GPU-assisted 3D Pose Estimation under realistic illumination

Anne Braun
Fraunhofer FIT
Department for Collaborative and Augmented Environments
53754 Sankt Augustin, Germany
anne-kathrin.braun@fit.fraunhofer.de

Stefan Müller
University of Koblenz-Landau
Institute for Computational Visualistics
56070 Koblenz, Germany
stefanm@uni-koblenz.de

ABSTRACT

This paper describes an approach which combines computer vision methods with techniques from the area of computer graphics. This method which is called analysis-by-synthesis explicitly seeks for consideration of environmental information in order to improve the resulting estimation of the 3D camera pose. In this paper, two different kinds of pose estimation will be presented. The first approach uses intensity-based methods and the second one is a feature point-based approach. The described approaches are based on a GPU-assisted rendering considering the real world illumination. These real world lighting conditions are captured using a HDR sampling technique. The results of this GPU-assisted approach are that both methods, the intensity-based as well as the feature point-based method, achieve better results in terms of a more robust and stable 3D camera pose under consideration of the real environmental information.

Keywords

Markerless Tracking, Analysis by Synthesis, Deferred Shading, 3D Pose Estimation

1. INTRODUCTION

Common computer vision approaches for markerless pose estimation typically exploit features in the current video image, like edges or point features, without further knowledge of the physical process of image generation. Computer graphics on the other side considers the process of illumination and light material interaction and generates realistic looking synthetic images. The approach described in this paper combines both research areas in order to improve the creation and detection of features for robust and accurate tracking. In other words, the objective of the approach described in this paper is to estimate the camera pose under consideration of environmental information and object properties.

This additional information will be used to support a computer vision-based technique for instance to match a virtual image against the real image. Since this approach explicitly seeks for consideration of environmental information the properties and conditions which will be covered are the 3D geometry of the object, the texture, and the environmental lighting condition which results in the surface color of the object. The approach will consider these properties in the way that the resulting tracking result will be improved. It doesn't require special infrastructure in the environment and it avoids drift. This method is called Analysis-by-synthesis. We developed and tested two different approaches: an intensity-based approach and a feature-based approach. Both methods will be described and compared.

This paper is structured as follows: in the next section we review some related work before we introduce the concepts of our approach. In section 5 we describe the realization of the several methods and continue in section 6 to present the results. We finally conclude and provide a glimpse into our future work.

2. RELATED WORK

The concept of intensity-based image registration is a widely used method in the area of medical imaging. In [Hip02] an intensity-based registration approach is described to match 3D images from a magnetic resonance angiography (MRA) to 2D x-ray angiography images. They realized their approach testing six different similarity measures, whereas the pattern intensity, gradient difference and gradient correlation performed consistently accurate and robust results. An intensity-based registration method to estimate the 3D camera pose is presented by Stricker [Stri01]. He computes a 2D transformation, which registers the current frame as a whole on a reference pattern. A Euclidian transformation between the live video image and one from a set of calibrated reference images is computed using the Fourier-Mellin Transform.

Feature-based approaches were much more explored in the past. Relevant work which made use of synthetic images can be found in [Lep03a] and [Rei06a]. Lepetit et al. describe a point-based approach, which uses 3D coordinates of registered keyframes from an underlying 3D model. These coordinates will be matched with the current camera image. Reitmayr et al. apply an edge-based approach on a textured 3D model. Another method using a textured model is presented by [Ros05a]. Their approach uses a textured surface mesh which is rendered in a virtual image. A modified block matching algorithm is applied to determine correspondences between patches of the surface mesh and points in an image. [Ble09a] presented a similar approach also based on a simplified textured CAD model of the environment. But in contrast she tracks feature points instead of edges and applies a sensor-fusion-based approach. Wuest et al. [Wue05a] use an OpenGL-based rendering for visibility testing. After each render step, the framebuffer and the depth buffer will be stored. A Canny edge-detector will be applied and on the framebuffer and the z-value of the depth buffer will be used to test the visibility.

In [Sch09a], an intensity-based and a feature-based approach is presented. But in contrast to the work described in this paper, their paper describes several matching and metric techniques instead of exploring the environmental influences.

3. ANALYSIS-BY-SYNTHESIS

The work described in this paper is based on the method which is called analysis-by-synthesis. Analysis-by-synthesis can be described as an optimized comparison of objects with a given model. A priori knowledge will be used to create this model. In case of 3D pose estimation, the a priori knowledge is the environmental information which will be used in form of computer graphics-based knowledge. Therefore, a 3D model of the target scene, typically a CAD model, is used to predict the appearances of features in the camera images, usually by projecting the model from the predicted camera pose. The analysis-by-synthesis technique has many advantages. One advantage is that the graphic card can be used to generate a rendered image from the predicted pose efficiently. Using the rendered image as reference, the correct level of detail is guaranteed. A further advantage of the analysis-by-synthesis approach in contrast to the frame-to-frame tracking is that the light conditions and occlusions of features don't disturb the results. By creating a synthetic reference image for every frame, the disadvantages like drift will be avoided.

The aim is to improve the process of 3D pose estimation with all the information the computer graphics render process can provide. The assumption for our approach was that the degree of realism of the synthetic image will influence the results. This means, the more realistic the synthetic image looks like the more accurate is the resulting pose. Therefore, beside the 3D geometry, the a priori knowledge includes also the texture of the material and the realistic lighting condition of the environment respectively.

4. GPU-BASED TRACKING

The foundation of the analysis-by-synthesis approach provides the deferred shading technique. Deferred shading is a technique, where the parameters which are required by a shader, like position, material, normal etc., are rendered to buffers and the lighting calculation will be realized as 2D post-processing using the information stored in these buffers. This shading process heavily relies on multiple render targets (MRT) to perform in real-time. These

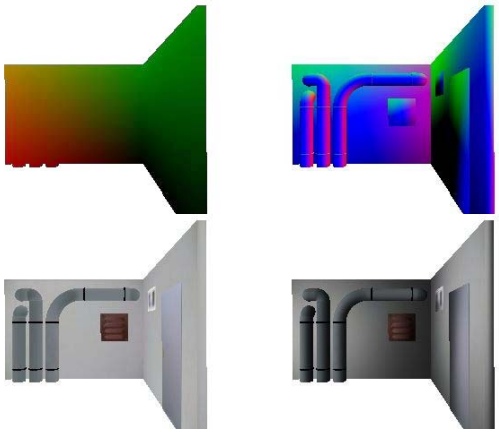

Figure 1. Multiple render targets (MRT) for Deferred Shading:vertex positions, normals, texture and final rendering

multiple render targets are called geometry buffer (G-buffer). An example of multiple render target and the final result is shown in figure 1. Initial research on deferred shading has been done by Deering et al. in 1988 [Dee88a] and Saito et al. [Sai90a] who introduced the geometric buffer (G-Buffer) in 1990.

As already mentioned, two different approaches will be described in this paper. The first approach is an intensity-based approach and the second is a feature-based method. Each of the approaches makes use of the G-buffer in a different way. The intensity-based approach applies the described deferred shading technique to create a synthetic image using the current object parameters. This identical virtual copy is used to detect the appropriate camera pose by a simple comparison of the real image and the virtual copy. Applying an optimization method, the parameters of the vector for the camera pose will converge to the solution. To determine the distance function, which is the difference between the real and synthetic image, the intensity-based approach will compare the images using the pixel intensities. The basic input data to the registration process are two images: the live video image and the rendered image. Registration is treated as an optimization problem with the goal of finding the spatial mapping that will bring the video image into alignment with the rendered image. A metric provides a measure of how well the video image is matched by the rendered image. This measure forms the quantitative criterion to be optimized by the optimizer over the search space defined by the parameters of the camera transform.

Like the intensity-based approach, the feature-based concept makes also use of the G-buffer. The features used for this approach can be either point-based, edge-based or a hybrid approach, which combines point and edge features. Running a feature detector on the video image, significant 2D point features will be obtained. These 2D point features together with the corresponding 3D point in the world will be used to calculate the 3D camera pose using a 2D-3D registration. According to the detected 2D features in the video image, the 3D information will be obtained from the G-buffer where the positions are stored at the corresponding position in the vertex buffer as color vector. Using the information of the render targets, 2D-3D correspondences of the features can be created and used for pose estimation. The analysis by synthesis approach will be applied by using the material buffer, vertex buffer and normal buffer to render a realistic lighting simulation to compare the features accordingly.

5. REALIZATION

The first step of the analysis by synthesis approach is the creation of the virtual image. This synthesis will be realized by acquiring the model data which are the geometry, the texture and the lighting information.
In this approach the geometry and texture were reconstructed manually using a laser scanner and the Photo Modeler software[1].
The lighting condition was reconstructed using a sampling method. The sampling technique doesn't require the entire image information. In contrast to environment mapping, it uses only a small amount of selected pixels. Although the number of light source will be reduced, the general amount of luminous flux will be guaranteed for several sampling densities. The sampling method is based on the K-mean clustering process. It requires a light-probe image in longitude –latitude format. Figure 2 shows an exposure series of nine images which was used to create a light probe.

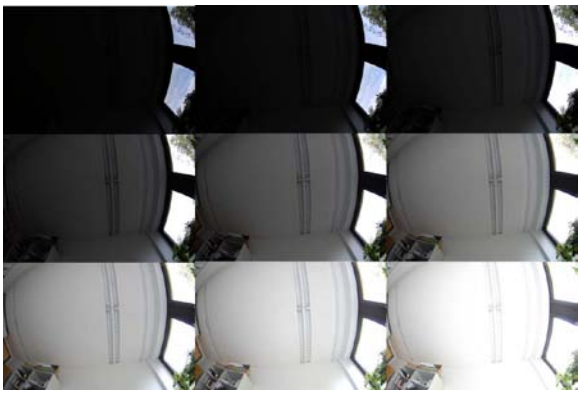

Figure 2. Nine images captured with a fish-eye lens at different exposure levels

The output is a user defined number of light sources (s. figure 3). This kind of importance sampling [Kol03a] starts with one initial light source, which is randomly chosen. The pixel in the light probe image will be partitioned into sets according to the distance to the light source. Each of the *k* light sources is moved to the center of mass of its set. The pixels are repartitioned according to the new light source direction and the process will repeated until convergence.

Using the gathered data, the virtual model of the video image will be rendered under consideration of the real environmental lighting.
The analysis-by-synthesis method can be described as an optimization problem. Figure 4 shows the general process of a registration approach based on analysis-by-synthesis. The first step is the acquisition

[1] http://www.photomodeler.com/

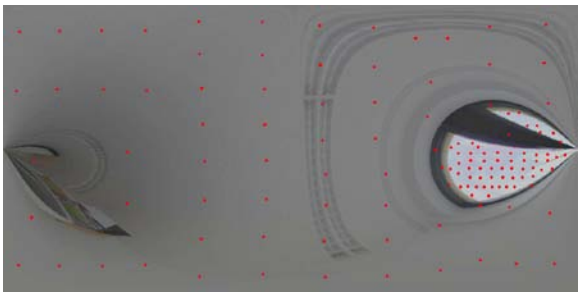

Figure 3. Light-probe image, generated of the images in figure 2, in longitude-latitude format, sampled with 128 light sources.

of image data before the similarity-features need to be decided. Similarity-features can be significant geometric primitives or pixel intensities. Furthermore, a similarity metric has to be selected. This metric describes the feature matches in the real and synthetic images. Finally, this metric will be optimized to get the best parameters. The kind of parameters depends on selected features and therefore on the different approach. Two different approaches will be distinguished in this work: Intensity-based and feature-based methods.

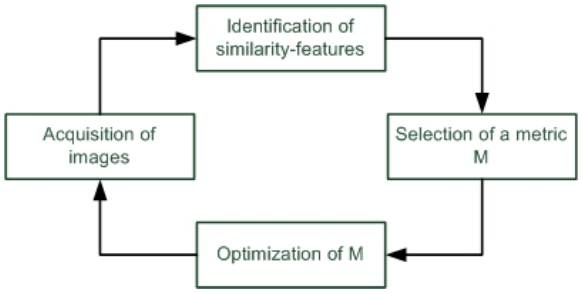

Figure 4. General process of pose estimation using an analysis-by-synthesis method

The deferred shading was realized with the OpenGL shading language GLSL. The rendering was divided in three phases: A geometry phase, a lighting phase and a post-processing phase. The G-buffer will be created in the geometry phase. The vertex shader of this phase receives the 3D model data which will be transformed to the view space and output it to the fragment shader. The fragment shader is responsible for filling the G-buffer's data. In the lighting phase, the according fragment shader uses the sampled light sources and light colors together with the data in the G-buffer to calculate the realistic illumination. The final solution texture map will be rendered to the main frame buffer in the post-processing phase.

The realization of the two computer vision-based approaches will be explained in the remainder of this section.

Feature-based approach

Applying the analysis-by-synthesis approach using features based on interest points, the technique will be used to predict the appearances of the natural interest points by rendering a 3D copy of the scene. A real-time rendered reference image has several advantages against a static feature map. By rendering the virtual image, the appearance of each feature can be adapted to illumination changes and therefore improve the registration of features in the two images which corresponds to each other. The feature point-based approach uses natural geometric primitives like points or corners which will be detected and tracked in the live image. Theses interest points are combined with a descriptor describing its appearance. Using this descriptor the features will be matched with a reference set of features or tracked in a new image. All feature point-based approaches have in common that they are invariant for rotation and translation. To be scale invariant, the features will be searched in different scale spaces.

For the work described in this paper, a SURF-based approach was used. The SURF (speed up robust features) feature method was developed by Bay et al. [Bay08a]. The SURF features are faster to detect and to match and the detector is more efficient as other feature detectors ([Bay08a] [Cheng07a]).
Starting from the initial camera pose, the features will be detected in both images. Using the G-buffer, the 3D position of the features in the rendered image can be reconstructed, since the coordinate is stored as RGB color in the vertex buffer. A detected feature in the video image will be matched using a similarity metric with all features in the rendered image. To measure the similarity the summed squared distance metric will be applied. This metric is based on the minimization of the differences of intensities in both images.

If the resulting error is below a certain threshold, the 2D feature in the video image and the 3D feature of the rendered image will set as 2D-3D correspondence.

$$M = \arg\min \sum_i \left\| m_i - f_i(M) \right\|^2$$

The pose estimation will be realized by minimizing the projection error between the 3D feature point M, projected with the projection function f_i and the 2D feature m_i in frame i.

Intensity-based approach

In contrast to the feature-based approach, the intensity-based method neither requires feature extraction nor is the search for correspondences necessary. The pixel values of two images

correspond in their coordinates and will be compared pair-wise. The realization of the analysis-by-synthesis approach using a direct intensity-based measurement was based on the Insight Segmentation and Registration Toolkit (ITK) [2]. In ITK, registration is performed within a framework of pluggable components that can easily be interchanged. The framework provides several functions to calculate the similarity metric. In case of the camera pose estimation, the object in 3D space and a 2D image have to be matched. Since the ITK framework is actually for medical imaging, it doesn't provide the full functionality for these circumstances. The classes for the metric and transformation functions are therefore derivated from the original ITK classes and reimplemented according to the requirements of 3D pose estimation. The modification in the reimplemented metric classes affects especially the *GetValueAndDerivative()* function, since this function will be called in every optimization step and within that function, all the metric values and parameters will be returned to the optimizer.

There are seven parameters to be optimized: Four parameters for the rotation (axis plus angle) and three parameters for the translation. To calculate the metric values, these parameters will be modified. The resulting metric value will be compared to the former metric value and according to the difference, the derivation of the optimizer will be estimated. After the parameters have been changed, the image needs to be re-rendered with the modified camera pose and synthetic image of the registration process needs to be renewed according the current frame buffer. This process will be repeated until the metric value is minimal. An optimizer is required to explore the parameter space of the transformation and search for optimal values of the metric. As optimizer, the *itk::RegularStepGradientDescentOptimizer* class was selected. This optimizer belongs to the class of gradient descent methods. The size of the step lengths will be reduced depending of the direction of movement in the parametric space. For the intensity-based approach two similarity measurement methods were used: Sum of Squared Differences (SSD) and Normalized Cross-Correlation (NCC). The resulting values of the normalized SSD range between [0,1], where 0 is the highest similarity. The range of values for the NCC is between [-1, 1], where 1 is the highest similarity.

The normalized SSD and NCC metric are defined with:

$$SSD(I,M) = \frac{\sum_{x,y}(I(x,y)-M(x,y))^2}{\sqrt{\sum_{x,y} I(x,y)^2}\sqrt{\sum_{x,y} M(x,y)^2}}$$

$$NCC(I,M) = \frac{\sum_{x,y}(I(x,y)-\bar{I})\cdot(M(x,y)-\bar{M})}{\sqrt{\sum_{x,y}(I(x,y)-\bar{I})^2}\sqrt{\sum_{x,y}(M(x,y)-\bar{M})^2}}$$

where $I(x,y)$ is the value of the image pixel in the video image I and $M(x,y)$ is the pixel value in the image of the synthetic model M. $\tilde{M}$ and $\tilde{I}$ are the mean value of the video and the rendered image.

6. RESULTS

In order to evaluate the performance of the presented approaches, we estimated the pose of two different test objects and compared the results. The two tested scenarios were placed indoor. Figure 4 shows the two test objects as rendered image and as real image.

The experiments were performed under a controlled condition in an office environment. For the rendered image we sampled the environmental light probe with 128 light sources

Figure 4. The two test objects: in the video image (left) and rendered (right)

For the intensity-based approach we took a small wooden model of a basement room. This room consists of two sides with mostly homogenous colored walls, a door and a window, three pipes and a ventilation box. In the experiment the pose was estimated under three different conditions. The first condition was to match a rendered image with the same rendered image as video image. This setting was used as reference for the implemented approach. The second condition was the matching of the live video image with a rendered image. However, the virtual image was not rendered under realistic light conditions. The third condition tested the approach described in this paper. A real video image was matched with a synthetic image rendered using the real light conditions. For each of the three conditions,

[2] http://www.itk.org/

the two metric approaches as described in the previous section were applied to measure the similarity.

Figure 5 shows the results for the two similarity metrics according the three test conditions. The diagrams show that the optimization converged already after four iterations. The initial step width of the optimizer for the four parameters of the rotation was 0.001 and the step width for the three parameters of the translation was 0.2. Regarding the diagrams, it's also obvious that the condition with the realistic rendered image provides better results than the synthetic reference image which was not rendered realistically. This can be seen on the high metric value of the NCC metric respectively the low metric value of the SSD metric. Comparing the SSD and the NCC metric, the NCC leads to a more precise solution.

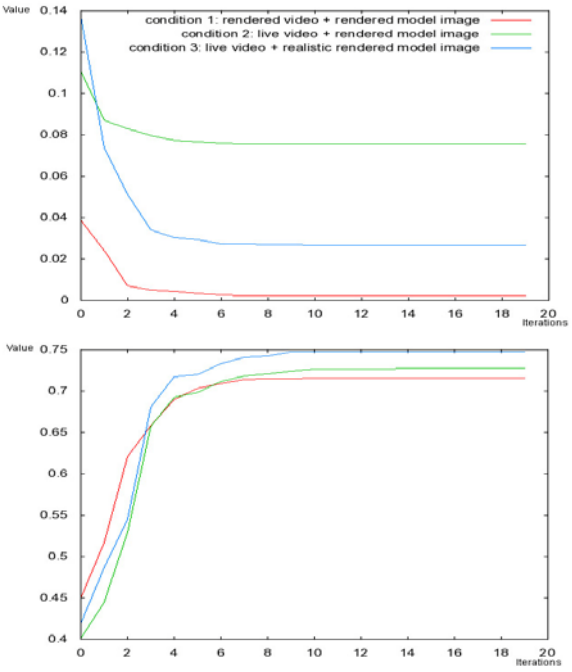

Figure 5. Results of the metric value for the two similarity metrics: The normalized NCC (bottom) and the normalized SSD (top). (according to the number of iterations for optimization).

Figure 7 shows the final results of the estimated poses under the three conditions. The first row shows the result of the first condition, where two identical rendered images were matched against each other. The resulting image and therefore the estimated pose is equal to the one of the input image. The second row presents the estimated pose of the second condition which registered a live video image with a rendered image without realistic illumination. The estimated pose was not precise and correct. Especially the z-direction of the translation was incorrect. The realistic rendering showed better results, which can be seen in the third row. Compared to the condition 1, the approach converged with very good results. The translation is very precise, and the values of the rotation are only slightly different.

To sum up the results for the intensity-based approach, the analysis-by-synthesis approach using a realistic rendered image as synthetic model provides a better estimated pose as a synthetic image which was rendered without considering the realistic illumination. Light sources which are not precise or even incorrect lead to errors in pose estimation.

For the feature-based approach, we tested another object instead of the wooden model of the basement. Since this model consists of mainly homogenous surfaces, the feature point-based approach didn't succeed due to the lack of feature points in the environment. This markerless computer vision approach using feature points is therefore not suitable to estimate the pose in rooms which are poor of features. Instead of the model we have chosen a pattern with an earth texture on it for the second test to demonstrate and test the second approach.

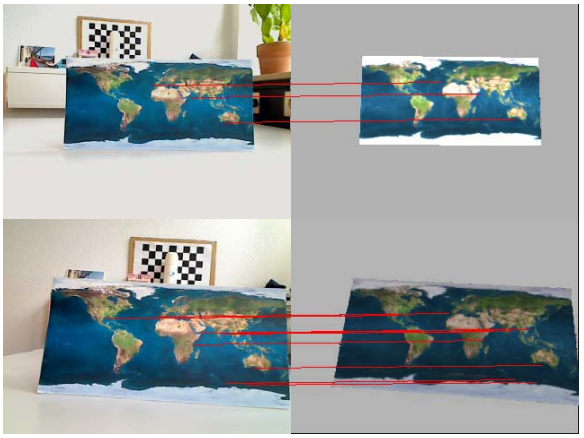

Figure 6. Results for the feature point-based approach: The scenario with the realistic lighting conditions provides more feature correspondences than the other scenario.

Testing this pattern-based scenario with the feature point-based approach, we also found out, that the more realistic the reference image is rendered the more robust and stable the pose will be estimated. The quality of the estimated pose depends on the number of correspondences, because possible outliers can be compensated by other correspondences. Furthermore, a large number of correspondence pairs reduce the risk of getting trapped in a local minimum.

We tested the earth pattern scenario under different conditions similar to the basement model scenario.

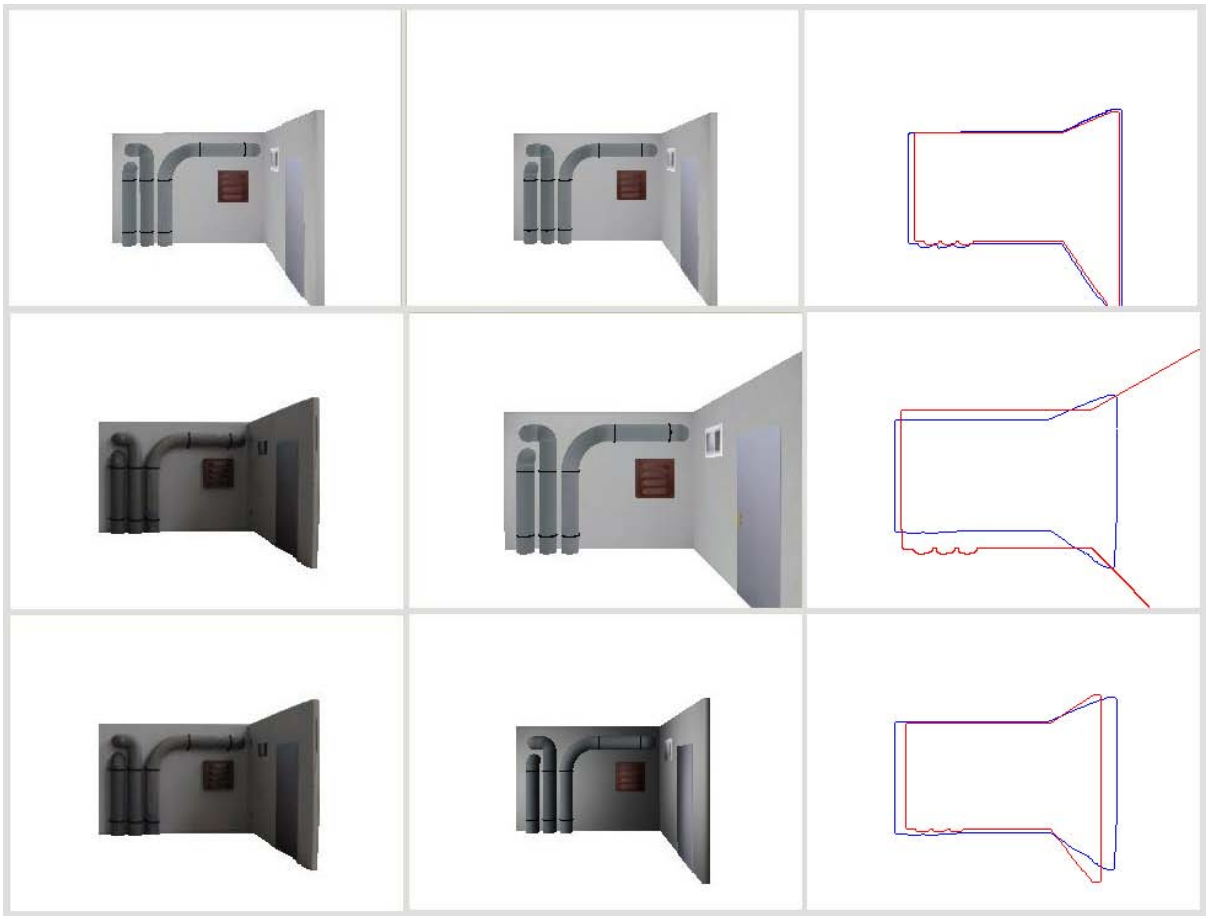

Figure 7. The final results of the estimated poses under three conditions: In the first row, two identical rendered images were matched against each other. The second row shows the estimated pose using a live video image with a rendered image without realistic illumination. The third row shows the rendered image rendered under realistic lighting conditions. The comparison in the third column shows the outlines of the input video image of the first column with blue lines and the synthetic reference image of the second column drawn with red lines.

We rendered a virtual copy of the pattern without considering the real lighting conditions for the first test. For the second test, we rendered the synthetic image using the real illumination. The result is shown in figure 6. The scenario with the realistic lighting conditions provides more feature correspondences than the other scenario.

We furthermore tested the amount of realism by comparing the number of correspondences. We therefore approach to realistic lighting conditions by increasing the number of light samples and respectively light sources. In figure 8 the results of the test is shown. An increasing number of light sources and therefore a more realistic appearance of the rendered image increase the number of feature correspondences.
The final result of the analysis-by-synthesis approach for the feature point-based method is that a more robust and stable pose will be estimated under realistic rendered light conditions.

7. CONCLUSION

In this paper we presented two approaches for 3D pose estimation using the analysis-by-synthesis approach. We therefore rendered a synthetic image of the real scenario to apply a registration method to estimate the 3D camera pose. We realized an intensity-based approach and a feature point-based approach. Both methods achieved better results using a realistic rendered image under real lighting conditions. The real lighting conditions were reconstructed based on an importance sampling approach using a HDR image of the environment captured with a fisheye lens.

The described methods require environmental data like a 3D model, the lighting and the texture to

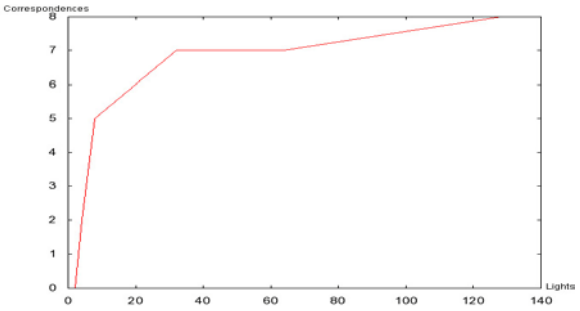

Figure 8. An increasing number of light sources and therefore a more realistic appearance of the rendered image increase the number of feature correspondences.

render a realistic virtual copy of the scenario. At first sight, this seems to be a disadvantage. But regarding the 3D model and the texture, many 3D models are already available for various situations and scenarios. Either it was reconstructed for public access for instance Google Earth® or the models exist anyway from previous design steps like in the construction industry. Furthermore, the reconstructed and sampled lighting condition can be used for improved rendering, for instance for photorealistic AR applications.

A significant contribution of the intensity-based approach is that the 3D camera pose can be estimated although the scenario doesn't provide many feature points. This advantage will overcome the problem of markerless computer vision based tracking in unconstrained environments without additional sensors. The advantage of the feature point-based approach is that errors and artifacts like drift will be avoided since this approach doesn't use a frame-to-frame tracking but a reference image.

8. FUTURE WORK

The future work will include several improvements. One task will be to replace the ITK based implementation of the intensity-based approach. Since this library is actually intended for medical registration, the modification for our approach results in slow computation times. A tailor-made solution for this problem will increase the frame rate. However, using this library enables us to try this approach and test first prototypes. Furthermore, the required data can be acquired automatically.

9. ACKNOWLEDGMENTS

We acknowledge the support from the European Commission for the research project CoSpaces under grant number FP7 IST-5-034245.

10. REFERENCES

[Bay08a] Bay Herbert, Ess Andreas, Tuytelaars Tinne, and Van Gool Luc. Speeded-up robust features (surf). Comput. Vis. Image Underst., 110(3):346–359, 2008.

[Ble09a] Bleser Gabriele, Stricker Didier. Advanced tracking through efficient image proce ssing and visual-inertial sensor fusion. Computer & Graphics, Vol. 33, Pages 59-72, Elsevier, New York, 2009.

[Cheng07a] Cheng D., Xie Shane, and Hiimmerle E.. Comparison of local descriptors for image registration of geometrically-complex 3d scenes. pages 140 –145, 2007.

[Dee88a] Deering Michael, Winner Stephanie, Bic Schediwy, Duffy Chris, and Neil Hunt. The triangle processor and normal vector shader: a vlsi system for high performance graphics. SIGGRAPH '88: Proceedings of the 15th annual conference, pages 21–30, New York, NY, USA, 1988. ACM.

[Hip02] Hipwell, J. H., Penney G. P., Cox T. C., Byrne J. V., and Hawkes D. J. 2D-3D Intensity Based Registration of DSA and MRA – A Comparison of Similarity Measures. Lecture Notes in Computer Science, 2489 (2002), pp. 501–508

[Kol03a] Kollig Thomas and Keller Alexander, editors. Efficient Illumination by High Dynamic Range Images, 2003.

[Lep03a] Lepetit V., Vacchetti L., Thalmann D., and Fua P.. Fully Automated and Stable Registration for Augmented Reality Applications, 2003.

[Rei06a] Reitmayr Gerhard and Drummond Tom W., editors. Going out: Robust Modelbased Tracking for Outdoor Augmented Reality. ISMAR, 2006.

[Ros05a] Rosenhahn B, Ho H, Klette R. Texture driven pose estimation. In: Proceedings of the International Conference on Computer Graphics, Imaging and Visualization (CGIV'05), Beijing, China, July 2005. pp 271–277.

[Sai90a] Saito Takafumi and Takahashi Tokiichiro. Comprehensible rendering of 3-d shapes. volume 24, pages 197–206, New York, NY, USA, 1990. ACM.

[Sch09a] Schumann, M. Achilles, S., and Mueller, S. Analysis by Synthesis Techniques for Markerless Tracking. Gi VR-AR workshop, 2009.

[Str01] Stricker, D. Tracking with Reference Images: A Real-Time and Markerless Tracking Solution for Out-Door Augmented Reality Applications. In Proc. of VAST, 2001

[Wue05a] Wuest Harald, Vial Florent, and Stricker Didier. Adaptive Line Tracking with Multiple Hypotheses for Augmented Reality.ISMAR, 2005