Estimating Scale using Depth From Focus for Mobile Augmented Reality

Klen Čopič Pucihar
School of Computing and Communications
InfoLab21, Lancaster University
Lancaster LA1 4WA UK
+44 1524 510393
k.copicpucihar@lancaster.as.uk

Paul Coulton
School of Computing and Communications
InfoLab21, Lancaster University
Lancaster LA1 4WA UK
+44 1524 510393
p.coulton@lancaster.ac.uk

ABSTRACT
Whilst there has been considerable progress in augmented reality over recent years it has principally been related to either marker based or apriori mapped systems which limits its opportunity for wide scale deployment. Recent advances in marker-less systems that have no apriori information using techniques borrowed from robotic vision are now finding their way into mobile augmented reality and are producing exciting results. However, unlike marker based and apriori tracking systems these techniques are independent of scale which is a vital component in ensuring that augmented objects are contextually sensitive to the environment they are projected upon. In this paper we address the problem of scale by adapting a Depth From Focus (DFF) technique, which has previously been limited to high-end cameras to a commercial mobile phone. The results clearly show that the technique is viable and with the ever-improving quality of camera phone optics, add considerably to the enhancement of mobile augmented reality solutions. Further as it simple require a platfrom with an auto-focusing camera the solution is applicable to other AR platforms.

Keywords
Mobile, scale, metric scale, camera, phone, augmented reality.

INTRODUCTION
One of the main challenges of Augmented Reality (AR) systems is camera tracking, which can be implemented using fiducial markers or natural-features. In fiducial based systems the scale ambiguity is not present as it can be easily derived by using markers of a known size, whereas in the natural feature based systems it is only possible if the system is of informed type where the apriori knowledge of the view being studied is available i.e. where a database of landmarks forming the map is created offline and the map creation process introduces metric scale. In the case of natural feature tracking where the 3D map is created online from natural features alone, the scale is unknown because it is impossible to determine the scale of the scene based on a sequence of images alone [5].

In fiducial marker and apriori feature tracking systems scale ambiguity is not a problem although such systems offer limited prospects of large scale deployment as they would require either wide scale augmentation of our physical space with fiducial markers or wide scale 3D mapping of our physical environment. The alternative options are maker-less AR systems that use online tracking approaches without apriori information the method of map creation and camera pose estimation can vary from the model-based to move-matching approaches.

In the case of online model-based approach the camera pose is always estimated by comparing the initial frame with the current camera frame. The initial frame is an image taken directly from above the plane, or one that is synthetically un-projected from additional sensor information, by which perspective distortions of camera projections are removed and the extracted landmarks can be used as an object model of the plane in the scene. As the same initial fame is always taken for pose estimation, such system is not incremental and does not have problems with drift or loop closures [1]. However, such systems are limited to planar scenes, as landmarks not lying on a plane cannot be initialized from only one observation thus making extraction of the depth information using stereo vision impossible. Furthermore, as the initial frame is always used for the camera pose estimation, all newly added features need to be referenced to the initial frame, which in practice means long term maps where features are tracked over a long period of time.

AR systems that use this approach, but differ in the sense that their maps are created offline have been created [12] [17], however, there is no reason why such systems could not be modified to act as uninformed tracking systems which would improve their use flexibility, but at the same time introduce the scale ambiguity. One such system running on a mobile phone is Nestor [4] in which curves of planar shapes are used for tracking and shape identification. The shapes are added to the database of known shapes by the method described above, and are then used as natural features for camera pose extraction as well

as to select 3D objects for augmentation. The drawback of this system is that it is also ambiguous up to scale.

In the later case of move-matching techniques, the camera pose is updated based on the frame-to-frame movement of tracked features. Such system [14] is incremental as after each frame is acquired the camera pose update from the previous frame is computed. This approach does not require long-term feature tracking and with it the requirement to maintain long-term maps, which makes it more flexible and faster as computationally expensive bundle adjustment of the map is not required. Furthermore it also enables depth estimation and with it the camera pose extraction from non-planar surfaces. However, as the method is incremental, it is hard to avoid drift, which also introduces the problem of loop closures [1]. In case of non-planar surfaces, one of the main problems is the initialization of the system where the camera pose and the map environment are unknown. To solve this problem, the Simultaneous Localization and Mapping (SLAM) technique were developed in the field of robotic exploration and were later adopted by AR systems.

Two such SLAM algorithms, EKF-SLAM [15] and FastSLAM2.0 [10] both use incremental mapping methods and were later adapted for hand held cameras [1] [2]. In these systems, the map is initialized by a fiducial marker through which the scale of the map becomes available. However in case of Eade and Drummond SLAM implementation [2] the map can also be initialized without the necessity of a marker but in this case the scale again becomes unknown. Note after the map initialization, natural features are used for expanding the map and tracking the camera pose.

Further developments of single hand held camera tracking were achieved using the Parallel Tracking and Mapping algorithm (PTAM) [6], which differentiates from others by separating the mapping and tracking tasks. In PTAM bundle adjustment is used as an alternative to incremental mapping in which long-term maps are created and features are frequently revisited. The map initialization is done with five-point stereo algorithm or in the later versions by homography decompositions. In both of these cases the metric scale is unknown if no additional information is available.

Currently the only presented alternative for estimating scale is performed during the process of stereo map initialization as demonstrated in PTAM [6] whereby users were asked to provide first two keyframes of the map by moving the camera sideways for approximately 10 cm during from which the metric scale of the map could be estimated as additional information was introduced to the captured video stream. However, according to Klein and Murray, this map initialisation method proved to be problematic as users tended to use pure rotation rather then lateral movement, thus the correct map initialization was heavily dependent on users understanding of the stereo baseline requirements [7]. Furthermore, introduction of scale in this manner is subjective as the user camera movement is approximate and subjectively assessed.

To date a highly modified variation of PTAM for the iPhone is the only implementation of six degrees of freedom camera tracking SLAM on a mobile phone where, according to Klein and Murray, stereo initialization was determined inadequate not only because of the introduction of the user error previously defined, but due to the limitations of the mobile phone platform, in particular the limited computational power and narrow camera field of view [7]. In the alternative map initialization, Klein and Murray, ask the user to only provide the first key-frame, therefore, the previously defined additional information is lost. This means that currently there are no marker-less mobile AR systems that provide an estimate of scale.

In this paper we introduce a possible solution for providing metric scale for marker-less AR systems with no apriori information by utilizing the Depth From Focus (DFF) technique. In the following section the theoretical background of the method will be presented followed by the design patterns section where two different generic scale implementations will be discussed. The solutions are then analyzed through an empirical study of a specific implemented on a commercially available mobile phone the Nokia N900. Note that the proposed solution is platform, as well as operating system, independent and could therefore be implemented on any auto-focusing system where access to the camera driver is available. Finally the implementation of a demo application will be presented followed by the conclusions and further work.

THEORETICAL FRAMEWORK

Digital cameras are generally auto-focused by searching for the lens position that gives the 'best' focused image, thus the lens position is dependent on the distance to the object as shown in Figure 1. If the focused lens position and the focal distance of the lens are known, the thin Gaussian lens equation (1) can be used to calculate DFF i.e. the distance to the object u.

$$1/f = 1/u + 1/v \quad (1)$$

This method has been mainly used in the domain of robotic vision as an alternative to stereo depth recovery. One of the main choices with this method is which focus measure to use in order to identify the best lens position [18]. An ideal focus measure is described as unimodal and monotonic in that it should have only one maximum at the point where the image is in focus [11] [16]. However, in practice any focal measure has many local maximums, therefore, the global peak of the focal measure is not easy to find. Furthermore, as it has been observed by [11] and [18], not only the texture and contrast of the scene, but also the depth of field (DOF) influence the maximums of focus measure function. It is preferable to have good texture with high contrast as well as the smallest possible DOF, which can be achieved by using the maximal focal distance of the camera as well as maximal aperture. Furthermore, with bigger focal distance the lens movement for focusing the image is

bigger which is expected to increases the resolution and precision of the lens positioning system.

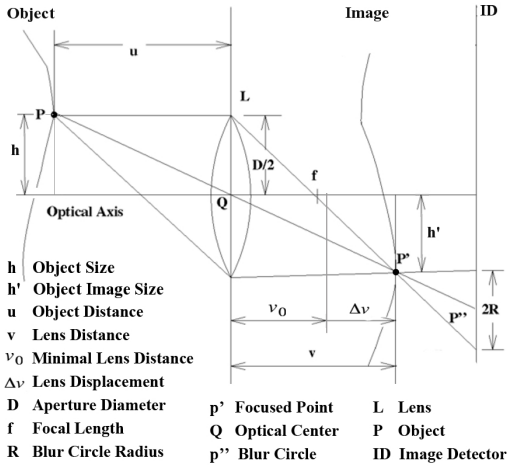

Figure 1: Image formation in a convex lens

In order to calculate the scale unit *s* of the scene, one needs to know the distance to the object plane *u*, the vertical or horizontal cameras field of view α and picture height or width in pixels *2h'*. The calculation of scale unit *s* is then based on simple trigonometry as shown in equation (2). In case of the augmented reality application, the user would need to focus on the plane where at least two map points are present. After defining the scale between two map points the scale of the whole map is known.

$$h = u \tan(\alpha/2) \quad \Rightarrow \quad s = h[mm]/h'[px] \qquad (2)$$

The measurable depth of DFF system is theoretical limited by the hyperfocal distance, which is a defined as a minimal object distance at which we need to focus in order to consider the points at infinity to be in focus. However, as DOF needs to be as small as possible in order to achieve reliable and accurate results, such distance lies much closer. The relation between the size of DOF and hyperfocal distance is inversely proportional with the ratio between object and hyperfocal distance, therefore the accuracy of the system is expected to be better in the close up range. In case of an AR application with the ability to expand the map, the scale estimation only imposes the limitation for close up system initialization, which can then expand the map to desirable proportions. Therefore, the size of the AR workspace is not limited by the requirement of a close-up initialization.

From the previous discussion it is obvious that the scale could be introduced to uninformed marker-less AR systems if the system has a camera with auto-focusing capability and allows access to the camera driver. In the following section the generic implementation for scale estimation will be designed.

DESIGN PATTERNS FOR GENERIC IMPLMENTATION

The motor count captured from the camera driver is assumed to represent the relative distance of the lens in the motor step domain. In order to use motor count with the Gaussian lens equation (1), the conversion to absolute distance in metric space (on Figure 1 shown as *v*) is required. An alternative is to capture measurements across the whole focusing range and define an approximation function that will define the transformation from motor count to object distance. This research analyzes both cases as it has some significant implications for the user interaction requirements as well as the flexibility of the system.

In the first mode of operation, the camera system is assumed to be unknown. In order to convert the motor count to object distance, the lens equation (1) can be used, however, as already indicated, the lens movement interval is usually unknown. The only information available about the lens position is its motor count, which needs to be converted to lens distance *v* in metric space.

The proposed solution is to focus the camera at the object, at two different, but known distances. The first measurement should be taken close to the minimal focusable object distance, and the second at approximately one sixth of the hyperfocal distance. The range is performed in the close up region of the camera as the accuracy of DFF system is expected to decline with object distance and it is important to ensure that the calibration of the motor step values is made in a way to best fit the lens equation in the close up region. As the distances to the object are known, the theoretical lens position can be calculated using the equation (1) by which the motor step *unit* is defined. The difference between the current motor count value *m* and the minimal motor count value m_0 make the conversion of motor step count to metric space possible by equation (3).

$$v = v_0 + \Delta m \cdot unit \qquad \Delta m = m - m_0 \qquad (3)$$

In the second mode of operation, the assumption of a known camera model is made, which enables the use of an approximation function to transform the motor count step to object distance. No user calibration is therefore required for this mode of operation, however, this would limit its applicability to known camera models for which the approximation curve has been predefined. In the next section, we will make a case study of the scale estimation accuracy using standard camera phone Nokia N900 running Maemo operating system. It is important to note, that the proposed solution is not limited to the specific operating system nor the device. In the following section the proposed solutions will be analyzed through an empirical study of a specific implemented on mobile phone the Nokia N900.

EMPERICAL STUDY OF SCALE ESTIMATION

The data presented in this section was captured with four phones where auto-focusing was performed by two focusing algorithms, namely, the native camera application algorithm and by utilising the 'gstreamer' library. The phone camera used is a 5-mega pixels camera with Carl. Zeiss optics with a focal length of 5.2 millimetres, aperture f/2.8 and a horizontal field of view of 56 degrees.

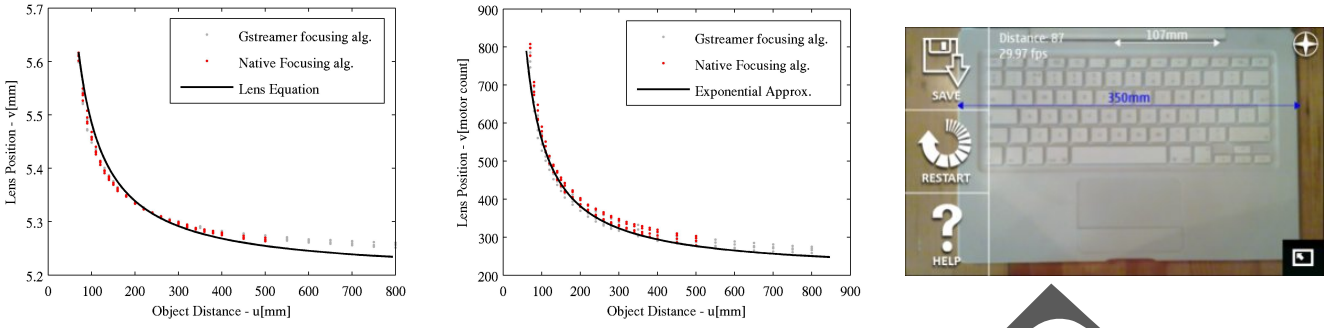

Figure 2: The left graph shows the average lens position in relation to object distance in mode one operation. The graph in the middle shows the average lens position in motor step space in relation to object distance for mode two operation and exponential approximation function (3). To the right, the screen shot of the Metre application measuring the laptop is shown.

In all measurements the same randomly selected A3 colour poster with good contrast and texture was used. The measurements were repeated 20 times at each given distance. In case of 'gstreamer' library, the phone was focused at an object at random distance before capturing each measurement. In the case of native application the measurements were taken in a sequence starting at minimal object distance. Using the data analysis of the 'gstreamer' dataset, a decision was made to only capture values in the range between 70 and 500 millimetres in the next experiment, as results above this distance were considered to be to unreliable. The motor count value was assumed to be represented by the 'V4L2_CID_ FOCUS_ABSOLUTE' variable of the Video4Linux2 camera driver and was captured after each successful focusing.

DFF Accuracy Using Gaussian Lens Equation (Mode 1)

In order to analyze how well the Gaussian lens equation (1) fits the captured data, the average values of measured lens displacements of each phone were plotted alongside the theoretical values obtained using the lens equation function and are shown in the left graph of . The camera was calibrated at object distances of 70 and 250 millimeters. Although the shape of the theoretical curve runs relatively close to the captured data set it is still considerably different. As expected, the dataset is best described close to the value of v_0 which is the lens position of the far point used in the calibration procedure of mode one operation. Furthermore, it can be observed that the two focusing algorithms produce similar results and that the deviation of results between four different phones is small. The results prove that the assumptions made are correct and that the captured value from the camera driver is interpreted correctly.

The accuracy of depth measurement can be best described with the relative depth error (shown on in the left graph of Figure 3), which is also the relative error of the scale introduced to the AR system because the only variable in scale calculation of equation (2) is the object distance. The maximal relative depth error at distances below 300 millimetres ranges from 9.5 up to 15.8 percent, which is compared to results acquired with precise camera systems (0.098% acquired at 1.2 meters) still very high [18]. However, it should be taken into account that the focal length, lens mechanics and quality of such high precision camera systems limit the direct comparability to the mobile phone camera. These limitations could be also seen as one of the reasons for deviation of the dataset from the theoretical lens equation. Furthermore, as discussed in the method section, some of these parameters have significant effect on the focus measure that is a crucial component of the auto-focusing accuracy.

DFF Accuracy Using Approximation (Mode 2)

In order to improve the accuracy of the system and to remove need for user calibration, the mode two solution proposed the use of approximation function for mapping the transformation from motor count space to object distance. The exponential approximation curve (3) was determined from the average data set of all measurements taken by the 'gstreamer' focusing algorithm. To make the function best represent the data at a close range, only measurements up to 400 millimetres were considered.

$$v(u) = a/u + b \qquad (3)$$

The centre graph of shows that this new curve better represents the measurements, especially in the close up region. The maximal relative depth error can be seen in the centre graph of Figure 3 and shows that the maximal relative depth error does not drop but stays at comparable levels to mode one operation, however it is obvious, that the approximation function describes the data far more consistently as in the case of lens equation because the relative depth error graph does not show the distinctive minima at the calibration value v_0 that can be observed in mode one operation. It can also be seen that the standard deviation of results between different phones seams to have grown compared to mode one operation. This is due to the fact that motor count values have not been normalized as in the case of mode one operation.

To explore the full potential of the system, a decision was made to analyze behaviour of the system where a simple one step calibration was added to the mode two operation. As it is easier for the user to calibrate the device at the close up region, the user is asked to calibrate their phone close to the minimal focusable distance, in our case at 70 millimetres. The exponential approximation curve needs to

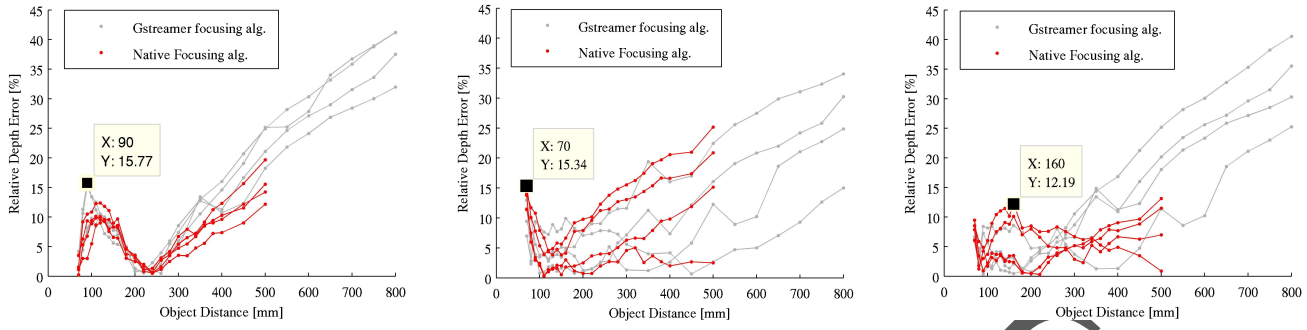

Figure 3: The graphs in this section all show Relative Depth Error of object distance. In the graph to the left mode one operation results are shown, followed by the mode two results and finally on the right the results for a single step user calibration of mode two are presented. Each line of the graph represents one of the 4 phones., and the two colours denote different focusing algorithms.

be recalculated for the measurements which are normalized at the maximum motor count captured by user calibration. Again only the measurements captured with 'gstreamer' focusing algorithm in the region up to 400 millimetres were used. In operation each time the new motor count value is captured, it is subtracted from the user calibrated maximal motor count value and converted to distance using the new fitting function. By doing this the maximal relative depth error shown at the middle of Figure 3 has dropped to a range between 6.5 to 12.2 percent.

If the two different focusing algorithms are compared, their performance does not differ significantly in the close up region, however this is not the case in the regions further away from the camera. The accuracy of the native camera focusing algorithm in distant regions is better then what would be expected. The possible reason for this could be the already mentioned difference in the procedure how the two experiments were executed. Contrary to the 'gstreamer' measurement, the native camera was not refocused at a randomly distant object before capturing each measurement, but was rather capturing the measurements in a sequence.

It is important to note that the captured measurements were taken under the controlled environment in good lighting conditions with good focusing surface, with no user factor error, therefore, the accuracy in real world scenario could be expected to decrease. Furthermore, as incremental SLAM techniques continuously update the map and camera pose with increments, the overall scale could be affected by accruing the local scale errors [3], however, this would not be the case in SLAM approaches where batch methods are used to maintain long term maps. Furthermore, accruing of scale error would also not be present in the marker-less object-based tracking systems as those systems are not incremental and drift is not a problem.

To sum up, this data analysis shows that the proposed solution is valid and can produce reasonable scale estimation with relative error raging from 6.5 to 12.2 percent in the region between 70 and 300 millimetres. In the following section, a demo application called Metre will be discussed in order to demonstrate a use case of the proposed solution.

APPLICATION SPECIFIC IMPLEMENTATION

To highlight the technique rather than content a simpler application providing scale to a captured picture was developed. The demo application is called Metre and enables users to measure objects on a taken picture. The application was implemented on Nokia N900 phone where the tracking part of the application was implemented using the OpenCV library, the video capturing and auto-focusing were implementation using "gstreamer" library and the scale was initialized by the solution of mode two operation.

In order to increase the maximal size of the objects and to improve accuracy of the measurement, the application enables users to introduce scale close to the object they want to measure and then move back to get the full view of the object. In the scale initialization process, two natural features that are chosen based on Shi and Tomasi good corner definition [13], are being tracked using optical flow, which is calculated in the small window region of selected points by the Pyramid Lucas-Kanade algorithm [9]. As the distance between the two points is known from the scale initialization step the scale is known as long as the two features are successfully tracked. The screen shot of the application can be seen in the right corner of .

CONCLUSION AND FUTURE WORK

The results show that auto-focusing capability of the camera phone can be used to effectively introduce the scale estimate into the marker-less AR workspace without apriori information. However, currently the method is limited to the close up initialization (in our case distances up to 300 millimeters) as in this region the maximal relative scale error is expected to stay in the range of 6.5-12.2 percent.

The limitation in range and accuracy is mainly due to the small focal length (5.2 millimetres) of a camera phone, which results in short hyperfocal distance and therefore a small DFF range. It was discovered that the region up to 1/9 of the hyperfocal distance was to be accurate enough. It is important to note that marker-less AR systems which create 3D maps online have the potential to dynamically expand these maps. This means the requirement of a close up initialization is only necessary at the start of the mapping process after which the map can be expand to desired proportions. Furthermore, range capability

limitations are likely to be overcome by the next generation camera phones in which the focal distance is expected to raise by the introduction of optical zoom lenses.

However, it is important to identify that the ideal AR platform would use a camera with a wide field of view, which in practice means even smaller focal distances than the one used in this case study. Furthermore, most camera pose tracking systems use a camera projection model where the intrinsic parameters are assumed to be known and fixed [8]. As zooming changes the intrinsic parameters of the camera, it is not permitted. However, this problem could be overcome, by moving the zoom back to the original position after initializing for scale.

In the future, a fully featured uninformed marker-less augmented reality application with the proposed scale estimation will be implemented in order to explore the user interaction and to test applications that could take advantage of the newly added scale information. Finally, as the proposed scale estimation is device and platform independent, a more detailed feasibility study for implementing the proposed system on other suitable AR platforms should be preformed.

To conclude, the proposed method can be used to introduce scale into marker less AR systems without the requirement of apiori knowledge of the workspace, however, such scale estimation is currently limited to a small close up range. Nevertheless, as AR systems have potential to dynamically expend their maps, the close up initialization does not limit the size of their workspace. Furthermore, by introducing better lens optics, and optical zoom lenses to mobile devices, the accuracy and range will inevitably improve.

ACKNOWLEDGMENTS

The authors would like to thank Nokia for the provision of software and hardware to the Mobile Radicals research group at Lancaster University which was used in this project.

REFERENCES

1. Davison, A.J., Reid, I.D., Molton, N.D. and Stasse, O. MonoSLAM: Real-Time Single Camera SLAM. *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, *29* (6). 1052-1067.
2. Eade, E. and Drummond, T., Scalable Monocular SLAM. in *Computer Vision and Pattern Recognition, 2006 IEEE Computer Society Conference on*, (2006), 469-476.
3. Ethan, E. and Tom, D. Presentation of paper: Scalable Monocular SLAM, 2006.
4. Hagbi, N., Bergig, O., El-Sana, J. and Billinghurst, M. Shape recognition and pose estimation for mobile augmented reality *Proceedings of the 2009 8th IEEE International Symposium on Mixed and Augmented Reality*, IEEE Computer Society, 2009, 65-71.
5. Hartley, R.I. and Zisserman, A. *Multiple View Geometry in Computer Vision*. Cambridge University Press, ISBN: 0521540518, 2004.
6. Klein, G. and Murray, D. Parallel Tracking and Mapping for Small AR Workspaces *Proc. Sixth IEEE and ACM International Symposium on Mixed and Augmented Reality (ISMAR'07)*, Nara, Japan, 2007.
7. Klein, G. and Murray, D. Parallel Tracking and Mapping on a Camera Phone *Proc. Eigth IEEE and ACM International Symposium on Mixed and Augmented Reality (ISMAR'09)*, Orlando, 2009.
8. Lepetit, V. and Fua, P. Monocular model-based 3D tracking of rigid objects. *Found. Trends. Comput. Graph. Vis.*, *1* (1). 1-89.
9. Lucas, B.D. and Kanade, T., An Iterative Image Registration Technique with an Application to Stereo Vision. in *IJCAI81*, (1981), 674-679.
10. Montemerlo, M. and Thrun, S. FastSLAM 2.0: An Improved Particle Filtering Algorithm for Simultaneous Localization and Mapping that Provably Converges, Conference on Artificial Intelligence (ICCV'07), Rio de Janeiro, 2003, 1151–1156.
11. Nayar, S.K. and Nakagawa, Y., Shape from focus: an effective approach for rough surfaces. in *Robotics and Automation, 1990. Proceedings., 1990 IEEE International Conference on*, (1990), 218-225 vol.212.
12. Prince, S.J.D., Xu, K. and Cheok, A.D. Augmented reality camera tracking with homographies. *Ieee Computer Graphics and Applications*, *22* (6). 39-45.
13. Shi, J. and Tomasi, C. Good features to track *Computer Vision and Pattern Recognition, 1994. Proceedings CVPR '94., 1994 IEEE Computer Society Conference on*, 1994, 593 -600.
14. Simon, G., Fitzgibbon, A.W. and Zisserman, A. Markerless tracking using planar structures in the scene *Augmented Reality, 2000. (ISAR 2000). Proceedings. IEEE and ACM International Symposium on*, 2000, 120 -128.
15. Smith, R.C. and Cheeseman, P. On the Representation and Estimation of Spatial Uncertainty. *The International Journal of Robotics Research*, *5* (4). 56-68.
16. Subbarao, M., Choi, T. and Nikzad, A. Focusing Techniques. *Optical Engineering*, *32* (11). 2824-2836.
17. Wagner, D., Reitmayr, G., Mulloni, A., Drummond, T. and Schmalstieg, D. Pose tracking from natural features on mobile phones *ISMAR '08: Proceedings of the 7th IEEE/ACM International Symposium on Mixed and Augmented Reality*, IEEE Computer Society, Washington, DC, USA, 2008, 125-134.
18. Xiong, Y. and Shafer, S.A., Depth from focusing and defocusing. in *Computer Vision and Pattern Recognition, 1993. Proceedings CVPR '93., 1993 IEEE Computer Society Conference on*, (1993), 68-73.