Markerless Outdoor Localisation Based on SIFT Descriptors for Mobile Applications

Frank Lorenz Wendt, Stéphane Bres, Bruno Tellez, and Robert Laurini

LIRIS UMR 5205 CNRS, INSA-Lyon, F-69621, France
{lorenz.wendt, stephane.bres, bruno.tellez,
robert.laurini}@liris.cnrs.fr
http://liris.cnrs.fr/

Abstract. This study proposes augmented reality from mobile devices based on SIFT (Scale Invariant Feature Transform) features for markerless outdoor augmented reality application. The proposed application is navigation help in a city. These SIFT features are projected on a digital model of the building façades of the square to obtain 3D co-ordinates for each feature point. The algorithms implemented calculate the camera pose for frame of a video from 3D-2D point correspondences between features extracted in the current video frame and points in the reference dataset. The algorithms were successfully tested on video films of city squares. Although they do not operate in real-time, they are capable of a correct pose estimation and projection of artificial data into the scene. In case of a loss of track, the algorithms recover automatically. The study shows the potential of SIFT features for purely image based markerless outdoor augmented reality applications. This study takes place in the MoSAIC[1] project.

Keywords: Content-based image retrieval, image matching, augmented reality, SIFT, building recognition, pose estimation.

1 Introduction

The Internet has become the most important tool for information access and distribution. However, up to now, Internet access was restricted to stationary computers based in an office or at home, and linked to the Internet via cable. Existing options to access the Internet from a mobile client have been either restricted to selected places with a WLAN hotspot, or have provided only limited transfer rates via a mobile telephone network, or have been too costly to gain a significant market acceptance. In future, this is changing. A new network infrastructure is being built up, that grants mobile Internet access at transfer speeds comparable to home-based solutions. Mobile telephone manufacturers are developing optimized mobile handsets with more processing power and bigger displays, and Internet search engines offer services optimized for a mobile use. Moreover, nowadays standard equipment of any mobile device includes camera, GPS, and probably more interesting things in the future, which can be used as

[1] MoSAIC: Mobile Search and Annotation using Images in Context, ICT ASIA project (Ministère des Affaires Etrangères - MAE, France), 2006-2008.

A. Elmoataz et al. (Eds.): ICISP 2008, LNCS 5099, pp. 439 – 446, 2008.

new input medium to formulate search engine queries instead of using the more tedious textual input.

This study proposes a mobile, markerless augmented reality as a solution for a convenient and intuitive way to launch Internet queries with little or no need to enter a query as text, and to display query results in a simple and clear manner. The user would simply point with his camera at the building or object he is interested in (a restaurant, a theatre, a statue…), and the system would query a database or the Internet about the object of choice, and display the results in the live video of the object the user is just filming. Concepts for multimodal input options as proposed for example by Lim et al. [7] where the user takes a photo of the object of interest with his camera-equipped telephone are naturally included within this framework.Obviously, such a tool would require a combination of solutions to work properly – this study focuses only on the augmented reality solution itself, identifying the object in the view of the camera, and tracking it through the sequence of video images.

This system would be equally suited for navigation, tourism, gaming or advertisement.

Section 2 will give an overview to solutions to this problem proposed in literature. Section 3 cover our actual research and section 4 describe the results we have through a prototype of an augmented reality. Section 5 concludes the paper and raises further research questions.

2 Some Previous Works

In this study, an outdoor use of a hand-held augmented reality is proposed. This setting has some distinctive difficulties to overcome compared to an indoor application in a controlled environment. These difficulties can be summarized as follows : in an urban outdoor scene, abundant moving objects like cars and people can disturb the tracking process. The result is a camera movement that follows the cars, and not, as intended, the buildings. Likewise may a scene be partially occluded by objects that have not been at that position when the scene was modelled. The algorithm should nevertheless be able to recover the camera position from the remaining information. Plants are a further difficulty: They change their appearance over time, and can therefore not be used as “landmarks”. The fourth error source is lighting: in an outdoor application, light-ing can not be controlled. Therefore, the visual appearance of objects can change considerably, making tracking difficult. Shadows may also produce edges that distract tracking.

2.1 Proposed Solutions

Several studies support or replace the visual tracking by additional sensors like gyroscopes, accelerometers, GPS modules and digital compasses to overcome the aforementioned difficulties. Reitmayr and Drummond [10] propose combination of inertial sensors and vision-based point and edge trackers for a handheld outdoor augmented reality. In their approach, inertial sensors provide a first estimation of the camera pose. A textured building model is then rendered according to the estimated camera pose, and an edge tracking algorithm determines the exact pose by matching edges in the video image with edges in the textured model. Ribo et al. [11] and Jiang et al. [5]

likewise use hybrid tracking systems of sensors and visual tracking techniques for robust outdoor augmented reality applications. For this study, focus was set on purely image based techniques only. Ferrari et al [2] propose tracking of parallelogram shaped or elliptic planar features, de-scribed in an invariant way. Their work shows interesting results, but lack generality, as suitable planes are never present in a high number, making this concept prone to occlusions, or could be missing completely in a given scene. For a virtual reconstruction of antique Greek monuments shown in an AR system, Vlahakis et al [14] use a high number of keyframes to keep differences between the current frame and the most similar keyframe low. This allows the use of simple and fast matching techniques, but effectively restricts the movement of the user to a few selected standpoints, as not every possible perspective could be anticipated and taken as key-frame in advance.

Gordon and Lowe [3] use Lowe's SIFT detector and descriptor [8] for an augmented reality application. SIFT stands for Scale Invariant Feature Transform, and is an invariant point detector and descriptor. In a first step, Gordon and Lowe take a set of reference images to create a sparse representation of the object or scene to be recognized and tracked. SIFT points are then extracted from each image, and matched against each other. During the augmented reality application, SIFT points are extracted in each frame and matched against the points of the point cloud. This establishes 2D-3D correspondences, which are used to calculate the camera position and pose. This approach has two drawbacks: The extraction of SIFT features is computationally demanding, which restricted the frame rate to 4 frames per second. Secondly, as the camera pose and position are calculated for each frame individually, the resulting virtual overlay jitters against the real world background. The method was documented for a small indoor scene and a very restricted outdoor scene comprising only one building front. Its performance on a larger scale is therefore unknown.

Vacchetti et al. [13] combine relative orientation between frames and absolute orientation towards a very low number of keyframes to reduce both drift and jitter. They choose the Harris interest point detector (Harris and Stephens, 1988 [4]) and image patches as their descriptor to match both between subsequent video frames and between video frames and keyframes.

3 Our Proposition

Our work was inspired by both the work of Gordon and Lowe [3] and of Vacchetti et al. [13]. For its simplicity, a similar approach as in Gordon and Lowe was chosen, that uses SIFT keypoints as means to establish correspondences between 3D and 2D points. Other invariant point descriptors exist, which are computationally lighter, including PCA-SIFT (Principal com-ponent analysis SIFT) [6] , SURF (Speeded up robust features) [1], and GLOH (Gradient location-orientation histogram) [9], some of which are faster in computation. As the SIFT detector has proven its superior matching performance in a comparative study of Mikolajczyk and Schmid, [9], it was also chosen for this study. For more references on SIFT interest points and associated signature, see [8].

The aim of this study was to set up a prototype for an outdoor markerless augmented reality, based on the SIFT keypoint detector and descriptor. The application

outlined in section 1 of this paper requires a real-time operation on a mobile device. This was beyond the scope of this study. Instead, an implementation on a standard PC was realized that does not operate at real-time speed. It does however respect the requirements of a live application in that sense that only for each frame, only the information of previous frames or offline information was used.

The work for this study comprises 3 components: the calibration of the mobile device camera ; a 3D façade model of a small urban area and a 3D point cloud of SIFT points as the reference data ; a matching of video frames to the reference data and deduction of the camera location.

3.1 Camera Calibration

The camera of the mobile device has to be calibrated to know precisely its characteristics. The procedure is very simple. The user has to photograph a chessboard like calibration pattern from a number of different viewpoints. He then marks the outermost corners of the calibration pattern manually, and the calibration procedure automatically finds the different parameters of the camera, like focal length and lenses distortions.

3.2 3D Reference Model

To create the reference set of SIFT points, images of the building façades where taken. Each image was oriented absolutely with respect to the 3D model by manually providing control points. From these 3D -2D correspondences, the camera position and pose was calculated. In the next step, SIFT points were extracted from the image, and the rays from the centre of projection through the SIFT features in the image were constructed. Finally, the world coordinates of the points were obtained by intersecting the rays with the building façades. At early stages of the project, a manual region of interest was defined for each image to make sure that only points correctly originating from the façades are projected on the façade. The example on figure 1 presents the projection of SIFT points on the model.

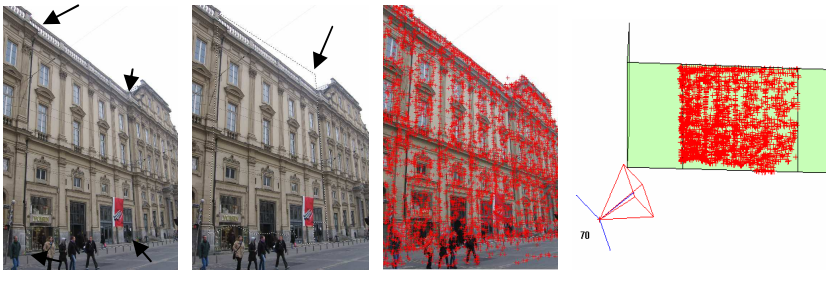

Fig. 1. Production of the reference set of SIFT points: Manual input of control points (arrows, left) – Manual definition of region-of-interest (center left) – extraction of SIFT features (center right) – All points within the region of interest are projected on the surface of the 3D model (right)

3.3 Matching of Video Frames to the Reference Data

During this study, four markerless augmented reality algorithms based on Lowe`s SIFT [8] point detector and descriptor were implemented. All four algorithms share the same principle: Point correspondences between features in the current frame and a point cloud of reference points are used to calculate the pose of the camera in a robust fashion. The reference set was produced by projecting SIFT point features from a series of reference images onto the surface of a simplified 3D building model (see figure 2).

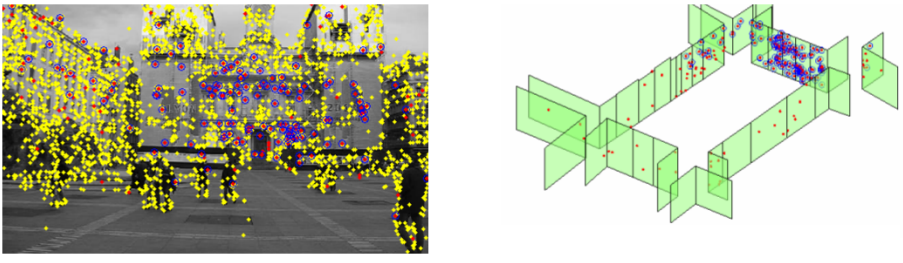

Fig. 2. Extraction of SIFT points (left) and matching to the 3D model (right). Only the points that match in sufficient number on the same area are kept. The other ones are considered as outliers.

The pose was calculated by minimizing the reprojection error of world points into the image in a least-squares sense within a RANSAC loop. This algorithm requires starting values for the camera's position and attitude. In this algorithm, a planar homography between four image points and four coplanar world points is used to initiate the pose estimation. This method has shown to deliver correct results reliably, provided the scene contains planar objects, a justified assumption for a building recognition application. The aforementioned methods to derive an initial pose were only used for the first frame. In the remaining video images, the pose of the previous frame was taken as the starting value, provided it was based on at least 15 point matches. This threshold was chosen arbitrarily and not tested in detail.

The camera pose was calculated in RANSAC loop. This algorithm is controlled by two values, namely the inlier threshold which refers to the maximum deviation a world point projected into the image plane may show from the true image point position to be regarded as an inlier point, and the number of iterations or camera poses that are calculated and tested. Extensive testing has shown that the shatter of the calculated position decreases with increasing iteration numbers, but remains almost constant if the number of iterations exceeds 250. Therefore, the value of 250 iterations is proposed to be used. Alternatively, an assumed ratio of correct points over all matched points of 25 % may be used.

Another method is implemented to reduce jitter. In this algorithm, a second set of points is detected in each image using the FAST detector [12]. These points are projected onto the 3D model. In the next frame, these points are found again by matching between the neighbouring frames using an image patch as a simple descriptor. The resulting additional 3D-2D correspondences are then used to increase the number of

point matches. This method effectively reduces jitter of the virtual image content with respect to the real objects, as shown by standard deviations of the calculated camera positions dropping to only a few decimetres in some sequences of the test video. Although the FAST detector, the projection and the matching method used here are simple calculations, they make only sense if they replace the costly SIFT calculation and matching, instead of being used additionally to SIFT.

4 Some Results

In the videos produced with the settings described above, the main reason for a complete loss of orientation was motion blur (see figure 3) and even in that case the algorithm recovers as soon as clear images are available again.

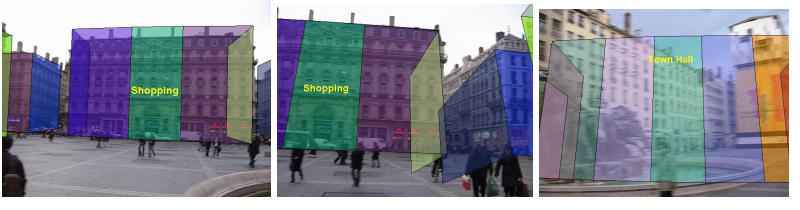

Fig. 3. Some images extracted from a computed augmented reality video. On these samples, we superimpose the 3D model to the real images to judge the accuracy of the matching. Correct matching (left) Wrong point of view estimation (center) Wrong matching due to motion blur (right).

Although in many of the frames produced for this study the virtual content jitters strongly, the algorithm has always recognized the buildings in the image correctly. This shows that the SIFT point detector and descriptor has the potential to be used for markerless augmented reality, provided its calculation can be accelerated. Errors in the reference dataset have been the most important reason for a bad fit of virtual content and filmed buildings, and not missmatches of the SIFT algorithm itself. This holds even for the uniform building façades that contain a lot of repeating structures. Although the reference images had a size of only 300 by 400 pixels, an average of 134 matches were found. The low ratio of inliers among these matches of approximately 25% is probably also caused by the low quality of the reference dataset. The 3D models used here contained important errors like wrong building heights and each reference image was oriented individually, mostly with only four control points. A better approach to produce the reference dataset would have been to use a more detailed 3D model, and to make a bundle adjustment over all input images to produce a consistent dataset

5 Conclusion and Further Researches

While this research has shown that SIFT features are well suited for an augmented reality application under the given conditions, lower performance due to blur in the

frames coming from fast displacement of the camera, or complete different point of view or zoom on details, can be compensated by a bigger set of reference images, or by altering the reference images synthetically. This raises the question: How must the reference set be made ? The number of reference images depends on the scale of the objects to be contained in the reference set, which again depends on the application of the augmented reality application. If information about buildings is to be displayed, it is sufficient to cover the buildings with images as done in this study. However, if smaller objects like, for example, building details are to be detected, the number of needed reference images increases. An increased number of reference images has unfavourable consequences: The effort to produce the reference set is increased, the size of the reference set gets bigger, which makes storage and transfer of it more difficult, and the matching process takes more time when the search domain increases.

Similar question applies for the lighting conditions. Is this dataset still sufficient in dawn or at night time? Once the requirements on the reference set are known in more detail, automatic methods to derive the reference dataset would be of great advantage.

Finally, further investigations are necessary to transform the augmented reality algorithms into an intuitive tool that helps the user to fulfil his information demand in a simple and easy-to-use fashion.

References

[1] Bay, H., Tuytelaars, T., Van Gool, L.: SURF: Speeded Up Robust Features. In: Proceedings of the ninth European Conference on Computer Vision, May, pp. 404–417 (2006)

[2] Ferrari, V., Tuytelaars, T., Van Gool, L.: Markerless augmented reality with a real-time affine region tracker. In: Proceedings of the IEEE and ACM International Symposium on Augmented Reality, pp. 87–96 (2001)

[3] Gordon, I., Lowe, D.: Scene modelling, recognition and tracking with invariant image features. In: Proc. 3rd IEEE and ACM International Symposium on Mixed and Augmented Reality (ISMAR 2004), Arlington, pp. 110–119 (2004)

[4] Harris, C., Stephens, M.: A combined corner and edge detector. In: Proc. of The Fourth Alvey Vision Conference, 1988, Manchester, UK, pp. 147–151 (1998)

[5] Jiang, B., Neumann, U., You, S.: A Robust Hybrid Tracking System for Outdoor Augmented Reality. In: Proc. of the IEEE Virtual Reality Conference 2004 (VR 2004), pp. 3–10 (2004)

[6] Ke, Y., Sukthankar, R.: PCA-SIFT: A more distinctive representation for local image descriptors. In: Proc. of the IEEE Conf. on Computer Vision and Pattern Recognition (CVPR), vol. 2, pp. 506–513 (2004)

[7] Lim, J.-H., Chevallet, J.-P., Merah, S.N.: SnapToTell: Ubiquitous Information Access from Camera. In: A Picture-Driven Tourist Information Directory Service in Mobile & Ubiquitous Information Access (MUIA 2004) Workshop as part of the conference Mobile Human Computer Interaction with Mobile Devices and Services (Mobile HCI 2004), Glasgow University of Strathclyde, Scotland, September 2004, pp. 21–27 (2004)

[8] Lowe, D.: Distinctive image features from scale-invariant keypoints. International Journal of Computer Vision 60, 91–100 (2004)

[9] Mikolajczyk, K., Schmid, C.A.: Performance evaluation of local descriptors. In: Proc. of IEEE Computer Vision and Pattern Recognition, vol. 2, pp. 257–263 (2003)

[10] Reitmayr, G., Drummond, T.: Going Out: Robust Model-based Tracking for Outdoor Augmented Reality. In: Proc. IEEE ISMAR 2006, Santa Barbara, USA, pp. 109–118 (2006)
[11] Ribo, M., Lang, P., Ganster, H., Brandner, M., Stock, C., Pinz, A.: Hybrid tracking for outdoor augmented reality applications. IEEE Comp. Graph. Appl. 22(6), 54–63 (2002)
[12] Rosten, E., Drummond, T.: Fusing points and lines for high performance tracking. In: Proc. 10th IEEE International Conference on Computer Vision (ICCV 2005), Beijing, vol. 2, pp. 1508–1515 (2005)
[13] Vacchetti, L., Lepetit, V., Fua, P.: Combining edge and texture information for real-time accurate 3D camera tracking. In: Proc. 3rd IEEE and ACM International Symposium on Mixed and Augmented Reality (ISMAR 2004), Arlington, VA, pp. 48–57 (2004)
[14] Vlahakis, V., Ioannidis, N., Karigiannis, J., Tsotros, M., Gounaris, M., Stricker, D., Gleue, T., Daehne, P., Almeida, L.: Archeoguide: An Augmented Reality Guide for Archaeological Site. IEEE Computer Graphics and Applications 22(5), 52–60 (2002)