Vision-Based Interaction – A First Glance at Playing MR Games in the Real-World Around Us

Volker Paelke
University of Hannover, IKG
Appelstraße 9a
30167 Hannover
+49 511 762 2472
Volker.Paelke@ikg.uni-hannover.de

Christian Reimann
Paderborn University, C-LAB
Fürstenallee 11
33102 Paderborn
+49 5251 606118
Christian.Reimann@c-lab.de

ABSTRACT
Mixed-reality games have the potential to let users play in the world surrounding them. However, to exploit this new approaches to game content creation, content presentation techniques and interaction techniques are required. In this paper we explore the potential of computer-vision on mobile devices with a camera as an interaction modality. Based on a theoretical review of the available design space potential interaction techniques are discussed. Some of these were implemented in an experimental game to enable practical evaluation. We provide an overview of the game and present intial experiences with the vision-based interaction techniques employed.

Categories and Subject Descriptors
I.3.m [**Computer Graphics**]: miscallenous

General Terms
Design, Human Factors.

Keywords
Mobile gaming, Computer Vision

1. INTRODUCTION
The great commercial success of computer gaming in the last decade has changed the common understanding of "games" significantly: While traditionally "games" and "play" described activities ranging from board games over outdoor activities to sports, it is now mostly associated with computer games in which a player sits in front of a computer screen and interacts with a mouse, keyboard or joystick. While current computer games have great attraction for a limited audience they lack several of the appealing aspects of traditional games, e.g. to serve as a catalyst for social interaction, to make the hands-on acquisition of real world knowledge enjoyable and to incorporate the training of practical skills.

Emerging technologies from the domains of ubiquitous and mobile computing, augmented and mixed reality and spatio-temporal sensors have the potential to evolve the user interface of computer games from the keyboard/mouse/monitor environment into a more natural and intuitive interaction environment, where multiple players interact in a real-world indoor or outdoor environment through physical multi-modal actions. This style of mixed-reality (MR) games will eventually allow to combine the merits of traditional games with those of computer games to create new forms of game experiences. Although, some well known experiments have been conducted in the domain of MR games (e.g. AR Quake) research in the domain is still in an early stage. For this paper we have focused on the special requirements of interaction techniques for MR games (section 2), specifically on the use of interaction techniques that exploit the camera of mobile devices as their primary sensor (section 3). To conduct meaningful evaluations of our interaction techniques these have been integrated into experimental MR game applications that are described in section 6. Section 7 closes with initial resulty and observation and provides an outline of future work.

2. THE DESIGN SPACE OF MR GAMES
Most existing computer games are completely virtual environments. As the game world is created from scratch, game designer have complete control and enjoy many degrees of freedom in the design. However, this complete separation from reality also prevents the use of real-world objects and features within the game, constraining interaction to the joystick/display interface. The use of emerging sensor and interaction technologies allows extending this design space significantly by incorporating real-world environments into games. As Figure 1 shows games taking place in a real-world outdoor environment form the other end of the spectrum, where game designers have only minimal influence on the environment.

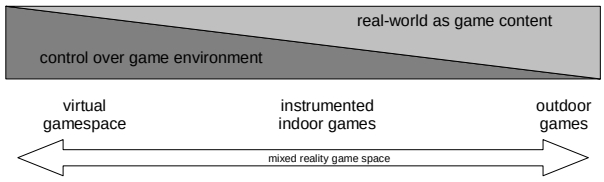

Figure 1: From Virtual Gamespace to Oudoor Games

The spectrum of MR games covers the complete area between these two extremes. The creation of MR games that integrate the game experience into real-world environments introduces a number of requirements that differ significantly from "conventional" computer games:

1) New approaches to content authoring and modelling are required as well as game concepts that exploit real-world features

in the game. For example, 3D models are a basic constituent of most computer games. For conventional computer games the 3D models of the game world are created with 3D modelling tools. However, once real-world environments are to be integrated into games this introduces several problems: Outdoor MR games require accurate and up-to date 3D environment models both for graphics generation and as the spatial basis for augmentation, which is difficult and cost-intensive to achieve with traditional modelling tools, especially for larger environments. In indoor MR games the same requirements arise with somewhat reduced correctness criteria. Correct 3D models are also essential if blended multiplayer gaming with indoor and outdoor players is intended.

2) Appropriate presentation styles are required for the creation of game output that ensures perceptibility of relevant information under the constraints of current MR display devices. The optimal graphics solution would provide users of different MR devices with detailed high quality graphics that integrates seamlessly with the surrounding environment and places only limited requirements on storage and transmission. Since current hardware still imposes mayor limitations in this domain, game designers have to develop effective work-arounds, e.g. through the use of illustration techniques and abstracted presentation styles.

3) Interaction on mobile devices is severly constraint by the available input modalities. The challenge here is not only to find usable and effective replacements for the interaction techniques available in conventional computer games, but also to develop means that exploit the user's real-world context to influence gameplay, in order to turn the world around the user effectively into his "game board".

3. INSIDE-OUT VISION

Our choice of inside-out vision as an interaction modality is motivated by the widespread availability of camera equiped PDAs, smartphones and similar devices. Due to the formfactor of the devices, into which the camera is embedded, these are typically used in an inside-out setup. This means that the camera itself is manipulated in space to effect some interaction. The videostream captured by the camera is analyzed to derive high-level interaction events that control the application.

The additional input mechanism available on the mobile device (e.g. buttons) can be combined with the camera input to create more complex composite interaction techniques. So far, such interaction techniques have mostly been created on an ad-hoc basis by computer vision experts for use in technology demonstators. Reuse has taken place largely based on availability, e.g. techniques used in publicly available demo programs have sometimes been reused in other programs based on implementational convenience, not on informed choices in the user interface design. Currently, little is known about the usability of inside-out vision (IOV) techniques, no librarys exist, and the exploration of IOV techniques and their application is still at an early stage. To structure our rearch and development efforts we have structured the design space of IOV techniques. Such approaches have proven to be useful for the general study of interaction techniques in the past (e.g. [2]). In the following sections we identify the influences and constraints inherent in the.

4. INFLUENCES AND CONSTRAINTS

The constraints that influence the design of interaction techniques based on inside-out vision can be separated into two categories: those that are due to the sensor and those that are due the human user and his environment.

Card's design space of input devices [2] is based on the physical properties that are used by input devices (absolute and relative position, absolute and relative force, both in linear and rotary form) and composition operators (merge, layout, connection). Interaction techniques are constructed by combining several physical properties accessible to sensors through composition operators and mapping the resulting input domain to a logical parameter space suitable for applications. In order to integrate IOV into this framework it is necessary to identify what properties can be sensed using a camera in the inside-out configuration. Differing from direct physical sensors the input properties must be extracted from noisy high-bandwidth image sequence. Table 1 shows what properties can be derived from image sequences. In practice, the requirement of interaction techniques to operate in real-time with minimal lag is often in conflict with the high processing requirements of computer vision techniques, especially if local processing on a mobile device is intended, so only a subset of these possibilities can be used.

TABLE 1: POSSIBLE INPUT PROPERTIES

Absolute position: Absolute positioning is only possible if a point of origin is provided that allows establishing a spatial relation between the environment and the image captured by the camera. A possible solution that allows for fast and relatively precise positioning is the use of markers/fiducials at known positions. Several software packages support 6DOF positioning using cameras and markers (e.g. ARToolkit [1]).

Alternative "marker-less" approaches (e.g. [7, 11]) use a geometric model of the environment instead of markers. The main advantage is that no artificial markers in the environment are required, making them more appropriate for mobile and wearable systems. However, "marker-less" approaches are often more sensitive to environmental effects like changes in lighting, depend on the strucure and "content" of the environment and the more complex image and model processing typically results in higher latency in the interaction. If no geometric model of the environment can be provided in advance, as is typically the case in mobile applications, it is necessary to construct the model on the fly, which is an active area of research ([11]).

These absolute positioning techniques can be used to determine the position and orientation of the IOV camera in all six degrees of freedom (6DOF), thus proving access to all three linear and three rotary degrees of freedom in Card's design space. However, the precision of the information can vary significantly.

The detection of the presence/absence of objects is another useful information that can be exploited in IOV. Because of its similarity to button-presses in conventional interfaces it is grouped under

absolute positiong, although it does not require a point of origin. Again the detection of prepared objects like barcodes and markers is simpler than that of generic real-world objects, but solution exits for both.

Relative position (motion): Motion can be sensed in three linear (x, y, z) and three rotary degrees of freedom by processing the incoming video stream. No point of origin is required for the detection of motion from image sequences, allowing the use in unprepared environments. However, in practice the precission that can be attained in unprepared environments is limited. While 2DOF motion detection is suitable for the limited processing power of current mobile devices (and for which special purpose hardware used in optical mice and video compression could eventually be adapted) 6DOF motion tracking is much more difficult and computationally intensive. If the environment is specially prepared, e.g. by placing and tracking fiducials, procesessing on mobile devices becomes possiblity (e.g. [14]); otherwise the processing often has to take place on more powerful hardware, using a client-server approach that can introduce problematic latencies.

Absolute and relative force: Information about force can not be extracted from image data without addition transducer hardware.

To identify the influences and constraints introduced by the human user and his environment the following questions must be considered when constructing an IOV interaction technique:

1. Is the required positioning and motion of the camera possible for the user? This refers both to constraints on possible positions due to user anatomy, as well as to physical constraints imposed by the surroundings (e.g. use in an office vs. use on a plane).
2. Is the required positioning and motion of the camera comfortable for the user? IOVs will only be used if users prefer them to alternative techniques so that criteria like fatigue, precission and speed must be considered.
3. Are the required positioning and motion of the camera acceptable? For most applications IOVs will not be used if the required motions are embarrasing in public.
4. Are the required input properties sensable with the available hardware? As discussed previously, only a subset of the theoretically available input properties can be used in practice. It has to be ensured that the required input properties can be provided with appropriate accuracy, speed and latency under the conditions of use.
5. Is it possible to differentiate intentional inputs from unintentional camera movements? To avoid the “midas touch”-problem means to distinguish input from unintentional noise must be provided, e.g. by explicit input confirmation.
6. Is the mapping from inputs to interaction events unambigous?

5. POSSIBLE USES OF INSIDE-OUT VISION

The following discussion of (possible) uses of IOV is structured according to the interaction tasks select, position, quantify and gesture. It is based on the popular taxonomy of Foley et al. [3]. Due to the characteristics of IOV we have replaced the text task in the original taxonomy with a generic gesture recognition task. Interaction tasks specify what a users can try to achieve in an application on an abstract level - for the implementation in an actual user interface a concrete realization in the form of an interaction technique is required. Exemplary interaction techniques based on IOV are presented for the interaction tasks:

Select: The select task refers to symbolic selection from a set of options. Different approaches to symbolic selection are enabled by IOV: An interesting approach based on the tangible computing paradigm can be used if the set of options can be represented by associated physical objects. Then selection can be effected simply by placing the camera so that the object is in the camera’s field of view. Examples for this include the use of barcodes which are easy to recognize even on performance limited hardware, the use of more complex markers (that also enable more complex tasks) or the use of geometry or image based object recognition.

While selection based on physical objects has interesting properties for some applications it often can not be used either because the application has to operate in unprepared environments or because the set of options is to large or changes dynamically. In these cases approaches based on virtual representions of the set of options similar to menus in a desktop interface can be used. Figure 2 shows the use of “Kick-Up-Menus” ([9]). Here simple motion detection is used on the image sequence provided by a camera facing downward from a PDA or Smartphone to detect "kicking" movements of the user's feet. When a collision between the user's "kick" and an interaction object shown on the screen of the mobile device is detected, a corresponding selection event for the application is generated. As Figure 2 shows “Kick-Up-Menus” can be structured hierarchically to enable access to large sets of options.

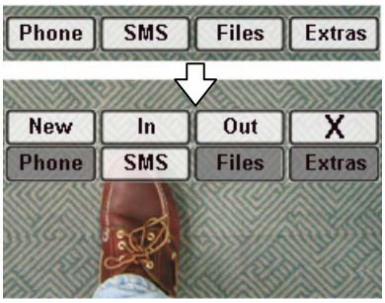

Figure 2: Kick-Up-Menus and PDA with IOV camera setup

A common selection task in 3D applications is spatial selection. While spatial selection of physical objects can be realized as described previously, spatial selection of virtual objects typically has to be constructed from one or more positioning tasks as described in the following subsection.

Position: Different from desktop environments where positioning usually refers to xy-positiong using the mouse VR and AR applications often require postioning with up to 6 degrees of freedom. As discussed in chapter 3 absolute positioning in 6DOF is possible using IOV if a point of origin is provided.

Figure 3: The Mozzies game on the SX1 smartphone with IOV camera

In these cases the 6DOF positiong data provided by the computer vision algorithm can be mapped (possibly through some transfer function) to the application domain. To provide positing data with adequate precission and lag most existing applications use marker based approaches, e.g. ARToolkit [1] and Sony's Cybercode [10]. If not all 6DOF are required simpler, faster and more robust algorithms can be used that are suitable for mobile devices. Figure 3 shows the Mozzies game on the Siemens SX1 smartphone that uses simple 2D motion detection and a crosshair for 2D xy-positing.

Quantify: The quantify interaction task is used to specify numeric values as input parameters to the application. In mouse-based interfaces potentiometer, slider and scollbar widgets are often employed for this task. A similar approach is used in Spotcodes [12]. Spotcodes is a system based on circular markers from which rotation information can be derived. Interaction techniques are provided for the specification of rotation angles and values. Sometimes a direct mapping from the input to the application domain is possible without the need for widgets as an intermediary. In this way the pitch angle of the camera has been used to control scrolling (instead of a scrollbar widget). Figure 4 shows ARSoccer, a mobile soccer application [4]. Here the direction and speed of a motion vector generated by a kicking foot are used to control a simple soccer game, resulting in an intuitive mapping between the input and application domains. The interaction techniques of AR-Soccer are now used in a commercial game implementation [5].

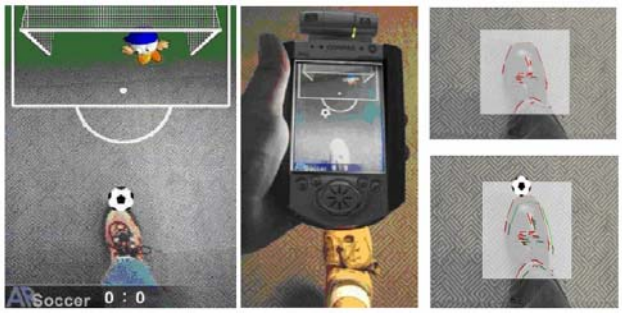

Figure 4: The AR-Soccer Application with simple edge tracking

Gesture: Gestures refer to the symbolic interpretation of camera motion. This can range from simple yes/no gestures over a simple gesture vocabulary (similar to mouse gestures in some applications) to complex sign languages. Here a careful tradeoff between the learning required of the user to become proficient with the gestures, the requirement for unambigous gesture identification, the required processing power and the expressiveness of the gesture set is required. So far most application use only simple gestures but techniques and gestures developed for the domain of head-gestures that shares may properties with IOV (e.g. [6]) could in principle be adapted to IOV.

6. EXAMPLE: IOV IN THE FORGOTTEN-VALLEY ADVENTURE GAME

To explore some of possibilities of IOV in games we have developed a small adventure-style game using our MobEE game-engine. The adventure "Forgotten Valley" demonstrates the capabilities and possibilities of IOV that are currently supported by MobEE in a blended mixed-reality setup that enables both indoor and outdoor use.

Starting the adventure the user is offered the opportunity to either start a new game or continue a previously played storyline. By choosing to play a new game he finds his Avatar placed in the middle of an unknown map (figure 5), not knowing where he is or how he got here. In mixed-reality mode the user can start physically anywhere on the university campus that is our real-world "game board" for "Forgotten Valley".

Figure 5: Starting point

In conventional mode the user can use the pointing device (which can vary between different mobile devices) to move across the map which is scrolling according to the avatars' movements so that the avatar represented by a small person always stays in the centre of the screen. Exploring the surroundings in this manner, the player encounters different places where he may find hints about his whereabouts and how to move on in the game. In mixed-reality mode the user physically walks around on the university campus to discover the places relevant for the game.

Figure 6: Riddles to solve ("Gate" left and "Oracle" right)

The user has to solve several little puzzles (see figure 6) and talk to the people populating the valley to eventually find his way out.

All actions of the user and corresponding "experiences" of his avatar are recorded by the program and saved into a file. This information can later be used as the basis for a context refresh when the user wants to re-enter a previous played game.

When the user chooses to continue a game that he started at an earlier time, he is presented with an automatically generated re-narration of his previous adventures in the game world (see figure 7). The context refresh shows the most important events in the storyline (as specified by the game designer). The context refresh or scenes therein can be skipped by the user by pressing the "fast-forward" button.

Figure 7: Context Refresh, showing an important part of the story

The game uses background music, spoken parts and written text to tell a story that is designed to be interesting and captivating. Clicking on the menu-bar the user can choose between different combinations of output modalities (e.g. text, graphics, audio, or mixed-reality). The same adventure can thus be played as a pure text-adventure, as a 2D graphics game or a mixed-reality experience using the same game-engine. To ensure an enjoyable game experience in text-only mode more detailed descriptions of the locations could be added to substitute for the graphics and a linked map in order for the avatar to move around. The following sub-section describes the mixed-reality mode in more detail.

6.1 IOV in the Mixed-Reality Mode

Gameplay in mixed-reality mode is similar to that in normal mode as described before: While navigating the user is presented with a scrolling raster map of the university campus onto which icons representing the "game locations" are added if the user has explored the corresponding part of the game. At a "game-location" the user can interact with the real-environment using the camera on the PDA. Our current version of the mixed-reality setup is implemented on a HP iPaq Pocket PC PDA with a plug-in camera (FlyCam). To track the user's position in the real world while he is walking around, we use a GPS-sensor (Holux GR-230), which has a wireless Bluetooth connection to the PDA. To avoid problems caused by the low update rate of the GPS the navigation has two main states: "walking" and "waiting". While in "walking"-state the game is continuously updated approximately three times a second with extrapolated data from the GPS. The "waiting"-state is entered, when the user is interacting with the game at a "game-location", e.g. solving a riddle. While in "waiting"-state all the information from the GPS is ignored. The main reason for ignoring the GPS in this state, is that the GPS-data can drift, meaning that the GPS-position could move even when the user is not. The "waiting"-state is left, when the user explicitly finishes interacting with the actual "game-location" (e.g. has solved the riddle and gathered the information) or when he simply walks away (when the position difference exceeds a preset threshold).

If the user is at a "game-location" he can use the camera of his mobile device to capture an image of his surroundings that is then augmented with the graphical game content. At the "game-locations" (or hotspots in the conventional presentation) the user interacts with the game more intensively than just navigating. Here he meets NPCs (Non-Player-Characters), solves riddles, fights dragons and so on. While GPS data is sufficiently accurate to determine if the user is approaching a "game-location" and inform him accordingly, it does not provide the required accuracy for augmenting images of the user's surroundings spatially correct with game information. As there is no other sensor available on the PDA IOV is used. Therefore, the current prototype uses ARToolKit [1], a computer-vision fiducial based tracking system for AR-applications for the actual augmentation. As vision based tracking is too computational expensive for most devices currently available we have implemented a "snap-shot" AR approach: The user takes a single picture with the PDA's camera, which is then analyzed and taken as a static background for rendering. Since only the augmentation graphics have to be rendered the impact of the hardware constraints are reduced since the user has high-fidelity context information from his real-surroundings. This way interactive framerate (>10 fps) with appealing graphics can be realized on most Pocket PC PDAs. We have found that the static image is usually sufficient to establish the link between the game content and the environment, although

real-time 3D tracking and augmentation remain a desirable goal.Depending on the game content taking snapshots of specific markers is also used as in interaction technique to trigger actions within the game.

Figure 8 shows the same riddles as in Figure 6 within the physical environment on the campus.

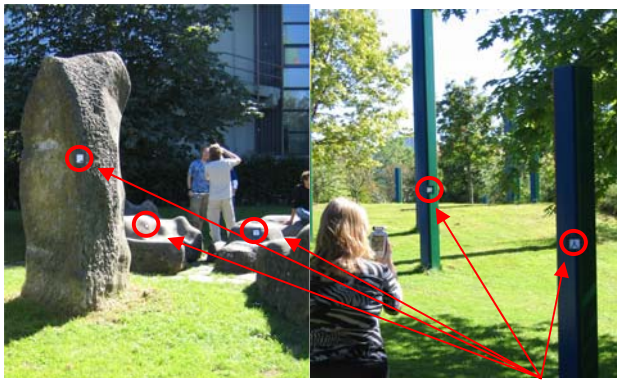

Figure 8: MR-locations "Oracle" and "Gate" (with Markers)

When the user approaches the group of stones (Figure 8, left) the scrolling map on the PDA signals a possible "game-location". When the user takes a picture of one of the markers the "Oracle-Riddle" starts, similar to the one in the 2D-Version. After a short explanation of the riddle the user has to take pictures of the markers on the stones in the right order to solve the riddle. When he succeeds, additional information is displayed, that tells him about "a dangerous dragon of huge ancient wisdom" and the story continues.

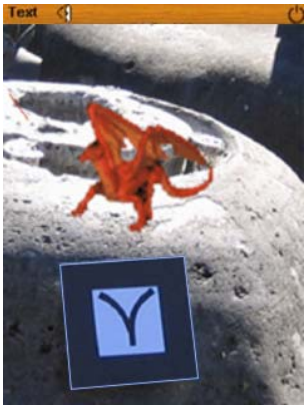

Figure 9: Dragon in MR mode

7. OUTLOOK

Work on IOV based interaction techniques is still at an early stage. We have tried to provide an overview of the available design space and illustrated it with examples. Several areas are of interest for future work:

On the theoretical side the combination of IOV with other input modalities is an interesting domain to explore. PDAs and smartphones typically provide a number of buttons or even a touch screen. Using Card's design space the resulting possibilities can be explored systematically. The construction of specialised IOV input devices consisting of a camera and extra sensors could also be interesting. For example, pressure sensors could be added to make the properties of relative/absolute force accessible to cover the complete design space.

On the practical side the viability and usability of IOV based interaction techniques is best explored by experiment. However, computer vision is a hard problem even with existing libraries (e.g. [8]). A problem with many existing computer vision algorithms is that they were designed for other purposes and that "intermediate results" that can often be exploited in IOV based interaction techniques, are not accessible to the user. The adaption of computer vision techniques to the requirements of designing IOV interaction techniques is therefore necessary. Possible hardware support for these computer vision techniques is another interesting research problem. We have found mixed-reality games to be an attractive test platform for IOV techniques, since the gaming aspect is attractive for test users and the shortcomings of interaction techniques that are inevitable in prototypes of interaction techniques are typically handled as part of the game challenge, leading to valuable feedback even from early and rudimentary prototypes. As the design space of IOV based interaction techniques awaits further exploration, games could play an important part of exploring it and making it accessible to real-world users.

8. REFERENCES

[1] ARToolkit: http://hitl.wahington.edu/research/ shared_space, accessed 28. Jan. 2004

[2] Card, S. K.; Mackinlay, J.D. and Robertson, G.G.: A Morphological Analysis of the Design Space of Input Devices, ACM Transactions on Information Systems, Vol. 9, No. 2, April 1991, pp. 99 - 122

[3] Foley, J. D. ; van Dam, A.; Feiner, S.K. and Hughes, J. F.: Computer Graphics - Principles and Practice, Second Edition in C, Addison Wesley, 1996.

[4] Geiger, C.; Paelke, V. and Reimann, C. :Mobile Entertainment Computing, Lecture Notes in Computer Science, Vol. 3105 / 2004, Springer Verlag 2004, pp. 142 - 147

[5] KickReal: http://www.kickreal.de/, accessed 28. Jan. 2004

[6] Kjeldsen, R.: Head Gestures for Computer Control, Proc. IEEE RATFG-RTS Workshop on Recognition And Tracking of Face and Gesture, Vancouver, Canada, July 2001, pp. 61-67

[7] Neumann, U. and You, S.: Natural Feature Tracking for Augmented-Reality, IEEE Transactions on Multimedia,

[8] OpenCV: OpenSource Computer Vision Library, http://www.intel.com/research/mrl/research/opencv/, accessed 28. Jan. 2004

[9] Paelke, V.; Reimann, C. and Stichling, D.: Kick-Up-Menus, in: Extended abstracts of ACM CHI 2004, Vienna, 2004

[10] Rekimoto, J. and and Ayatsuka, Y.: CyberCode: Designing Augmented Reality Environments with Visual Tags, Proc. Designing Augmented Reality Environments DARE 2000, Elsinore, Denmark, April 2000

[11] Simon, G. and Berger, M-O.: Reconstructing while registering: A novel approach for markerless augmented reality, in: Proc. IEEE and ACM International Symposium on Mixed and Augmented Reality, pp.285-294, 2002

[12] Spotcode: http://www.highenergymagic.com, accessed 28. Jan. 2004

[13] Stichling, D. and Kleinjohann, B.: Edge Vectorization for Embedded Real-Time Systems using the CV-SDF Model, Proc. Vision Interface 2003 , Halifax, Canada

[14] Wagner, D.: Porting the Core ARToolKit library onto the PocketPC Platform, Proc. 2nd IEEE International Augmented Reality Toolkit Workshop, October 2003, Tokyo, Japan