

Robust Visual Odometry for Complex Urban Environments

Ignacio Parra, Miguel Ángel Sotelo, Ljubo Vlacic

Abstract—This paper describes a new approach for estimating the vehicle motion trajectory in complex urban environments by means of visual odometry. A new strategy for robust feature extraction and data post-processing is developed and tested on-road. Scale-invariant Image Features (SIFT) are used in order to cope with the complexity of urban environments. The obtained results are discussed and compared to previous works. In the prototype system, the ego-motion of the vehicle is computed using a stereo-vision system mounted next to the rear view mirror of the car. Feature points are matched between pairs of frames and linked into 3D trajectories. The distance between estimations is dynamically adapted based on re-projection and estimation errors. Vehicle motion is estimated using the non-linear, photogrammetric approach based on RANSAC (RANdom SAMple Consensus). The obvious application of the method is to provide on-board driver assistance in navigation tasks, or to provide a means of autonomously navigating a vehicle. The method has been tested in real traffic conditions without using prior knowledge about the scene or the vehicle motion. An example of how to estimate a vehicle's trajectory is provided along with suggestions for possible further improvement of the proposed odometry algorithm.

I. INTRODUCTION

Accurate estimation of the vehicle global position is a key issue, not only for developing useful driver assistance systems, but also for achieving autonomous driving. Using stereo-vision for computing the position of obstacles or estimating road lane markers is a popular technique in intelligent vehicle applications. The challenge now is to extend stereo-vision capabilities to also provide accurate estimation of the vehicle's ego-motion with respect to the road, and thus to compute its global position. This is becoming more and more tractable to implement on standard PC-based systems.

In this paper, a new approach for ego-motion computing based on stereo-vision is proposed. The use of stereo-vision has the advantage of disambiguating the 3D position of detected features in the scene at a given frame. Based on that, feature points are matched between pairs of frames and linked into 3D trajectories. The idea of estimating displacements from two 3-D frames using stereo vision has been previously used in [1] [2] and [3]. A common feature of these studies is the use of robust estimation and outliers rejection using RANSAC (RANdom SAMple Consensus)[4]. In [2] a so-called firewall mechanism is implemented in order to reset the system to remove cumulative error. Both monocular and stereo-based versions of visual odometry were developed in [2], although the monocular version needs

I. Parra and M.A. Sotelo are with the Department of Electronics, Escuela Politécnica Superior, University of Alcalá, Alcalá de Henares, Madrid, Spain. parra, sotelo@depeca.uah.es

L. Vlacic is with the Intelligent Control Systems Laboratory (ICSL), Griffith University, Brisbane, Australia l.vlacic@griffith.edu.au

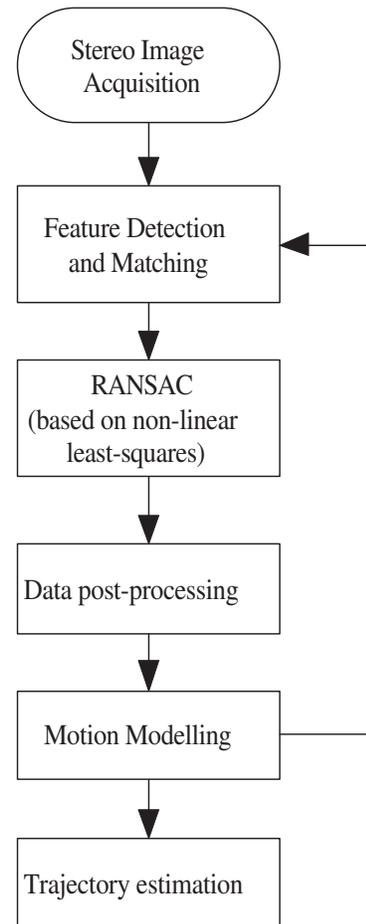


Fig. 1. General layout of the visual odometry method based on RANSAC.

additional improvements to run in real time, and the stereo version is limited to a frame rate of 13 images per second. In [5] a stereo system composed of two wide Field of View cameras was installed on a mobile robot together with a GPS receiver and classical encoders. The system was tested in outdoor scenarios on different runs of up to 150 meters each. In [6], trajectory estimation is carried out using visual cues for the sake of autonomously driving a car in inner-city conditions.

In the present work, the solution of the non-linear system equations describing the vehicle motion at each frame is computed under the non-linear, photogrammetric approach using RANSAC. The use of RANSAC [2] allows for outliers

rejection in 2D images corresponding to real traffic scenes, providing a method for carrying out visual odometry on-board a road vehicle.

The rest of the paper is organized as follows: in section II the new feature detection and matching technique is presented; section III provides a description of the proposed non-linear method for estimating the vehicle's ego-motion and the 3D vehicle trajectory; implementation and results are provided in section IV; finally, section V is devoted to conclusions and discussion on how to improve the current system performance in the future.

II. FEATURES DETECTION AND MATCHING

In most previous research on visual odometry, features are used for establishing correspondences between consecutive frames in a video sequence. Some of the most common choices are Harris corner detector [7] and the Kanadi-Lucas-Tomasi detector (KLT)[8].

Harris corners have been found to yield detections that are relatively stable under small to moderate image distortions [9]. As stated in [2], distortions between consecutive frames can be regarded as fairly small when using video input. However, Harris corners are not always the best choice for landmark matching when the environment is cluttered and repetitive superimposed objects appear on the images. This is the situation for urban visual odometry systems. Although Harris corners can yield distinctive features, they are not always the best candidates for stereo and temporal matching. Among the wide spectrum of matching techniques that can be used to solve the correspondence problem, the *Zero Mean Normalized Cross Correlation* [10] is more frequently used thanks to its robustness.

In order to minimize the number of outliers, a mutual consistency check is usually employed (as described in [2]). Accordingly, only pairs of features that yield mutual matching are accepted as a valid match. The accepted matches are

used both in 3D feature detection (based on stereo images) and in feature tracking (between consecutive frames).

In urban cluttered environments repetitive patterns such as zebra crossings, building windows, fences, etc. can be found. In Fig. 2 the typical correlation response along the epipolar line for a repetitive pattern is shown. Multiple maxima or even higher responses for badly matched points are frequent. Although some of these correlation mistakes can be detected using techniques such as the mutual consistency check or the unique maximum criterion, the input data for the ego-motion estimation will be regularly corrupted by these outliers which will decrease the accuracy of the estimation.

Moreover, superimposed objects limit observed from different viewpoints are a source of correlation errors for the system. In Fig. 3 we can see a typical example of an urban environment in which a car's bonnet is superimposed on the image of the next car's license plate and bumper. As can be seen in Fig. 3(a), the Harris corner extractor chooses, as feature points, the conjuncture in the image between the car's bonnet and the next car's license plate and bumper. In the plane image these are, apparently, good features to track, but the different depths of the superimposed objects will cause a misdetection due to the different viewpoints. In Fig. 3(b) and 3(c) it can be seen how the conjuncture in the image between the number 1 on the license plate and the bonnet is matched but they do not correspond to the same point in the 3D space. We can see the same kind of misdetection in the conjuncture between the car's bonnet and the bumper. The error in the 3D reconstruction of these points is not big enough to be rejected by the RANSAC algorithm so they will corrupt the final solution.

In practice, these errors lead to local minima in the solution space and thus to inaccurate and unstable estimations. A more reliable matching technique is needed in order to cope with the complexity of the urban environments.

In this system we apply a similar approach to [11], in which scale-invariant image features are used for Simultaneous Localization And Map Building (SLAMB) in unmodified (no artificial landmarks) dynamic environments. To do so they use a trinocular stereo system [12] to estimate the 3D position of the landmarks and to build a 3D map where the robot can be localized simultaneously. Our approach uses a calibrated stereo rig mounted next to the rear view mirror of a car to compute the ego-motion of the vehicle.

In our system, at each frame, SIFT features are extracted from each of the four images (stereo pair at time 1 and stereo pair at time 2), and stereo matched among the stereo pairs (Fig. 4). The resulting matches for the stereo pairs are then, matched again among them. Only the features finding a matching pair in the three matching processes will be used for the computation of the ego-motion.

SIFT (Scale Invariant Feature Transform) was developed by Lowe [13] for image feature generation in object recognition applications. The features are invariant to image translation, scaling, rotation, and partially invariant to illumination changes and affine or 3D projection. These characteristics make them good feature points for robust visual odometry

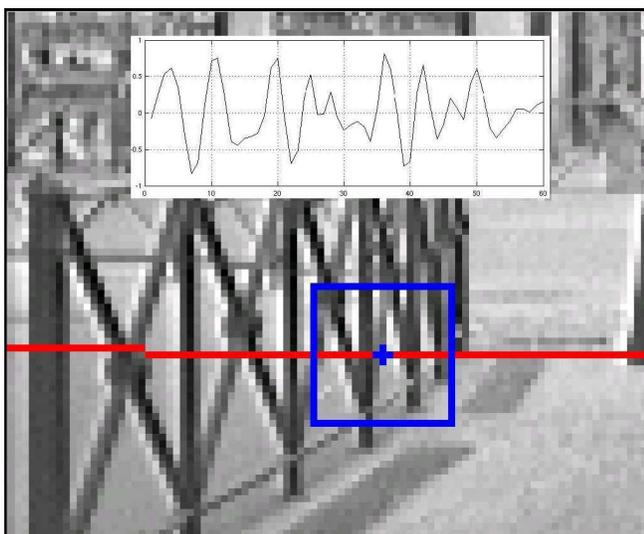
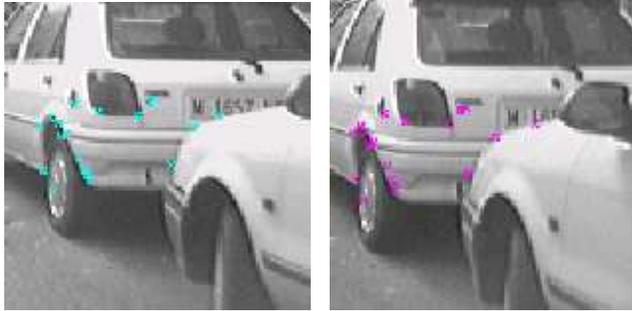


Fig. 2. Correlation response along the epipolar line for a repetitive pattern



(a) Left image at time 1 Harris points. (b) Right image at time 1. Matched Harris points



(c) Left image at time 2. Harris points matched with Harris points from Left image at time 1. Outliers from Left image at time 2. in orange

Fig. 3. Examples of matches for superimposed objects

systems, since when mobile vehicles are moving around in an environment, landmarks are observed over time, but from different angles and distances.

As described in [14] the best matching candidate for a SIFT feature is its nearest neighbour, defined as the feature with the minimum Euclidean distance between descriptor vectors.

The large number of features generated from images, as well as the high dimensionality of their descriptors, make an exhaustive search for closest matches inefficient. Therefore the Best-Bin-First (BBF) algorithm based on a k-d tree search [15] is used. This can give speedup by factor of 1000 while finding the nearest neighbor (of interest) 95% of the time.

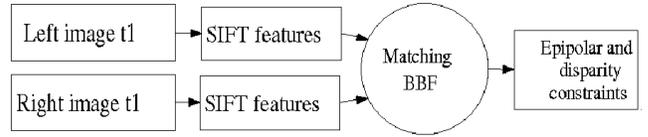
As the SIFT best candidate search is not based on epipolar geometry, the reliability of matches can be improved by applying an epipolar geometry constraint to remove remaining outliers. This is a great advantage with respect to other techniques which rely on epipolar geometry for the best candidate search. For each selected image pair this constraint can be expressed as:

$$x_l^T \cdot F \cdot x_r = 0 \quad (1)$$

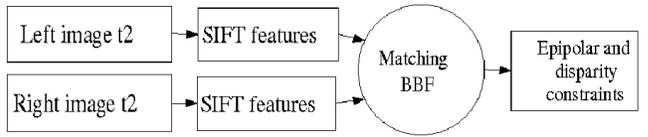
where F is the Fundamental matrix previously computed in an off-line calibration process and x_l^T , x_r are respectively the homogeneous image coordinates of the matched features in image *left* transposed and the homogeneous image coordinates of the matched features in image *right*. Also matches are only allowed between two disparity limits. Sub-pixel

Sift temporal and stereo matching process

Stereo matching at time 1



Stereo Matching at time 2



Temporal matching

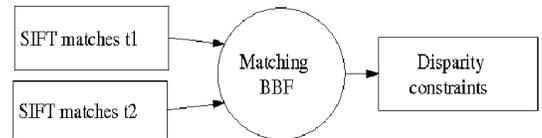


Fig. 4. Diagram of the features extraction method for the proposed system

horizontal disparity is obtained for each match. This will improve the 3D reconstruction accuracy and therefore the ego-motion estimation accuracy.

The resulting stereo matches between the first two stereo images are then similarly matched with the stereo matches in the next stereo pair. No epipolar geometry constraint is applied at this step and an extra vertical disparity constraint is used. If a feature has more than one match satisfying these criteria, it is ambiguous and discarded so that the resulting matching is more consistent and reliable.

From the positions of the matches and knowing the cameras' parameters, we can compute the 3D world coordinates (X, Y, Z) relative to the left camera for each feature in this final set.

Relaxing some of the constraints above does not necessarily increase the number of final matches (matches in the two stereo pairs and in time) because some SIFT features will then have multiple potential matches and therefore be discarded.

From the 3D coordinates of a SIFT landmark and the visual odometry estimation, we can compute the expected 3D relative position and hence the expected image coordinates and disparity in the new view. This information is used to search for the appropriate SIFT feature match within a region in the next frame.

Once the matches are obtained, the ego-motion is determined by finding the camera movement that would bring each projected SIFT landmark into the best alignment with its matching observed feature.

III. VISUAL ODOMETRY USING NON-LINEAR ESTIMATION

The problem of estimating the trajectory followed by a moving vehicle can be defined as that of determining at frame i the rotation matrix $R_{i-1,i}$ and the translational vector $T_{i-1,i}$ that characterize the relative vehicle movement between two consecutive frames. For this purpose a RANSAC based

on non linear least-squares method was developed for a previous visual odometry system. A complete description of this method can be found on [16]. An overview is given in sections III, III-A and III-B (also see Fig. 1).

The estimation of the rotation angles must be undertaken by using an iterative, least squares-based algorithm [4] that yields the solution of the non-linear equations system that must be solved in this motion estimation application. Otherwise, the linear approach can lead to a non-realistic solution where the rotation matrix is not orthonormal.

A. RANSAC

RANSAC (Random Sample Consensus) [17] [18] is an alternative to modifying the generative model to have heavier tails to search the collection of data points S for good points that reject points containing large errors, namely “outliers”.

RANSAC is used in this work to estimate the Rotation Matrix R and the translational vector T that characterize the relative movement of a vehicle between two consecutive frames. The input data to the algorithm are the 3D coordinates of the selected points at times t and $t + 1$.

B. 2D Approximation

Under the assumption that only 2D representations of the global trajectory are needed, like in a bird’s-eye view, the system can be dramatically simplified by considering that the vehicle can only turn around the y axis (strictly true for planar roads). It implies that angles θ_x and θ_z are set to 0, being θ_y estimated at each iteration.

A non-linear equation with four unknown variables $\mathbf{w} = [\theta_y, t_x, t_y, t_z]^T$ is obtained where $T = [t_x, t_y, t_z]$ is the translational vector.

After an iterative process using all the points obtained from the matching step the algorithm yields the final solution $\mathbf{w} = [\theta_y, t_x, t_y, t_z]^T$ that describes the relative vehicle movement between two consecutive iterations.

C. Data Post-processing

This is the last stage of the algorithm. In most previous research on visual odometry, features are used for establishing correspondences between consecutive frames in a video sequence. However it is a good idea to skip the frames yielding physically incorrect estimations or with a high mean square error to get more accurate estimations.

We have found there to be two main sources of errors in the estimation step:

- 1) Solutions for small movements (5 centimeters or less) where the distance between features is also small (one or two pixels), are prone to yield inaccurate solutions due to the discretized resolution of the 3D reconstruction (Fig. 5(b)).
- 2) Solutions for images where the features are in the background of the image (Fig. 5(a)) are inaccurate for the same reason as before: 3D reconstruction resolution decreases as long as depth increases. Although the features extraction algorithm sorts the features depending

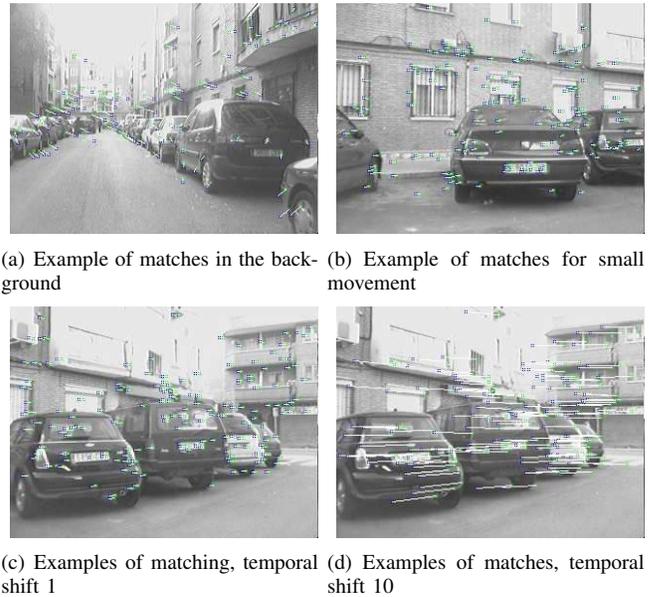


Fig. 5. Examples of SIFT matches. In green SIFT feature at time t_1 in blue matched feature at time t_2 , in white the movement of the feature.

on its depth and it uses the closest ones, at some frames it is not able to find enough features close to the car.

SIFT features have proven to be robust to pose and illumination changes, so they are good candidates for matching, even if there are some skipped frames between the matching stereo pairs and thus, the appearance of the features has changed (Fig. 5(d)). Also the fact that they don’t rely on the epipolar geometry for the matching process makes its computational time independent to the disparity between features. Using a correlation based matching process it would be necessary to increase the disparity limits in order to find the features which will probably be further away from each other.

According to this some ego-motion estimations are discarded using the following criteria.

- 1) High root mean square error e estimations are discarded.
- 2) Meaningless rotation angles estimations (non physically feasible) are discarded.

A maximum value of e has been set to 0.5. Similarly, a maximum rotation angle threshold is used to discard meaningless rotation estimations. In such cases, the ego-motion is computed again using frames t_i and $t(i + 1 + shift)$ where $shift$ is an integer which increases by one at every iteration. This process is repeated until an estimation meets the criteria explained above or the maximum temporal shift between frames is reached. The maximum temporal shift has been fixed to 5 so as the spatial distance between estimations remains small and thus the estimated trajectory is accurate. Using this maximum temporal shift the maximum spatial distance between estimations will be around 0.5-2.5m. If the system is not able to get a good estimation after 5 iterations the estimated vehicle motion is maintained according to

motion estimated in the previous correct frame assuming that the actual movement of the vehicle can not change abruptly.

The system is working at a video frame rate of 30fps which allows us to skip some frames without losing precision in the trajectory estimation.

IV. IMPLEMENTATION AND RESULTS

The visual odometry system described in this paper has been implemented on a Core II Duo at 2.16 GHz running Kubuntu GNU/Linux 6.1 with a 2.6.20-16 SMP kernel version. The algorithm is programmed in C using OpenCV libraries (version 0.9.9). A stereo vision platform based on Fire-i cameras (IEEE1394) was installed on a prototype vehicle. After calibrating the stereo vision system, several sequences were recorded in different locations including Alcalá de Henares and Arganda del Rey in Madrid (Spain). The stereo sequences were recorded using a non-compression algorithm at 30 frames/s with a resolution of 320x240 pixels. All sequences correspond to real traffic conditions in urban environments with pedestrians and other cars in the scene. In the experiments, the vehicle was driven below the maximum allowed velocity in cities, i.e., 50 Km/h.

A. 2D Visual Odometry Results

The results of a first experiment are depicted in Fig. 6. The vehicle starts on a trajectory in which it first turns slightly to the left. Then, the vehicle runs along a straight street and, finally, it turns right at a strong curve with some 90 degrees of variation in yaw. The upper part of Fig. 6 shows an aerial view of the area of the city (Alcalá de Henares) where the experiment was conducted (source: <http://maps.google.com>). The bottom part of the figure illustrates the 2D trajectory estimated by the visual odometry algorithm presented in this paper (no marker) and the previous version of the system using Harris corners and ZMNCC (triangles) [16].

As can be observed, the system provides reliable estimations of the path run by the vehicle in all the sections. As a matter of fact, the estimated length run in Fig. 6 is 147.37m, which is very similar to the ground truth (165.86m). Compared to the previous system the trajectory is more accurate and closer to the actual length of the run. Taking into account that 13.84% of the frames were discarded in the post-processing step, the actual length of the run is quite close to the real one.

In a second experiment, the car starts turning left and then runs along an almost straight path for a while. After that, a sharp right turn is executed. Then the vehicle moves straight for some meters and turns slightly right until the end of the street. Fig. 7 illustrates the real trajectory described by the vehicle (above) and the trajectory estimated by the visual odometry algorithm (below). The estimated trajectory reflects the exact shape of the real trajectory executed by the vehicle quite well. The system estimated a distance of 197.89m in a real run of 216.33m. Similarly to the first experiment 9.51% of the estimations were discarded by the post-processing step, thus the actual length of the run is again very close to the real one.

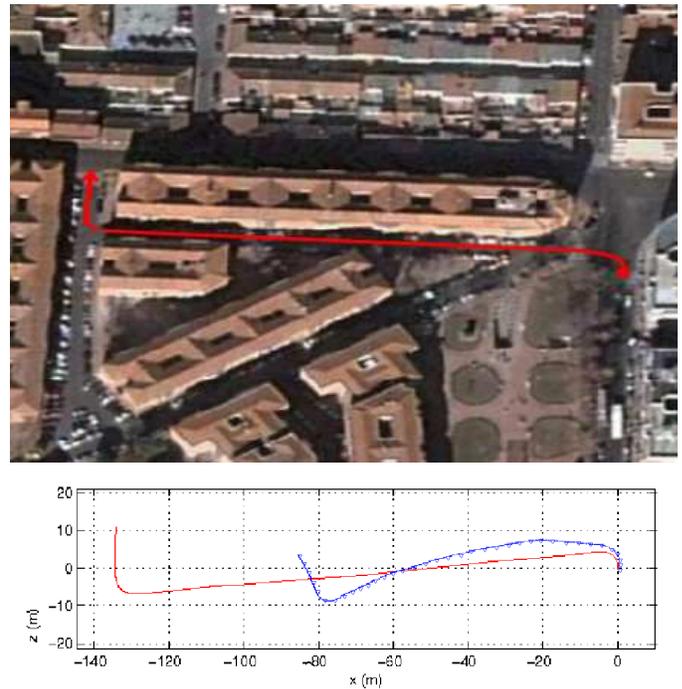


Fig. 6. Above, trajectory in the city for experiment 1. Below estimated trajectory for the previous Harris feature extractor (triangles) and for the new SIFT strategy (no markers)

V. CONCLUSIONS AND FUTURE WORK

We have described a method for improving the estimation of a vehicle's trajectory in a network of roads by means of visual odometry. To do so, SIFT feature points are extracted and matched along pairs of frames and linked into 3D trajectories. The resolution of the equations of the system at each frame is carried out under the non-linear, photogrammetric approach using least squares and RANSAC. This iterative technique enables the formulation of a robust method that can ignore large numbers of outliers as encountered in real traffic scenes. Fine grain outliers rejection methods have been experimented with, based on the root mean square error of the estimation and the vehicle dynamics. An adaptive temporal shift which tries to avoid bad estimations has also been developed. The resulting method is defined as visual odometry and can be used in conjunction with other sensors, such as GPS, to produce accurate estimates of the vehicle global position.

Real experiments have been conducted in urban environments in real traffic conditions with no prior knowledge of the vehicle movement or the environment structure. We provide examples of estimated vehicle trajectories using the proposed method. Although preliminary, the first results are encouraging since it has been demonstrated that the system is capable of providing approximate vehicle motion estimation.

As part of our future work we envision the development of a method for discriminating stationary points from those which are moving in the scene. Moving points can correspond to pedestrians or other vehicles circulating in the

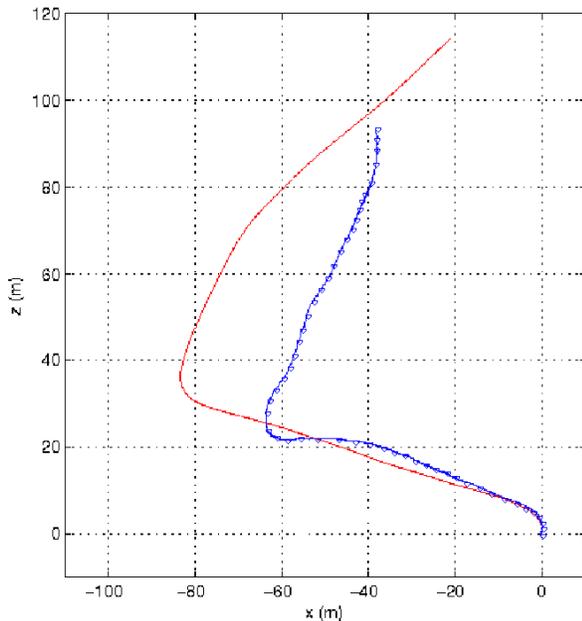


Fig. 7. Above, trajectory in the city for experiment 2. Below estimated trajectory for the previous Harris feature extractor (triangles) and for the new SIFT strategy (no markers)

same area. Vehicle motion estimation will mainly rely on stationary points. The system can benefit from other vision-based applications currently under development and refinement in our lab, such as pedestrian detection [19] and ACC (based on vehicle detection). The output of these systems can guide the search for stationary points in the 3D scene. Also a tracking of the features has to be addressed using the information of the movement estimations and a kalman filter which will estimate the feature's next position. This information will be used to determine a region of interest for the feature extraction algorithm and also to compute the features' probability of being stationary points. This will

allow to better deal with pedestrians, cars and other moving objects in the scene. This probability will be used for the resolution of the system using a weighted non-linear least squares method in which every point in the system will be weighted by its probability of being a stationary point.

The obvious application of the method is to provide on-board driver assistance in navigation tasks, or to provide a means for autonomously navigating a vehicle. For this purpose, fusion of GPS and vision data will be accomplished.

VI. ACKNOWLEDGMENTS

This work has been supported by the Spanish Ministry of Education and Science by means of Research Grant DPI2005-07980-C03-02 and the Regional Government of Madrid by means of Research Grant CCG06-UAH/DPI-0411.

REFERENCES

- [1] Z. Zhang and O. D. Faugeras, "Estimation of displacements from two 3-d frames obtained from stereo," in *IEEE Transactions on Pattern Analysis and Machine Intelligence*. Vol. 14, No. 12, December, 1992.
- [2] D. Nister, O. Naroditsky, and J. Beren, "Visual odometry," in *Proc. IEEE Conference on CVPR*. June, 2004.
- [3] A. Hagnelius, "Visual odometry," in *Masters Thesis in Computing Science*. Umea University, April, 2005.
- [4] D. A. Forsyth and J. Ponce, *Computer Vision. A Modern Approach*, international ed. Pearson Education International. Prentice Hall, 2003.
- [5] M. Agrawal and K. Konolige, "Real-time localization in outdoor environments using stereo vision and inexpensive gps," in *In 18th ICPR06*. pp. 1063-1068, 2006.
- [6] N. Simond and M. Parent, "Free space in front of an autonomous guided vehicle in inner-city conditions," in *In European Computer Aided Systems Theory Conference (Eurocast 2007)*. pp. 362-363, 2007.
- [7] C. Harris and M. Stephens, "A combined corner and edge detector," in *Proc. Fourth Alvey Vision Conference*. pp. 147-151, 1988.
- [8] B. Lucas and T. Kanade, "An iterative image registration technique with an application to stereo vision," in *Proceedings of the International Joint Conference on Artificial Intelligence*. pages 674-679, 1981.
- [9] C. Schmid, R. Mohr, and C. Bauckhage, "Evaluation of interest point detectors," in *IJCV*. Vol. 37, No. 2, pp. 151-172, 2000.
- [10] B. Boufama, "Reconstruction tridimensionnelle en vision par ordinateur: Cas des cameras non etalonnees," in *PhD thesis*. INP de Grenoble, France, 1994.
- [11] S. Se, D. Lowe, and J. Little, "Vision-based mobile robot localization and mapping using scale-invariant features," in *Proceedings of the IEEE ICRA*. pages 2051-2058, 2001.
- [12] D. Murray and J. Little, "Using real-time stereo vision for mobile robot navigation," in *Proceedings of the IEEE Workshop on Perception for Mobile Agents*, 1998.
- [13] D. G. Lowe, "Object recognition from local scale-invariant features," in *Proceedings of the Seventh ICCV*. pages 1150-1157, 1999.
- [14] I. Gordon and D. G. Lowe, "What and where: 3d object recognition with accurate pose," in *International Symposium on Mixed and Augmented Reality*, 2006.
- [15] J. S. Beis and D. G. Lowe, "Shape indexing using approximate nearest-neighbour search in high-dimensional spaces," in *Proceedings of the IEEE Conference on CVPR*. pages 1000-1006, 1997.
- [16] R. García, M. A. Sotelo, I. Parra, D. Fernández, and M. Gavilán, "3d visual odometry for gps navigation assistance," in *Proceedings of the IEEE Intelligent Vehicles Symposium*. pages 444-449, 2007.
- [17] M. A. Fischler and R. C. Bolles, "Random sample consensus: A paradigm for model fitting with applications to image analysis and automated cartography," in *Communications of the ACM*. June, 1981.
- [18] R. Hartley and A. Zisserman, *Multiple View Geometry in Computer Vision*. Cambridge University Press, 2004.
- [19] I. Parra, D. Fernández, M. A. Sotelo, L. M. Bergasa, P. Revenga, J. Nuevo, M. Ocana, and M. A. García, "A combination of feature extraction methods for svm pedestrian detection," in *IEEE Transactions on Intelligent Transportation Systems*. Vol. 8, No. 2, June, 2007.