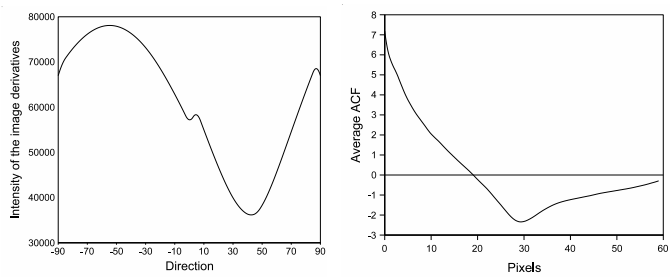
1) Orientation assignment: In order to assign orientations to detected interest points, we compute gradients orientations and magnitudes of 16×16 regularly spaced sample points into a square window centered around the interest point. The side length of the window is equal to $12 * scale$, where $scale$ is the interpolated characteristic scale (see Section III-A). Gradients magnitudes and orientation of sample points are computed using the stored first Gaussian derivatives in the discretised scale closed to the characteristic scale of the interest point:

$$\begin{aligned} m(x,y,\sigma) &= \sqrt{L_x(x,y,\sigma)^2 + L_y(x,y,\sigma)^2} \\ \theta(x,y,\sigma) &= tan^{-1}\left(\frac{L_y(x,y,\sigma)}{L_x(x,y,\sigma)}\right) \end{aligned} \quad (4)$$

The magnitudes are Gaussian-weighted with a circular bivariate Gaussian centered in the interest point with standard deviation equals to $2.5 * scale$. Magnitudes are then accumulated into an orientation histogram with 36 bins representing the discretised orientations of the gradients. After an histogram-smoothing step, the bins with values greater than 0.8 the global histogram maximum are selected: multiple interest points are created with the initial location and scale but with these different orientations (interpolated with histogram neighborhood).

2) Descriptor assignment: It is selected a square window centered around the interest point with side length of $20 * scale$ and oriented according with its orientation. This region is regularly divided into 4×4 smaller sub-regions, each containing 4×4 regularly spaced sample points. The gradients orientations and magnitudes of the sample points are computed as in Section III-B.1. Magnitudes are then Gaussian-weighted with a circular bivariate Gaussian with standard deviation equals to $6.7 * scale$ in order to increase stability of the descriptor towards small affine transformation and localization errors. To avoid high variations in the distribution of the gradients inside a sub-region caused by small pixels shift, magnitudes are further weighted with a weight of $1 - d$, where d is the distance of the sample point from the central value of the bin as measured in units of the histogram bin spacing [14]. Each sample point gradient is rotated according to the interest point orientation, then its magnitude is accumulated in a orientation histogram with 8 bins (i.e., 8 discretised orientations) characteristic of the sub-region. The 4×4 8-bins histograms form the 128-entry descriptor of the selected interest point. The descriptor is finally normalized to an unit vector in order to obtain invariance toward contrast variations.

IV. Motion Deblurring



(a) (b)

Fig. 2. (a) The motion-blur direction identification: The global minimum fall in the blur direction estimation (in degrees). (b) The average ACF used for estimation of the extent. The global minimum fall in the blur extent estimation (in pixels).

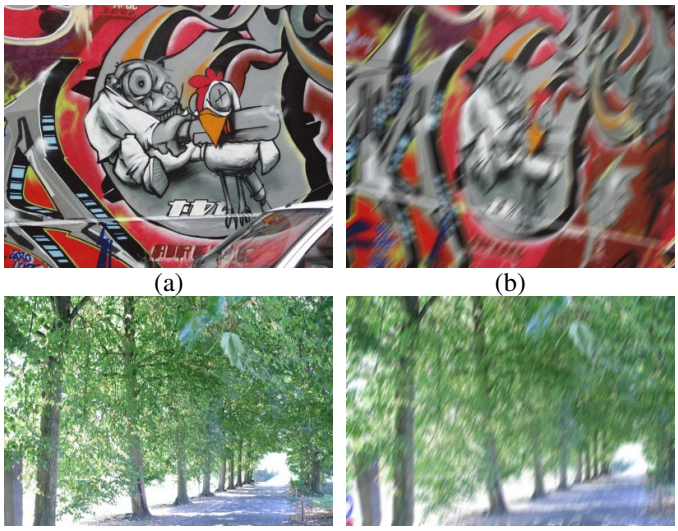
(a) (b)

Fig. 3. (a) The Blurred image. (b) The image restored with Wiener filter.

V. Experiments

We tested our detector and descriptor using both a standard dataset [2] with added synthetic motion-blur (some example in Figure 4) and real images with motion-blur effect grabbed by the CMOS camera that equip our humanoid robot (some example in Figure 5). Aim of our tests is to evaluate the effectiveness of the proposed method in matching images taken in a real humanoid robot scenario in presence of motion-blur phenomena. Testing image pairs are composed by two viewpoint of the same scene, the second frame present the motion-blur effect. Standard dataset we use is provided with *homographies* (plane projective transformations) between images: the map between images is known, every point in one frame corresponds exactly to one point in the other frame. We can determine in this case ground truth matches and also the accuracy (i.e. the localization error of the matches). For real images set , we count manually correct matches between frames (Figure 8). Our approach is compared to the SIFT features scheme [14] and to the SURF features scheme [4]. Comparisons are performed using well-known implementation of these methods[3] [4], without changing the algorithms standard

[2]http://www.robots.ox.ac.uk/~vgg/research/affine/
[3]http://web.engr.oregonstate.edu/~hess/index.html
[4]http://www.vision.ee.ethz.ch/~surf/



(a) (b) (c) (d)

Fig. 4. Some of the standard dataset images used in the experiments. Second image of every pair is blurred with a synthetic linear motion function with different directions and extents.

parameters. SIFT features implementations usually double the image size before starting features detection in order create more sample points. This step increase the computational cost and decrease robustness of matching in presence of motion-blur phenomena: for these reasons, we skip the resize step. We refer here to the proposed descriptor-detector scheme as *MoBIF* (Motion-Blur Invariant Features). For all tested approaches, we use the Nearest Neighbor Distance Ratio matching strategy (see [17]), with distance ratio equal to 0.5. In Figure 6 are presented matching results of the standard dataset image pairs in Figure 4. Images 4(b) and 4(d) are blurred with synthetic motion-blur function of directions -45 and 23 degrees and extents of 30 and 20, respectively. The matching accuracy is the distance in pixels between the ground truth match and the obtained match. MoBIF approach outperforms SIFT and SURF in both the number of correct matches and the localization accuracy. This is very important especially in visual odometry tasks, where the accuracy in matching affect significantly results in motion estimation. Results for some real images are presented in Figure 7: the X axis represent the image pairs used in matching process. Image pairs 1 and 2 are shown in Figure 5. Estimated motion-blur extents are in these cases 13, 14, 12, 20, 23 and 19, respectively. Also with real images MoBIF outperforms other approaches, with higher number of correct matches (Figure 7(a)) and a very high and stable correct matches ratio over all detected matches (Figure 7(b)). Especially with large motion-blur function extent (test image pair 5, estimated extent equals to 23) our approach preserves the reliability in matching (Figure 8) where SIFT and SURF techniques tend to fail.

VI. Conclusions

In this paper we presented an invariant features detector and descriptor scheme that outperforms the previous proposed

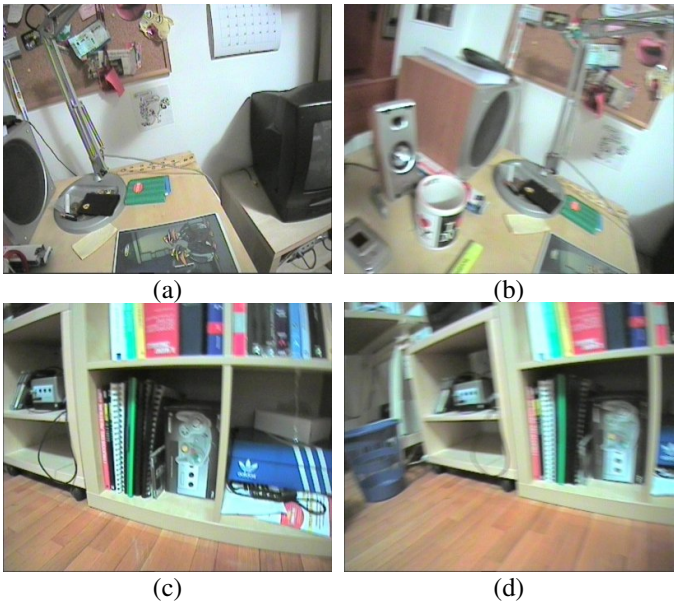


Fig. 5. Some of the real images used in the experiments. In (a),(b) the image pair 1, in (c),(d) the image pair 2. Second image of every pair present some motion-blur phenomena, PSF parameters are not a-priori known.

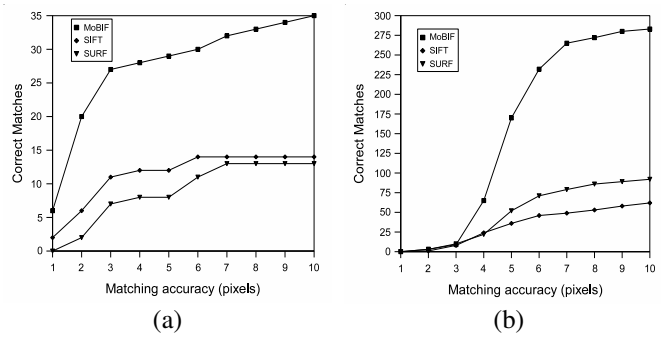


Fig. 6. Correct matches for the standard dataset images of Figure 4. Accuracy is the distance in pixels between the ground truth match and the obtained match.

methods in presence of motion-blur phenomena introduced by the camera movements. Experiment in artificially motion-blurred images and in real blurred images taken from the camera of a small humanoid robot shown the effectiveness and robustness of the proposed method. The basic idea began from the purpose of the authors to implement a robust visual-odometry algorithm for small humanoids robot: in future works we will present a visual-odometry framework for humanoids robot based on the proposed invariant features.

REFERENCES

[1] H. Andreasson, A. Treptow, and T. Duckett. Localization for mobile robots using panoramic vision, local features and particle filter. In *Proc. of the 2005 IEEE International Conference on Robotics and Automation (ICRA)*, 2005.

[2] Henrik Andreasson, Tom Duckett, and Achim Lilienthal. Mini-slam: Minimalistic visual slam in large-scale environments based on a new interpretation of image similarity. In *Proc. of the 2007 IEEE International Conference on Robotics and Automation (ICRA)*, 2007.

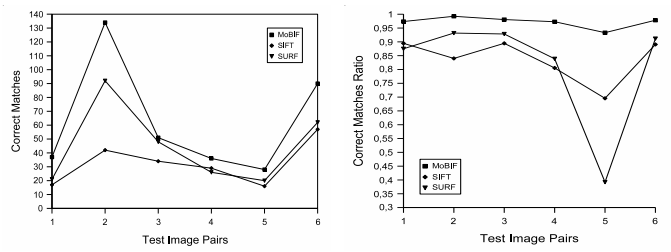


Fig. 7. (a) Correct matches for the real images set. In the x axis, the correspondent image pair. Estimated PSF extents are 13, 14, 12, 20, 23, 19, respectively. (b) Correct matches ratio over all detected matches

[3] T.D. Barfoot. Online visual motion estimation using fastslam with sift features. In *Proc. of the 2002 IEEE/RSJ Int. Conf. on Intelligent Robots and Systems (IROS)*, 2005.

[4] Herbert Bay, Tinne Tuytelaars, and Luc Van Gool. Surf: Speeded up robust features. In *Proc. of the ninth European Conference on Computer Vision (ECCV)*, 2006.

[5] S. Belongie, J. Malik, and J. Puzicha. Shape matching and object recognition using shape contexts. *IEEE Trans. Pattern Analysis and Machine Intelligence*, 2(4):509–522, 2005.

[6] Maren Bennewitz, Cyrill Stachniss, Wolfram Burgard, and Sven Behnke. Metric localization with sift features using a single camera. In *Proceedings of European Robotics Symposium (EUROS), Palermo / Italy*, 2006.

[7] M. Brown and D. Lowe. Invariant features from interest point groups. In *BMVC02*, 2002.

[8] Andrew J. Davison, Ian D. Reid, , Nicholas D. Molton, and Olivier Stasse. Monoslam: Real-time single camera slam. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 29(6):1–16, 2007.

[9] W. Freeman and E. Adelson. The design and use of steerable filters. *IEEE Trans. Pattern Analysis and Machine Intelligence*, 13(9):891–906, 1991.

[10] J. Koenderink and A. van Doorn. Representation of local geometry in the visual system. *Biological Cybernetics*, 55.

[11] J. Kosecka and F. Li. Vision based topological markov localization. In *Proc. of the 2004 IEEE International Conference on Robotics and Automation (ICRA)*, 2004.

[12] R.L. Lagendijk, J. Biemond, and D.E Boekee. Identification and restoration of noisy blurred images using the expectation-maximization algorithm. *IEEE Transactions on Acoustics, Speech, and Signal Processing*, 38(7):1180 – 1191, 1990.

[13] Tony Lindeberg. Scale-space theory: A basic tool for analysing structures at different scales. *Journal of Applied Statistics*, 21(2):224?270, 1994.

[14] David G. Lowe. Distinctive image features from scale-invariant keypoints. *International Journal of Computer Vision*, 60(2):91–110, 2004.

[15] L. B. Lucy. An iterative technique for the rectification of observed distributions. *Astronomical Journal*, 79.

[16] C. Mayntx, T. Aach, and D. Kunz. Blur identification using a spectral inertia tensor and spectralzeros. In *Proc. of 1999 International Conference on Image Processing (ICIP)*, 1999.

[17] Krystian Mikolajczyk and Cordelia Schmid. A performance evaluation of local descriptors. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 27(10):1615–1630, year = 2005,.

[18] Krystian Mikolajczyk and Cordelia Schmid. Scale and affine invariant interest point detectors. *International Journal of Computer Vision*, 60(1):63–86, 2004.

[19] M.E. Moghaddam and M Jamzad. Motion blur identification in noisy images using fuzzy sets. In *Proc. of the Fifth IEEE International Symposium on Signal Processing and Information Technology*, 2005.

[20] D. Nister, O. Naroditsky, and J. Bergen. Visual odometry. In *Proc. of the 2004 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR)*, 2004.

[21] A. Rosenfeld and A. C. Kak. *Digital Picture Processing*, volume 1. Academic, 1982.

[22] S. Se, D.G. Lowe, and J.J. Little. Global localization using distinctive

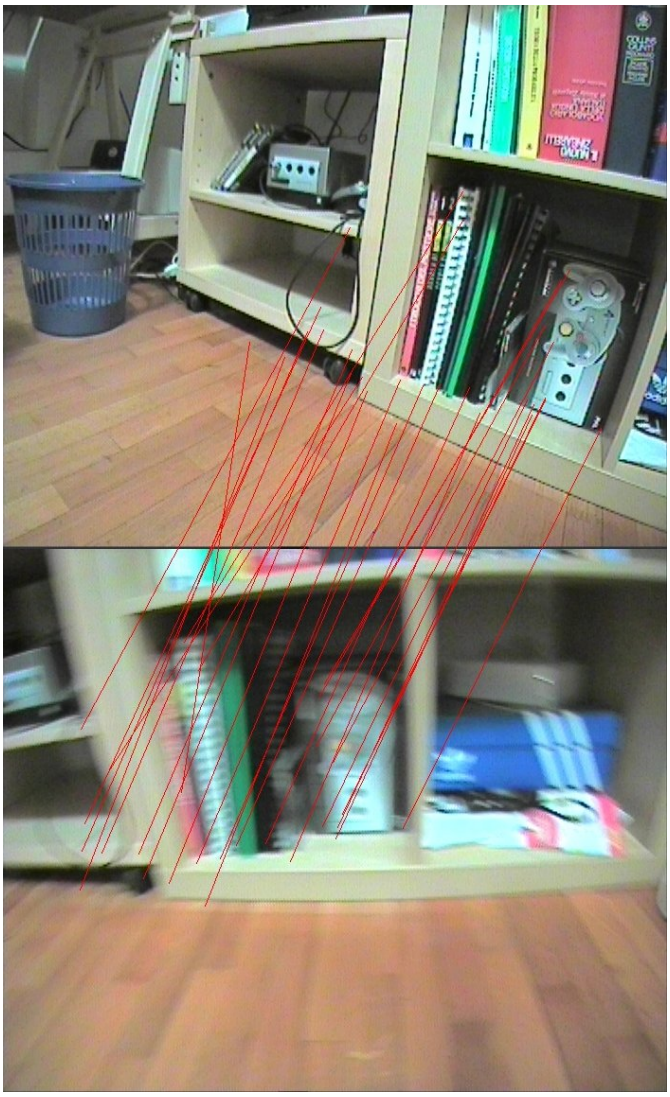


Fig. 8. MoBIF detected matches for the test image pair 5.

visual features. In *Proc. of the 2002 IEEE/RSJ Int. Conf. on Intelligent Robots and Systems (IROS)*, 2002.

[23] Olivier Stasse, Andrew J. Davison, Ramzi Sellaouti, and Kazuhito Yokoi. Real-time 3d slam for a humanoid robot considering pattern generator information. In *Proc. of the 2006 IEEE/RSJ Int. Conf. on Intelligent Robots and Systems (IROS)*, 2006.

[24] Adrian Stern, Inna Kruchakov, Eitan Yoavi, and Norman S. Kopeika. Recognition of motion-blurred images by use of the method of moments. *Applied Optics*, 41(11):2164–2171, 2002.

[25] Y. Yitzhaky, I. Mor, A. Lantzman, and N. S. Kopeika. Direct method for restoration of motion-blurred images. *Journal of the Optical Society of America A*, 15:1512–1519, June 1998.

[26] S. Zickler and M.M. Veloso. Detection and localization of multiple objects. In *Proc. of the 2006 IEEE-RAS International Conference on Humanoid Robots*, 2006.