HUMAN PRESENCE DETECTION BY SMART DEVICES

Bogdan Raducanu, Sriram Subramanian and Panos Markopoulos
UCE Group – Technical University of Eindhoven
Den Dolech 2, 5612 AZ Eindhoven
The Netherlands
E-mail: {b.m.raducanu; s.subramanian; p.markopoulos}@tue.nl

ABSTRACT

This paper discusses a computer vision based approach for enhancing a common device (display) with machine perception capabilities. Using techniques for assessing the distance and orientation of a target from the camera, a "smart device" becomes aware about user's presence and his interaction intentions. In other words, the "smart device" is aware when it becomes the user's focus of attention and it knows to respond accordingly. Our solution uses low-cost cameras adapted with infrared technology and is designed to be robust to lighting variations typical of home and work environments.

KEYWORDS

Ubiquitous computing, smart devices, computer vision, human-computer interaction, implicit input

INTRODUCTION

Considering ubiquitous computing from the standpoint of the user, an important departure compared to the current model of interaction with computing devices, concerns the notion of *implicit input.* Implicit input [1] entails that our natural interactions with the physical environment provide sufficient input to a variety of non-standard devices without any further user intervention. Such automatically captured input contrasts the current model of interaction where the user has to perform several secondary tasks relating to the operation of the computing device in order to achieve their primary task.

Implicit input can be achieved with a variety of technologies (pressure sensors, video cameras, radio-frequency tags, fingerprint readers), which are integrated in objects commonly found in our environment (chairs, tables, displays). This paper discusses an application of computer vision techniques to support implicit input in ubiquitous computing environments. In particular, we discuss the detection of a person proximity to an object of interest and whether he is facing towards the object. This helps to estimate whether that objects becomes user's focus of attention. The intuitive idea that people will face objects they are interested in, before starting the interaction itself, is supported by some empirical research at Microsoft [2].

There has been significant research for proximity detection indoors. Existing solutions address the problem at different scales and varying resolutions [3]: building level, room level and sub-room level. The solution proposed in this paper falls is suitable for the last of the three categories.

We found that the main technologies used to address this problem (localization at sub-room level) are based on ultrasonic and computer vision.

In [4] and [5], they propose a system called Active Bat that uses a grid of fixed receivers to detect ultrasonic pulses emitted by small badges that can be carried by users. The transmitters and receivers are synchronized using RF pulses. In order to estimate the orientation of the person, the signal from two transmitters can be used.

The Cricket system presented in [6] is also using a combination of ultrasonic and RF pulses. Transmitters placed in the environment periodically send their location as a radio signal. Receivers carried by the user can measure the delay in receiving the ultrasonic pulse and thus estimating person's position.

Computer vision solutions rely on the detection of a badge carried by the user or by face detection. The TRIP system described in [7] uses cameras to recognize circular identifying both the pattern and the location and orientation of the person with respect to the camera. The pattern encoded on the circular sectors can be used also for person identification.

In [8] and [9] two different solutions based on face detection are proposed. In [8], they use a stereovision system in order to make a 3D reconstruction of the face, and subsequently to track and recognize the face pose, thus estimating the user's focus of attention. In [9], a stereovision system enhanced with infrared technology is

used for head pose recognition. The tracking system provides a dynamic update of template images for tracking facial features (mouth, eyes) and thus to estimate head pose.

In this paper, we describe a new method based on computer vision to detect the presence of a person at sub-room level. Our camera is also enhanced with infrared technology. The presence is determined by locating a passive badge carried by a person. By measuring whether the distance and orientation of the badge to the camera is below a critical threshold we wish to assess whether the user's focus of attention is drawn to an object of interest.

The paper is structured as follows. Next section describes the solution proposed. Afterwards, we present and discuss the obtained results. Following section will discuss potential applications of our system. Final section summarizes our conclusions regarding this research and presents the guidelines for future work.

DESCRIPTION OF THE SYSTEM

Overview

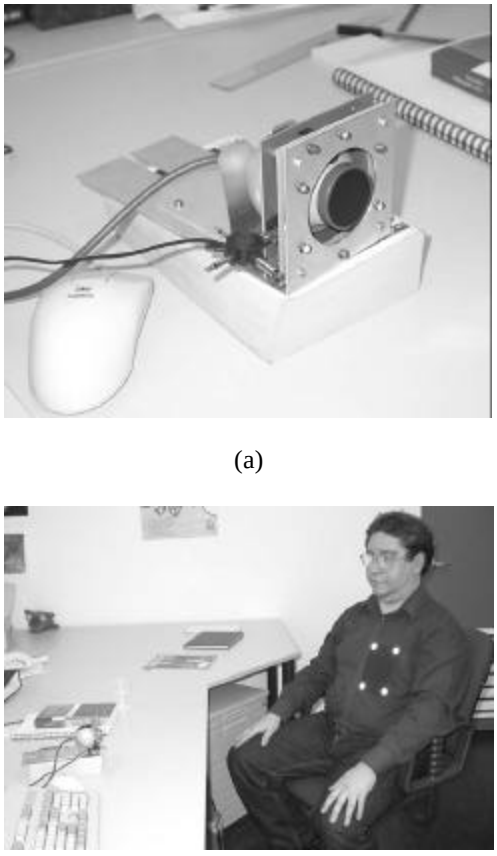
Our system consists of two parts (see figure 1):
- a perceiving component, represented by a Logitech™ webcam with an infrared filter and an array of infra-red LEDs placed around the camera in form of a ring;
- a passive component, represented by a badge, which has reflector tape patches attached on it.

The role of the filter is to let pass only those frequencies of the light that are close to the infrared rays spectrum. By using this kind of filter, the camera will perceive mostly the light that is reflected from these reflector patches. In consequence, background information is discarded from the beginning, making the image analysis process much simpler.

The badge is a rectangle made of rigid paper and has attached patches of reflector tape (in shape of discs), on each of its four corners. A very important property of this material is that it reflects the infrared light back, on the same direction it came from the source (infrared LEDs). The size of the badge is 10x15 centimetres and the discs have a diameter of 2 centimetres.

Existing solutions for depth estimation

Usually, in order to estimate the 3D coordinates of a point in space, a stereovision system is needed. Thus, the 3D coordinates are estimated based on the pixel disparity, i.e. the difference in object pixels' location in the pair of images captured by the two cameras. One of the disadvantages of using a stereovision system (besides its higher cost and necessity for specific hardware) is that accidental changes in the orientation of one of the cameras (which is very likely to happen in a dynamic environment like home or office), requires a recalibration of the whole system.



(a)

(b)

Figure 1. The two components of our system: (a) a Logitech webcam provided with an IR-light source and IR filter and (b) the target with IR reflector material

As an alternative, there is also the possibility to estimate the 3D coordinates of an object using a single camera. Several techniques exist, some of them being reviewed below: range imaging [10] and [11], focus/defocus [12] and inverse perspective transform [13].

The technique based on range images (also called depth maps) implies the existence of structured lights, i.e. scenes illuminated by a known geometrical pattern of light. To compute the depth of at all the points in the

image, the scene is illuminated one point at a time in a two-dimensional grid pattern. The 3D object coordinates are calculating computing the intersection of the camera's line of sight with the light plane.

The techniques based on focus/defocus consist of the existence of several images taken under different lens parameters. The image is modelled as a convolution of focused images, with a function determined by the camera parameters and the distance of the object from the camera. The depth is recovered by estimating the amount of blur in the image.

None of the above techniques is suitable for our problem. The first one, because our system is assumed to work in a dynamic environment with continuously changing illumination conditions. The second one because it is not a solution for real-time applications.

The inverse perspective transform allows the estimation of the 3D coordinates based on the information about the points and lines whose perspective projection we observe. Knowledge about the model of the object and relations with the perspective geometry constraints can often provide enough information to uniquely determine the 3D coordinates of the object.

Proposed solution

Following the inverse perspective transform approach, we found that in [7] for instance, a circular badge worn by subjects is located by the vision system. The distance to the target is estimated by fitting an ellipse around the projection of the target on the image plane and then by back-projecting this geometrical shape into its actual circular form, of known radius.

In our case, we use another technique described in [13], which allows the 3D reconstruction based on the observed perspective projection of two parallel line segments. The method presented below, does not use any information regarding camera's orientation. We always express the relative position of the person with respect to the camera. In consequence, small modifications in the camera position will not affect system's performance.

In a pre-processing step, we first binaries the input image (applying a fixed threshold) in order to segment the blobs corresponding to the infrared reflector patches from the background. After that, we apply an opening morphological filter in order to remove the "salt and pepper" noise that eventually exists after the binarization step. On the filtered image, we apply a connected component analysis to detect the centre of the four blobs.

In figure 2 we give a graphical representation of the 3D badge position and its projection on the image plane.

The 3D coordinate systems is represented by *(X,Y,Z)* axis, while 2D coordinate system of the image plane is denoted by *(u,v)*. Let us consider for instance the pair of parallel lines connecting the centres of the blobs that are situated along the vertical axis (height of the badge), *AC*

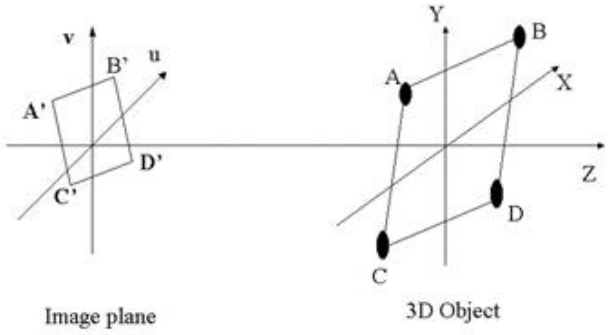


Figure 2. 3D position of the badge and its projection on the image plane

and *BD*. First, we calculate the cosines direction (m_1,m_2,m_3) of the lines segments *AC* and *BD*, i.e. the angle that is formed with each of the three axis *(X,Y,Z)*. Let be $A(x_1,y_1,z_1)$ the 3D coordinate of one end of the line segment and $C(x_2,y_2,z_2)$ the 3D coordinates of the other end. The length of the *AC* line segment is *L*. By $A'(u_1,v_1)$ and $C'(u_2,v_2)$ we denote the projected coordinates (in the image plane) of these two points. Then, the following equations are used to retrieve the 3D coordinates of the first point of the line segment (*f* – represents the focal length of the camera).

$$z_1 = \frac{L\left[(u_2-u_1)(fm_1-u_2m_3)+(v_2-v_1)(fm_2-v_2m_3)\right]}{(u_2-u_1)^2+(v_2-v_1)^2}$$

$$x_1 = \frac{u_1}{f} z_1 \qquad (1)$$

$$y_1 = \frac{v_1}{f} z_1$$

Then, the 3D coordinates of the other end of the segment line can be calculated straightforward:

$$\begin{pmatrix} x_2 \\ y_2 \\ z_2 \end{pmatrix} = \begin{pmatrix} x_1 \\ y_1 \\ z_1 \end{pmatrix} + L \begin{pmatrix} m_1 \\ m_2 \\ m_3 \end{pmatrix} \qquad (2)$$

In a similar way, the 3D coordinates of the centres of the remaining 2 blobs can be calculated. This approach is valid under rigid body transformations assumption [10], that the object can change its position or orientation, without changing its size and shape.

EXPERIMENTAL RESULTS

Since we looked for a low-cost solution to our problem, we tested the algorithm on a PC of modest performances with 128 MB of RAM and a processor of 730 Mhz. We set the frame rate of the webcam at 10 frames/sec, which is acceptable for real-time tracking. Our webcam is not specifically infrared-sensitive, and in order to get an image where the reflecting patches are highlighted we needed to increase the duration of the exposure [10]. The image size was 640x480 pixels.

The experiments performed were intended to assess the accuracy of the distance measured by the proposed algorithm. In the case of infrared technology, the only factor that can affect the performances of the algorithm described above is the amount of infrared radiation presented in the ambiance. We didn't have access to a spectrometer in order to measure this level of radiation. But empirical experiments demonstrated the robustness of our system in case of diffused natural light, considered during different moments of the day, and also artificial light. In our view, these are the most likely scenarios in an office or home environment. We found that only direct sunlight, presented in the scene covered by the camera, can affect system's performance.

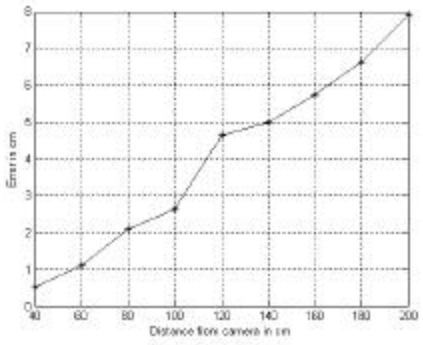
We estimated the accuracy of the distance measured when the target was positioned at orientations of 0, 30 and 45 degrees respect with the camera. The distance range was set between 40 cm and 2 m. For most applications inside a home or a small office space, it can reasonably be expected that the threshold distance relating to when a person is paying attention to a specific device, should fall within the mentioned range. The low-end of the distance range would corresponds when the user is in front of a PC, while the high-end would correspond for the case of "wall mounted display".

We collected the distance calculated by the proposed algorithm over a sequence of 350 frames, separately, at several distances and under several orientations of the target.

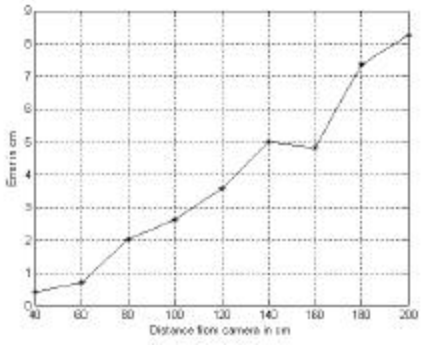
Figure 3 shows the error average in distance measured under the mentioned conditions. The distance D to the target is expressed by averaging the z-component of the four points corresponding to the centre of the four reflecting patches attached to the badge and estimated as has been presented in the previous section:

$$D = \frac{z_1 + z_2 + z_3 + z_4}{4} \qquad (3)$$

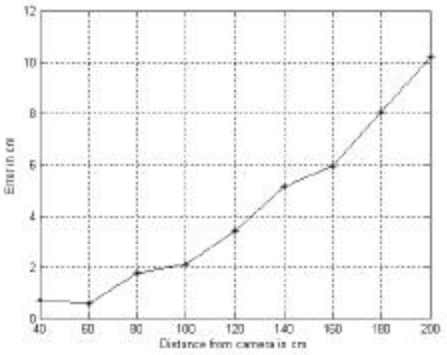
It can be appreciated from the below images, that the error in distance measurement at 2 meters is about 4%.



(a)



(b)

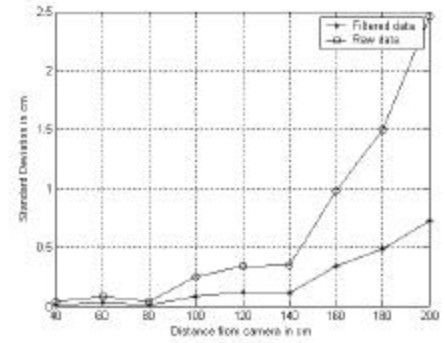


(c)

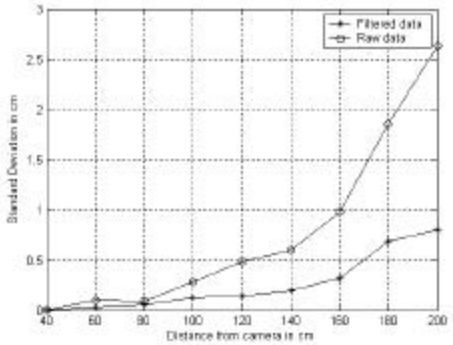
Figure 3. The distance accuracy estimated over 350 frames. Graphics (a), (b) and (c) correspond when the pattern is successively positioned at 0, 30 and 45 degrees with respect to the camera

Besides the absolute error (in cm) calculated in reference to the measured distance, we also estimated the standard deviation of the data collected for each distance. Due to sensor inaccuracy or other errors, the initial results present a large variation. A Kalman filter [14] was applied in order to smoothen the variable calculated (distance) and to reduce its variance. Figure 4 shows the standard

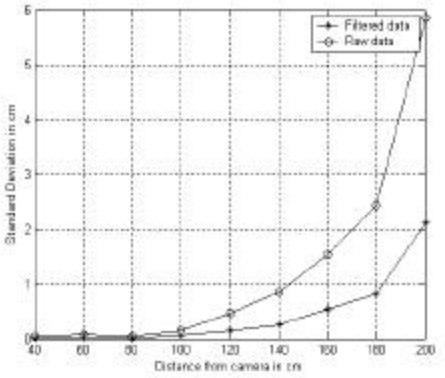
deviation of our data before and after the application of the filter. The improvements introduced by the application of Kalman filter are obvious.



(a)



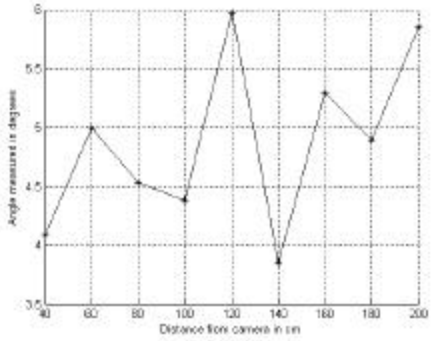
(b)



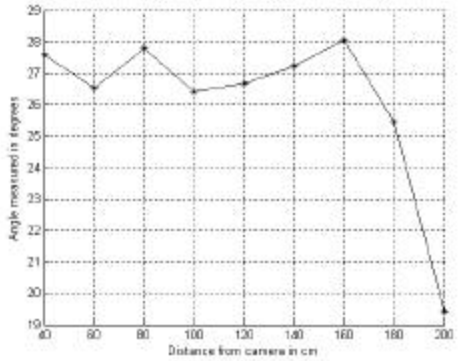
(c)

Figure 4. Standard deviation for the distance measured before and after the filtering when the pattern was positioned at 0, 30 and 45 degrees respect with the camera

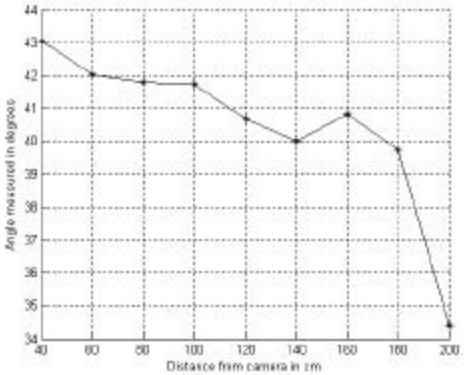
In order to express the orientation of the badge with respect to the camera, we calculated the rotation angle around the *Z* axis (we assumed that the camera and the badge are more or less at the same level). This angle, denoted by 'κ' is expressed in terms of the dot-product between the normal to the image plane and to the badge, respectively, as shown in the following formula:



(a)



(b)



(c)

Figure 5. The angle average accuracy (in degrees) estimated over 350 frames. Graphics (a), (b) and (c) correspond when the pattern is successively positioned a 0, 30 and 45 degrees with respect to the camera

$$\cos k = \frac{N_1 \bullet N_2}{\|N_1\| \cdot \|N_2\|} \quad (4)$$

The operator '•' denotes the dot-product between the normal vector N_1 and N_2, while the operator ||…|| expresses the norm of the vectors N_1 and N_2, respectively. Figure 5 shows the average error in angle measurements (expressed in degrees) under the mentioned conditions.

It can be seen that in general the error in angle measurement is around 5 degrees or less. The severe drop in angle precision that occurs at distances beyond 160 centimetres can be explained that starting with this range, the image losses resolution, and thus it present less 'structure'. The lost of resolution leads to an increment in the error of estimation of blob centres, and subsequently, in the calculation of angles and distances between them.

Discussion

The current implementation of the system allows only one person to be detected. This can be seen as an advantage, since multiple users in front of the camera would create confusion to the system, at the moment to whom it should interact with. The detection range (up to 2 meters) is intended to limit the operational area of the "smart device" in a neighborhood centred on it (otherwise, a bigger range could create ambiguity in response, if several "smart device" are close to each other). The purpose is to become aware, only when somebody is really close to it. It is assumed that the camera is placed in such a position in order to optimise the interaction between the user and the device that is attached to. We want to avoid situations in which, for objects of very large dimensions (like wall-mounted displays, for instance), when the user is facing towards the object, the user's presence (the badge) cannot be detected by the camera. Similar approach has been taken in the applications reported in [8] and [9].

With the current design of the badge, the users behave in an anonymous way (they are indistinguishable by the system). In some cases, anonymity may be preferred. This is the case of systems running in public areas (like information kiosks placed in museums for instance). In other cases, (like home environments) person identification is preferred in order to guarantee a personalized interaction with the user. The current limitation of the badge can be overcome, by extending it with another one, having encoded (through a number of dots) and thus giving an relative identity to the person who carries it.

For a better assessment of the performances of our system, we compared it with some existing locating systems employing different technologies. Since our approach is aimed to work for resolutions of sub-room scale granularity, the comparison is done with systems that fall within this category. The ultrasonic location systems like those described in [4], [5] and [6] have accuracy in distance measurement of 2-3 cm and 3 degrees in angle measurement. The system based on computer vision techniques [7] has an accuracy of about 6 cm in distance and 3 degrees in orientation.

Compared with other computer-vision based systems [7], our approach has the advantage that can operate under very different illumination conditions. Even under very poor lightning (that can be a perfectly realistic scenario in a home-based environment), our system proof to be very robust.

APPLICATIONS

With the system proposed in this paper, we implemented a prototype of a "smart device", namely an "aware display" endowed with visual perceiving capabilities (see figure 6) through the webcam embedded on the bottom. This device is envisioned to enclose a single-board PC [15]. By 'miniaturizing' its dimension, we pretend to create a "basic cell" for ubiquitous computing environments, transparent from the point of view of the user.

Figure 6. A "transparent" representation for an ubiquitous computing component: a touch-screen enhanced with perceptive capability due to the embedded webcam

With several units like the one from figure 6, we are currently developing a messaging application for a ubiquitous computing scenario. We pretend to have connected several "aware displays" (installed in different

locations in our laboratories) to a message delivery server. Each time an "aware display" perceives the presence of a person in its vicinity, it sends a notification to the server. When a new message is available, it is displayed on the unit that sent the notification. This way, the messages are always shown on the monitor that is closest to the user at a given moment, without the necessity, from part of the user, to go to a specific location to check for new messages. In figure 7 we depict a sketch of the intended application.

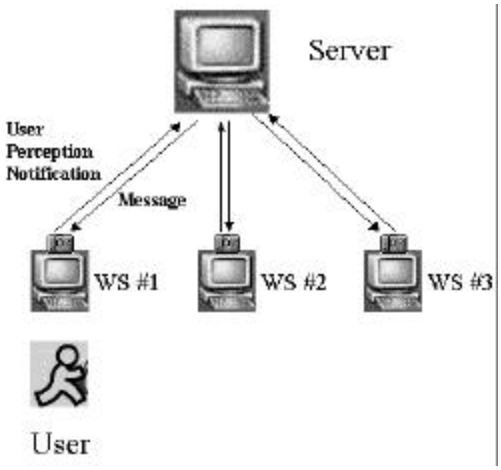


Figure 7. A sketch of the system architecture used for the distributed messaging application

We consider this is a realistic scenario, taking into account that is very common that a person can be present, throughout the day in several locations, in a building, not only in his/her office. As an example, we can refer to the university environment, where the researchers, besides their office, often has to go to a lab to do some experiments or have to attend a discussion session in a meeting room.

On the other hand, this application shows that the creation of an aware environment can be addressed incrementally, starting with one perceiving device and dynamically add others, as they become available.

CONCLUSIONS AND FUTURE WORK

In this paper we proposed a new, low-cost solution to detect user's presence at sub-room level resolutions. This approach presents a high robustness against varying lightning conditions. The detection range is between 40 centimetres and 2 meters, which makes it suitable for a large variety of applications. In consequence, this will allow a redefinition of the term "near", depending on the context the application will be developed for. A prototype of a "smart device" making use of our system was presented, together with an example of a messaging application that is currently under development.

While the presented method gave some very encouraging results, it obligates the person to wear the badge attached to his clothes. In everyday context this can be an onerous obligation for the user. On the other hand, it provides a direct mechanism to the user to control when his/her activities are monitored and responded to, by simply adding or removing it.

There are several alternative technologies to detect user proximity, e.g. using ultrasound signals. These approaches can work accurately in domestic environments, but interference with other electronic devices could affect their robustness. However, computer vision is better suited for the specific problem of detecting the direction the user is facing in. In our next step we shall investigate the feasibility of detecting user's attention without the need for reflecting badges, by directly detecting the human face and head pose in the scene.

ACKNOWLEDGMENTS

The authors want to thank to Martin Boschman for the design of the infrared circuitry and to Charles Mignot for making the "aware display".

REFERENCES

1. Abowd G.D, Mynatt E.D., "Charting Past, Present and Future Research in Ubiquitous Computing", *ACM Transactions on Computer-Human Interaction*, 7(1):29-58, 2000
2. Brumitt B., Cadiz J.J., "Let There Be Light: Comparing Interfaces for Homes of the Future", *Proceedings of Interact'01*, Japan, pp. 375-382
3. "Aware Home Research Initiative", Georgia Institute of Technology, *http://www.cc.gatech.edu/fce/ahri/projects/*
4. Ward A., Jones A., Hopper A., "A new location technique for the Active Office", *IEEE Personal Communications*, 4(5):42-47, 1997
5. Hazas M., Ward A., "A Novel Broadband Ultrasonic Location System", *Proceedings of Ubicomp2002*, pp. 264-280, Sweden, 2002
6. Priyantha N.B., Miu A.K.L., Balakrishnan H., Teller S., "The Cricket Compass for context-aware mobile applications", *Proceedings of Mobicom2001*, pp. 1-14, Italy, 2001

7. de Ipina D.L., Mendonca P.R.S., Hopper A., "TRIP: A Low-Cost Vision-Based Location System for Ubiquitous Computing", *Personal and Ubiquitous Computing Journal*, 6(3):206-219, 2002
8. Darrell T., Tollmar K., Bentley F., Checka N., Morency L.-P., Rahimi A., Oh A, "Face-Responsive Interfaces: From Direct Manipulation to Perceptive Presence, *Proceedings of Ubicomp 2002*, pp. 135-151, Sweden, 2002
9. Nakanishi Y., Fujii T., Kiatjima K., Sato Y., Koike H., "Vision-Based Face Tracking System for Large Displays", *Proceedings of Ubicomp2002*, pp. 152-159, Sweden, 2002
10. Jain R., Kasturi R., Schunck B.G., "Machine Vision", *McGraw-Hill*, New York, 1995
11. Davies E. R., "Machine Vision: Theory, Algorithms and Applications", *Academic Press*, San Diego, 1997
12. Xiong Y., Shafer S., "Depth from Focusing and Defocusing", *Technical Report CMU-RI-TR-93-07*, Robotics Institute, Carnegie Mellon University, 1993.
13. Haralick R.M., Shapiro L.G., "Computer and Robot Vision", *Addison-Wesley*, New York, 1993
14. Welch, G., Bishop G., "An Introduction to the Kalman Filter", *Technical Report TR 95-041*, University of North Carolina, 2002
15. Workbox Computer: *http://www.zerez.com/producten/workboxp3/techspecs.htm*