Development of a Real Time Image Based Object Recognition Method for Mobile AR-Devices

Prof. Dr.-Ing. Juergen Gausemeier,
Dipl. Inform. Juergen Fruend,
Dipl. Inform. Carsten Matysczok
Heinz Nixdorf Institute
33102 Paderborn, Germany
{gausemeier, fruend, onestone}@hni.uni-paderborn.de

Prof. Dr. Beat Bruederlin,
Dipl. Inform. David Beier
Technical University of Ilmenau
98693 Ilmenau, Germany
{beat.bruederlin, david.beier}@tu-ilmenau.de

Abstract

In this paper we describe an image based object recognition and tracking method for mobile AR-devices and the correlative process to generate the required data. The object recognition and tracking base on the 3D-geometries of the related objects. Correspondings between live camera images and 3D-models are generated and used to determine the location and orientation of objects in the current scene. The required data for the object recognition is generated from common 3D-CAD-files using a dedicated process model.

The object recognition and tracking method as well as the correlative generation process for the needed data are developed within the AR-PDA project. The AR-PDA is a personal digital assistant (e.g. PDA or 3rd generation mobile phone with an integrated camera), which uses AR technology to efficiently support consumers and service forces during their daily tasks.

Keywords: Augmented reality, AR, mobile devices, object recognition, object tracking, PDA, process model

1 Introduction

Within the last years Virtual Reality (VR) technology has successfully made its way from research institutes into industrial practice. Today VR is one of the upcoming key technologies in many fields of industrial production (e.g. engineering design, manufacturing, mechanical maintenance, training, marketing and sales support) [Gausemeier et al. 2001].

1.1 Augmented Reality

Augmented Reality (AR) is a new form of man-machine interface, which is closely related to VR [Behringer et al. 1999]. In contrast to Virtual Reality, where the user is totally immersed in the computer generated world, Augmented Reality joins com-puter generated word and reality (see figure 1).

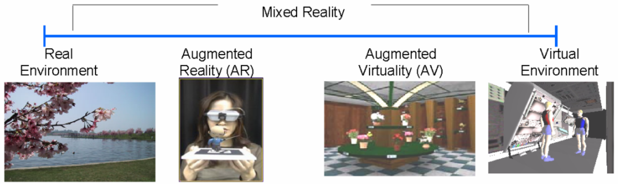

Figure 1: Reality - Virtuality (RV) Continuum [Milgram et al. 1994]

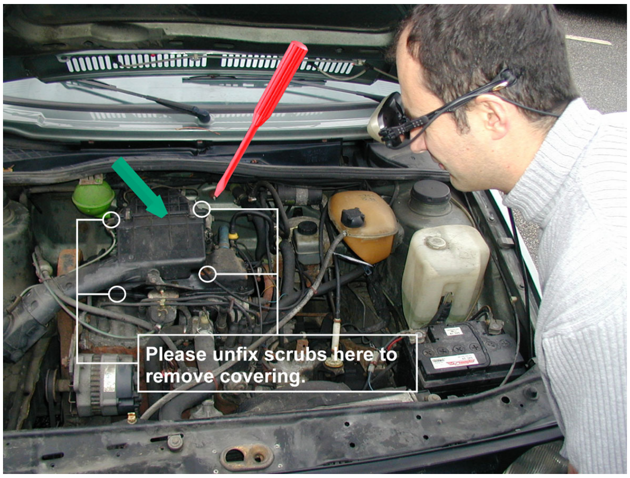

Figure 2: Assembly instructions are inserted into the users field of view via a head mounted display

In Augmented Reality the computer provides additional information, that enhances or augments the real world. The user can interact with the real world in a natural way, with the AR-system providing information and assistance. Therefore, AR enhances reality by superimposing information rather then completely replacing it.

The information can be inserted in a context-dependent way, i.e. derived appropriately from the visualized object. Among others, AR could therefore act as a substitute for the traditional assembly manual, for example by inserting assembly instructions into the field of view of a person per-forming an assembly task (see figure 2).

1.2 The AR-PDA Project

Recent advantages have shown, that the base technology of AR (computer vision, tracking [Kanade and Uenohara 1995], I/O devices, animated 3D-graphics) have matured the point of being usable only by specialists. While these advantages have laid the foundation for use of AR in application areas like maintenance or computer aided design [ARVIKA], there is still a significant lack of research projects, that consider the use of Augmented Reality for the consumer market.

In March 2000 the German Ministry for Research and Training (BMBF) initiated a competition for innovative VR/AR research projects. One of the funded projects is the AR-PDA project. The AR-PDA project involves the development of a hardware and software system for a mobile personal digital assistant (AR-PDA), which can be classified alongside mobile phones, notebooks and organizers [AR-PDA]. This assistant uses AR-technology to support users in their daily tasks. Therefore the user aims at real objects with the AR-PDA. An integrated camera takes the live video images and the AR-PDA sends the video stream by mobile radiocommunication (e.g. UMTS) to the AR-server. The server recognizes the objects by analyzing the image and establishes the relevant context-sensitive information, which is added to the video stream (like multimedia elements, such as sound, video, text, images or virtual objects) and then sent back to the AR-PDA.

We chose domestic appliances as a typical application scenario. In this context the AR-PDA provides the consumer with product information, tips on purchasing, starting up, using, repairing and servicing domestic appliances in a very straightforward way. The AR-PDA also allows direct con-tact with customer services so that an expert advice can be requested. It is also planned to use the AR-PDA in the framework of "mobile business" applications.

The following seven partners from academia and industry are involved in this project: UNITY AG, Siemens AG C-LAB, Lunatic Interactive Productions GmbH, Miele & Cie. GmbH & Co., Technical University of Ilmenau, University of Paderborn and the Heinz Nixdorf Institute.

1.3 AR-PDA System Architecture

Using the AR-PDA real camera images are enhanced by virtual illustration objects (e.g. 3D-objects, videos, pictures, text) and personalized user interactions with the augmented scene are supported.

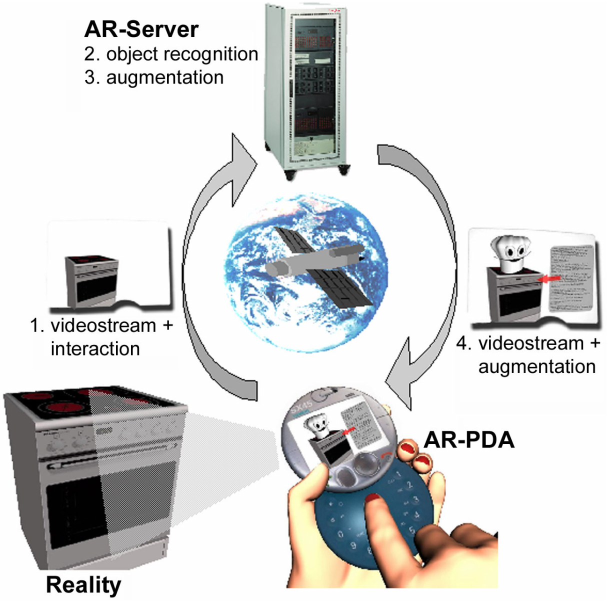

Figure 3: Transmission of the videostream between the AR-PDA and the server

Therefore a dedicated client/server-architecture will be developed [Gausemeier et al. 2002], which supports an outsourcing of the CPU-intensive optical tracking and scene augmentation on a server (see figure 3).

The clients reach from PDAs to mobile phones, which are equipped with cameras to film the reality with the relevant objects. Todays clients are not able to achieve the requirements for the CPU-intensive data processing in the area of augmented reality. Therefore the client primarily displays data and receives user interactions. The scalable design of the whole architecture stays abreast of changes in the development of future mobile devices. So tasks could be distributed flexible between client and server.

The server is responsible for the intensive computing in the field of image recognition, image processing and optical tracking. It receives the video data from the client via wireless broadband connection. Till now TCP/IP over wireless LAN (IEEE802.11b) and telecommunications standards for video conferencing like H.263 are used for the communication between client and server [Geiger et al. 2001; Geiger et al. 2001]. Certainly these standards will be supported in the future by new technologies like UMTS [Prasad et al. 2000].

2 Extraction of Model Characteristics

The image based object recognition system of the AR-PDA is premised on 3D-geometries of the corresponding objects. For this task, the suitable characteristics and attributes of these objects (e.g. dominant edges, homogenous/inhomogenous areas) have to analyzed and extracted form the related 3D-models (see figure 4).

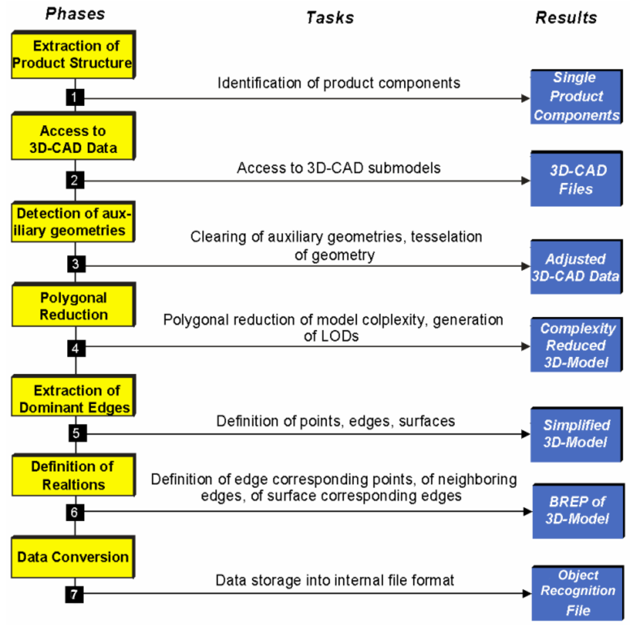
Phases
Tasks
Results
Extraction of Product Structure
1
Identification of product components
Single Product Components
Access to 3D-CAD Data
2
Access to 3D-CAD submodels
3D-CAD Files
Detection of aux-iliary geometries
3
Clearing of auxiliary geometries, tesselation of geometry
Adjusted 3D-CAD Data
Polygonal Reduction
4
Polygonal reduction of model colplexity, generation of LODs
Complexity Reduced 3D-Model
Extraction of Dominant Edges
5
Definition of points, edges, surfaces
Simplified 3D-Model
Definition of Realtions
6
Definition of edge corresponding points, of neighboring edges, of surface corresponding edges
BREP of 3D-Model
Data Conversion
7
Data storage into internal file format
Object Recognition File

Figure 4: Data generation process for the object recognition system

As pool, company appropriate 3D-CAD-systems are used. The image based object recognition and tracking system uses complexity reduced 3D-models. For the data generation process the starting point is the PDM[1]-system. Based on a PDM-file the internal product structure with all its components can be extracted. Each component has a corresponding 3D-CAD file, which is used as 3D-raw-data. The next step is the detection and clearing of auxiliary geometries, which describe the appearance of surfaces in the geometry (e.g. clearances, holes). These are used by some 3D-CAD-systems like *CATIA V4*. Subsequent a polygon reduction is performed, because the performance of the terminal device (PDA or mobile phone) has a significantly reduced 3D-graphics performance than a present desktop computer with a 3D-graphics accelerator card. After the polygon reduction an extraction of dominant edges of the 3D-model is performed. Based on the dominant edges a definition of the corresponding points, edges and surfaces takes place. Sub-sequent the following three relations are generated:

[1] Product Data Management (PDM) describes the management and classification of design data and specifications for an engineered product, and the management of change to this information [Mesihovic and Malmqvist 2000].

- **Points-to-edge relation**
  Definition of the points, which belong to an edge.
- **Edge-to-edge relation**
  Definition of the neighboring edges.
- **Edge-to-surface relation**
  Definition of the edges, which belong to a surface.

This geometrical description, which is similar to the binary representation (BREP) [Barsky et al. 1996] of the object, is converted and stored in the native file format of the object recognition system.The described data generation process was performed on 3D-CAD-data from the household appliance manufacturer *Miele & Cie. GmbH & Co*. Here a complete oven with more than 250.000 polygons is used to generate the related data for the object tracking system. An overview of the number of polygons, the number of scene graph nodes and the memory requirements are displayed in tables 1, 2, 3 and in the corresponding figures 5, 6, 7.

	Original Data	Polygonal Reduction	Dominant Edge Extraction
Baking Muffle	63.158	16.056	418
Single Trolley	26.506	5.272	1.102
Base Frame	16.589	3.146	70
Fat Pan	64.475	15.540	1.548
Grill	15.842	4.351	536
Door	37.830	10.661	579
Baking Sheet	9.171	2.186	56
Pick Up Grill	23.762	5.536	510

Table 1: Number of polygons for each component

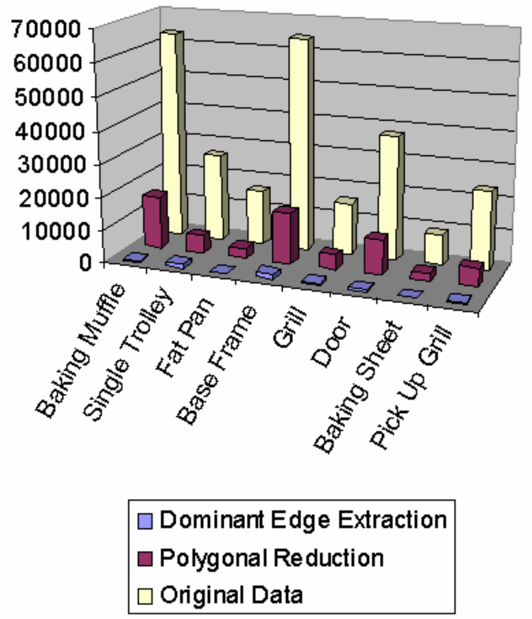

Figure 5: Number of polygons for each component

Based on the original data from common 3D-CAD-systems a polygonal reduction with the highest compression rate using standard polygonal reductions software (e.g. *IMEdit/IMCompress* from *InnovMetric*, *Rational Reducer* from *Systems in Motion*) results in only one fifth of the number of polygons (see table 1 and figure 5).

But this number is still too high of being suitable for the image based object recognition system. Therefore the ex-traction of the dominant edges reduces the number of polygons significantly.

During the polygonal reduction of the 3D-CAD-geometries the number of scene graph nodes is not considered (see table 2 and figure 6). Therefore within the dominant edge extraction a restructuring of the whole scene graph including an erasing of unused nodes is performed.

	Original Data	Polygonal Reduction	Dominant Edge Extraction
Baking Muffle	1.663	1.663	21
Single Trolley	777	777	115
Base Frame	359	359	9
Fat Pan	1.067	1.067	278
Grill	283	283	98
Door	1.222	1.222	50
Baking Sheet	237	237	5
Pick Up Grill	468	468	112

Table 2: Number of scene graph nodes for each component

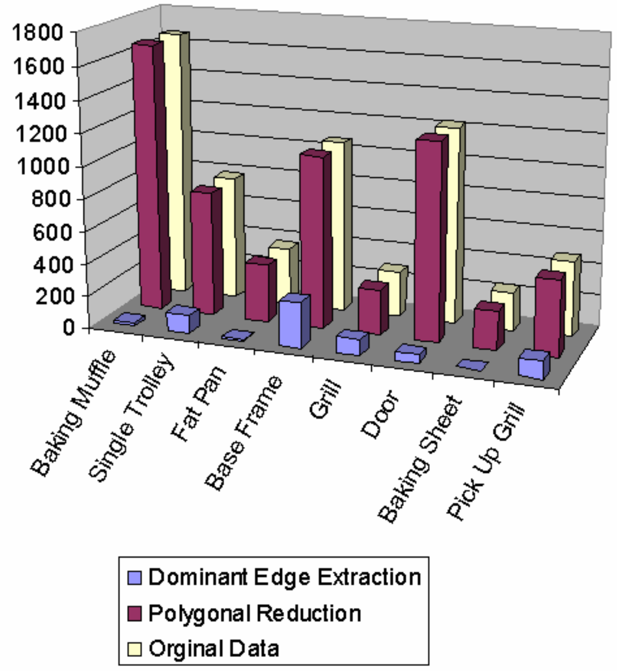

Figure 6: Number of scene graph nodes for each component

The reduction of the polygonal complexity of the models and the restructuring of the corresponding scene graphs results in a decrease of the memory requirements (see table 3 and figure 7).

	Original Data	Polygonal Reduction	Dominant Edge Extraction
Baking Muffle	11.978	2.320	32
Single Trolley	5.036	808	106
Base Frame	3.071	470	10
Fat Pan	11.832	2.030	255
Grill	2.900	547	69
Door	7.309	1.498	67
Baking Sheet	1.778	319	76
Pick Up Grill	4.528	746	6

Table 3: Memory requirements for each component (kilobyte)

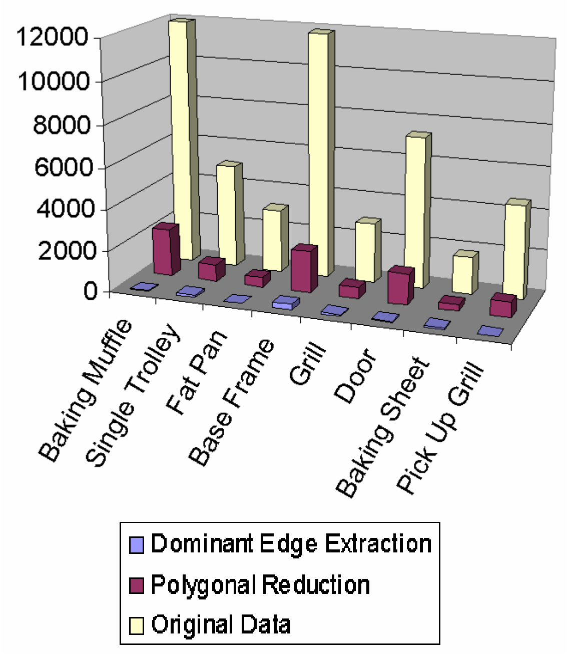

Figure 7: Memory requirements for each component (kilobyte)

The described process model for the extraction of model characteristics (like dominant edges) reduces the complexity of common 3D-CAD-models significantly. Thus the following vision based object recognition method can be performed in real time.

3 Development of an Image Based Object Recognition Method

For object recognition and image analysis many approaches have been developed in recent years [Girod et al. 2000; Parker 1997]. However most of these approaches and techniques are not adaptable to our problem, because they work only in limited conditions or use special devices, such as active vision systems or laser range finder.

For the image based object recognition method, features from the live video image are extracted and compared with features of the database of 3D-models. The models are based on simplified CAD data, containing highly visible edges and faces (see chapter *Extraction of Model Characteristics*).

From the live video images, edges and color information of adjacent faces are extracted, by using several filter operations. Additional texture information of faces may also be extracted. With correspondences of image features and 3D-model features, hypotheses are generate for the orientation of the model relative to the camera (see figure 8).

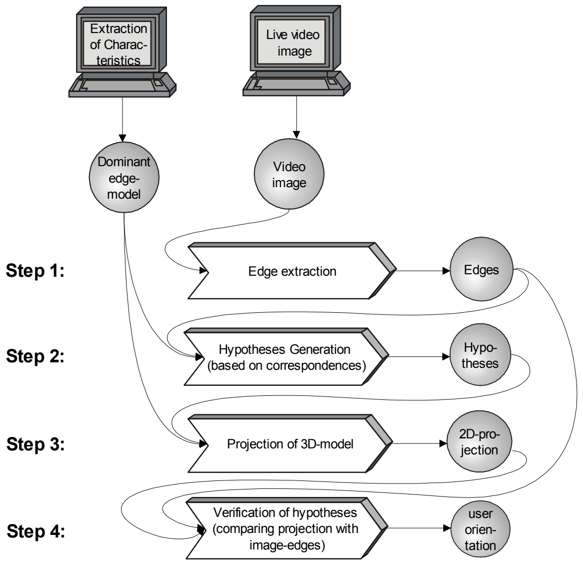

Figure 8: Processing of object tracking

Each generated hypothesis will be verified by projecting the 3D-model into the image plane. This projection is compared with the extracted edges of the live video images and the matching of both graphs is evaluated. The best matching hypothesis is taken as orientation of the model.

3.1 Hypothesis

To limit the number of hypotheses, we have to use several strategies. These strategies may be combined to reduce the number of hypotheses further.

Because of the recognition task of technical objects with many straight edges, we use edge correspondences to compute the orientation of these objects. With three 3D-2D-pairs it is possible to compute the relative orientation of the camera to the object [Rives et al. 1989]. Therefore each hypothesis consists of three edge-pairs.

We implemented the approach by *Dohme* [Rives et al. 1989], which finds the analytical solutions by solving an equation system for inverse perspective projection.

It uses a simplification of the pose determination problem by implicitly dividing it into lower dimensional subproblems. To generate hypotheses, correspondence pairs of edges and vertices are needed. One edge pair consists of a complete model edge and a potentially incomplete image edge.

In the world coordinate system an image edge defines a solution plane, in which the corresponding model edge lies (see figure 9).

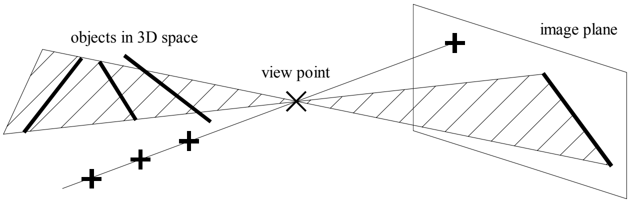

Figure 9: Projection of 3D-edges on image plane

Because of unsafe detection of image vertices, the positions of these vertices are not usable to compute the orientation. The view ray of an image vertex always intersects the corresponding model edge, but only the view ray of a complete image edge intersects the vertices of the model edge. Therefore, we use (infinite) lines instead of edges for the inverse projection of an edge into the world coordinate system.

Each hypothesis is evaluated by projecting the complete model into the image plane and comparing this projection with the input image features. For the amount of correspondence a value is determined: the more features are matched, the higher the value. If the correspondence value exceeds a threshold, this hypothesis is chosen and the testing of hypotheses is terminated, otherwise the best matching hypothesis of all is assumed for the orientation.

3.2 Heuristics

The testing of all possible hypotheses is a polynomial-time problem. For example, if we have *i* input edges and *m* 3D-model edges, the number of combinations *n* is:

$$n = i\ (i-1)\ (i-2)\ m\ (m-1)\ (m-2)$$

since we need at least 3 edge-pairs. *n* is approximately 10^6, if $i=10$ and $m=10$.

We observe, already for a relative small number of features, there are too many combinations for efficiently computing the orientation and performing the object recognition. If we need about 1ms per hypothesis, this would take several minutes.

Without reduction of the number of hypotheses it is impossible to recognize objects and estimate the orientation fast enough for interactive use. Thereby we have to limit the number of hypotheses, by excluding as much as possible from the list. To do this, we have to use several strategies. These strategies may be combined to reduce the number of hypotheses further.

For object recognition it is advantageous to use dominant long edges. Dominant edges are necessary for robust object recognition in different environments and different light sources. A reliable computation of orientation is only possible with long enough edges. With short edges the problem would be ill-conditioned and produce large errors. We therefore limit the edges in the image recognition to dominant, long edges [Mirmehdi et al. 1998]. The following steps are used:

- Merging of neighbouring collinear edges (with threshold for distance and differences in orientation)
- Reduction by limiting edge length (with threshold for minimum and maximum length)

Assuming the users view of the object, we can limit the modelling to include only features, that are visible in this view. For example, in our test scenario, we try to recognize an oven. Only the front face of the built-in oven is visible (see also figure 12). A reduction of model complexity also reduces number of hypotheses. When the orientation of the camera is limited then additional constraints are available for object recognition.

Classification is used for unique identification of edges and the finding of correspondences between image and model edges. Example criteria for classification are:

- Attributes for orientation: vertical, horizontal or diagonal
- Color information about adjacent faces

Lanser and *Zierl* describe in [Lanser and Zierl 1996] the use of topological relations for object recognition. These relations are quasi invariant under perspective transformations. In this sense quasi invariant means that these relations could be assumed to be invariant for some special conditions, e.g. very near parallel edges are also nearly parallel after perspective transformation. Examples of quasi invariant relations:

- Continuity: hypotheses about connection of image-lines, merging of collinear line segments
- Parallelism: looking for parallel lines
- Junctions: classification of vertices
  - By vertex type (see figure 10)
  - By orientation

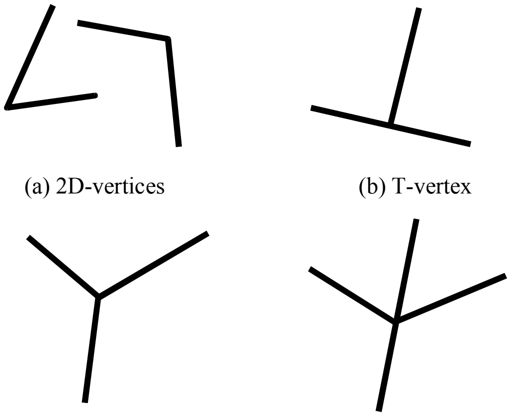
(a) 2D-vertices
(b) T-vertex
(c) 3D-vertex
(d) Other vertices

Figure 10: Different types of vertices

By using the described strategies we can limit the possible correspondences between image and model features. With the found correspondences we can build more accurate hypotheses.

The goal is to terminate the testing of hypotheses as fast as possible. Therefore we try to find more probable hypotheses first. This way, we can reduce the number of hypotheses to be verified.

To influence the order of computation and verification of hypotheses we introduce a weight to define a priority. This weight is determined by statistical values of the probability of finding a correct hypothesis. (For example the probability for unique feature structures is very high.) If the correspondence value of a matched hypothesis exceeds a limit, this hypothesis is used and it will not be necessary to test the remaining hypotheses.

The most expensive part of object recognition and pose determination is the finding of corresponding features. However, because the images are from a continuous video stream and the moving is generally continuous, image features can be tracked.

Regarding feature correspondences of previous images and the previous changing of orientation of the viewer, we can determine discontinuous changes (hiding or appearing features of the model) to update feature correspondences (see figure 11).

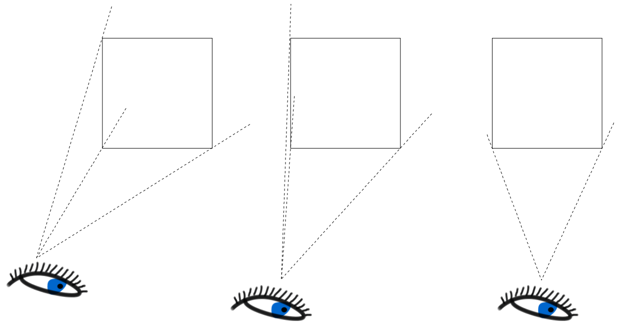

Figure 11: If the view point is moving from left to right, features are hided (middle), or features are appearing by moving from right to left

In most cases we can reuse previous determined feature correspondences for pose determination, as long as the correspondence value from the verification step remains above the threshold.

4 Implementation of the First Prototype

Till this day the design phase for the AR-PDA is almost completed. Now a demonstrator is implemented for the first milestone, in which the AR-PDA is used as a virtual sales assistant to illustrate and explain the details of a technical device (see figure 12). As AR-PDA we use an *Compaq iPAQ 3760* equipped with a wireless LAN adapter and a camera device. The optical object recognition is done by a PIII, 1,7 GHz PC, equipped with a *GForce III* graphics card and a connected Wireless LAN access point.

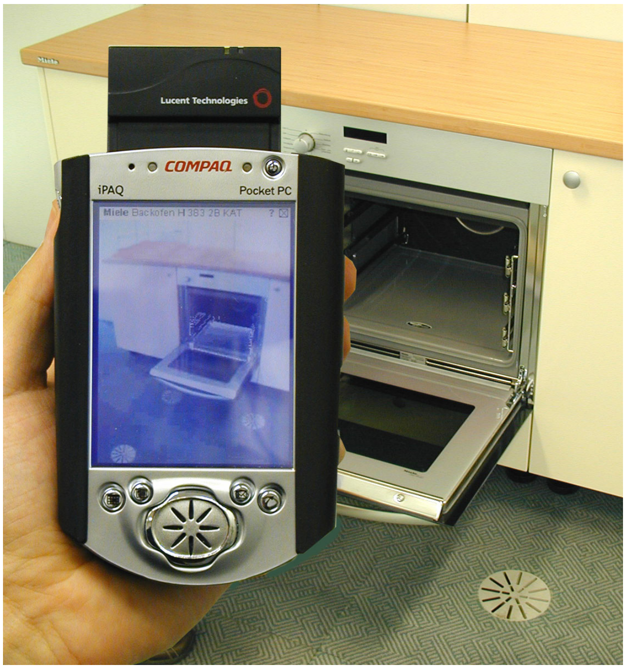

Figure 12: Prototype of the AR-PDA: The real image of an oven is augmented with a virtual telescope trolley

The actual prototype allows the object recognition and tracking of moving objects, an augmentation of the scene using animated 3D-objects or pictures and a personalized user interaction by touching the display with a pen.

One of the major challenges was the reduction of the high delay times - first it took about three seconds to refresh the picture when moving the AR-PDA or making an user inter-action.

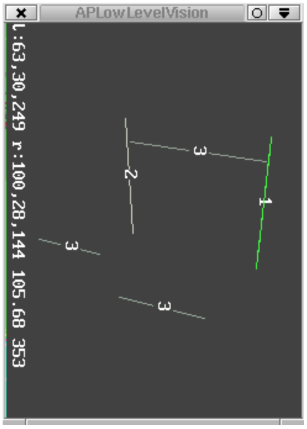

Figure 13a: Extracted edges from live video

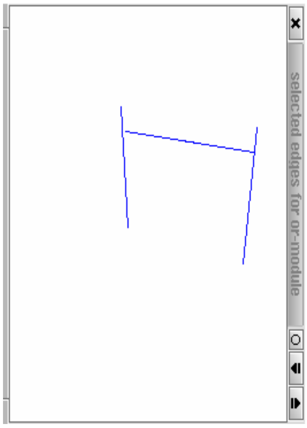

Figure 13b: Corresponding edges from 3D-CAD-model

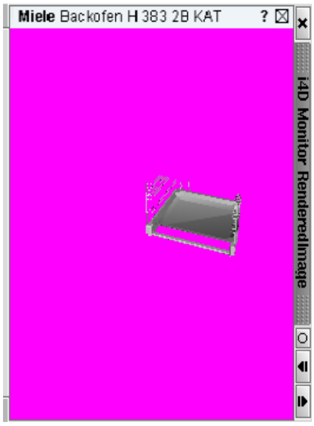

Figure 13c: Augmentation with 3D-illustration objects

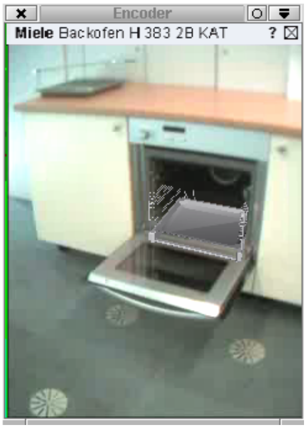

Figure 13d: Composition of video and augmentation

With the before described strategies for reducing of hypotheses, we can significantly lower the number of combinations from about 10^6 to less than 1000. For a good interactability, we need at least 16 frames per second (fps). This means, if it takes about 1ms for calculation and verifying per hypothesis, the number of hypotheses has to be limited to less than 63 (≈1000/16).

In our first prototype, with significant features and only one object to recognize, we reduced the number of hypotheses down to about 100. This results in a object recognition working with 10 fps.

In figure 13a the significant edges extracted from the camera image, marked with *1* and *2* are shown, which are compared with the significant edges (see figure 13b). However the success of the approach depends strongly on the recognition of the significant features. In difficult or changing lighting conditions the features are detected, but they can-not be classified by their color and therefore the number of hypotheses rises. Like every vision-based tracking system the shown approach cannot work, if the images have too little contrast.

We sped up the estimation of position of already recognized objects even more, by using feature tracking of previously used features. In cases where feature tracking works well, it takes a very short time (≤ 10ms) for computation of a new position which ensures interactive frame rates for the augmentation of the images. Figure 13d shows the camera image of an oven which is augmented by an animation of a virtual moving shelf (see figure 13c).

The video is also updated at interactive frame rates, e.g. upon changing the camera viewpoint in 3D. In this case an interaction in real time is now possible.

5 Summary and Outlook

The envisaged approach describes a real time image based object recognition method for mobile devices, which bases on the 3D-geometries of the related objects. The generation of the correspondences between the live video images and the 3D-geometries are used to determine the location and orientation of objects in the current scene. Till now this new object recognition method is fast enough to support an interaction in real time in a limited scope (recognition of only one object and a highly contrasting environment). But not all described methods of making assertions are still used in the current prototype.

The main focus of the next steps is the execution of further performance tests and an improvement of the systems performance as well as an inclusion of a higher number of attributes for the object tracking system (like homogenous and inhomogenous areas). Therefore the extraction of model characteristics must be adapted as well as the examination of further heuristics for edge and feature extraction from 3D-CAD-models. Moreover the occurred interferences (e.g. changing lighting conditions, low contrast environment) must be minimized to get a more robust object recognition and tracking system.

References

AR-PDA. http://www.ar-pda.de/

ARVIKA. http://www.arvika.de/

BARSKY, B., CHENG, F., SUN, J. AND WANG, X. 1996. Boundary Representation, Intersection, and Evaluation of NURB Based Non-manifold Objects. In *Proceedings of the 1996 ASME Design for Manufacturing Conference (ASME DTC-96 DFMC)*. August, Irvine, CA.

BEHRINGER, R., KLINKER, G. AND MIZELL, D. 1999. Augmented Reality - Placing Artificial Objects in Real Scenes. In *Proceedings of the IWAR 1998*. San Francisco, California.

GAUSEMEIER, J., EBBESMEYER, P. AND KALLMEYER, F. *Produktinnovation - Strategische Planung und Entwicklung der Produkte von morgen*. Carl Hanser Verlag.

GAUSEMEIER, J., FRUEND, J. AND MATYSCZOK, C. 2002. Development of a Process Model for Efficient Content Creation for Mobile Augmented Reality Applications. In *Proceedings of the CAD 2002*. Dresden.

GEIGER, C., KLEINJOHANN, B., REIMANN, C. AND STICHLING, D. 2001. Interaktive VR/AR-Inhalte auf mobilen Endgeraeten, In *Informatik 2001, Workshop 22, Synergien zwischen virtueller Realitaet und Computerspielen: Anforderungen, Design, Technologien*. Wien.

GEIGER, C., KLEINJOHANN, B., REIMANN, C. AND STICHLING, D. 2001. Mobile AR4All. In *Proceeding of the Second IEEE and ACM International Symposium on Augmented Reality*. New York.

GIROD, B., GREINER, G. AND NIEMANN, H. 2000. *Principles of 3D Image Analysis and Synthesis*. Kluwer Academic Publishers, Boston, USA.

KANADE, T. AND UENOHARA, M. 1995. Vision-Based Object Registration for Real-Time Image Overlay. In *Proceedings of Computer Vision, Virtual Reality and Robotics in Medicine*.

LANSER, S. AND ZIERL, C. 1996. On the Use of topological Constraints within Object Recognition Tasks. In *ICPR Volume 1*.

MESIHOVIC, S. AND MALMAQVIST, J. 2000. Product Data Management (PDM) System Support for the Engineering Configuration Process. In *ECAI 2000, Configuration Workshop*. Berlin.

MIRMEHDI, M., PALMER, P.L. AND KITTLER, J. 1998. Optimizing the Complete Image Feature Extraction Chain. In *Proceedings of the 3rd Asian Conference on Computer Vision*.

MILGRAM, P., TAKEMURA, H., UTSUMI, A. AND KISHINO, F. 1994. Augmented Reality: A Class of Displays on the Reality-Virtuality Continuum. In *Proceedings of SPIE Conference on Telemanipulator and Telepresence Technologies SPIE*. Boston, MA.

PARKER, J. 1997. *Algorithms for Images Processing and Computer Vision*. John Wiley and Sons, New York, USA.

PRASAD, R., MOHR, W. AND KONHAEUSER, W. 2000. *Third Generation Mobile Communication Systems*. Artech House Publishers.

RIVES, G., DOHME, M., RICHETIN, M. AND LAPRESTE, J.-T. 1989. Determination of the Attitude of 3D-Objects from a Single Perspective view.