INTECH
open science | open minds

OPEN ACCESS
ARTICLE

International Journal of Advanced Robotic Systems

Interactive and Audience Adaptive Digital Signage Using Real-Time Computer Vision

Regular Paper

Robert Ravnik[1,*] and Franc Solina[1]

1 Faculty of Computer and Information Science, University of Ljubljana, Slovenia
* Corresponding author E-mail: robert.ravnik@fri.uni-lj.si

Received 27 Jun 2012; Accepted 14 Dec 2012

DOI: 10.5772/55516

© 2013 Ravnik and Solina et al.; licensee InTech. This is an open access article distributed under the terms of the Creative Commons Attribution License (http://creativecommons.org/licenses/by/3.0), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.

Abstract In this paper we present the development of an interactive, content-aware and cost-effective digital signage system. Using a monocular camera installed within the frame of a digital signage display, we employ real-time computer vision algorithms to extract temporal, spatial and demographic features of the observers, which are further used for observer-specific broadcasting of digital signage content. The number of observers is obtained by the Viola and Jones face detection algorithm, whilst facial images are registered using multi-view Active Appearance Models. The distance of the observers from the system is estimated from the interpupillary distance of registered faces. Demographic features, including gender and age group, are determined using SVM classifiers to achieve individual observer-specific selection and adaption of the digital signage broadcasting content. The developed system was evaluated at the laboratory study level and in a field study performed for audience measurement research. Comparison of our monocular localization module with the Kinect stereo-system reveals a comparable level of accuracy. The facial characterization module is evaluated on the FERET database with 95% accuracy for gender classification and 92% for age group. Finally, the field study demonstrates the applicability of the developed system in real-life environments.

Keywords Face Localization, Digital Signage, Computer Vision, Information Interfaces

1. Introduction

Digital signage flat-panel displays are emerging as a new, efficient method for providing targeted information [1,2]. They are found in airports, hotels, universities, retail stores and various outdoor public spaces (Figure 1), all providing optimized information-and-appearance attractive multimedia content. Today, a large majority of the applications of digital signage are interfaces to public or internal information, advertising, brand building and influencing the customer's behaviour by enhancing the customer's experience [3].

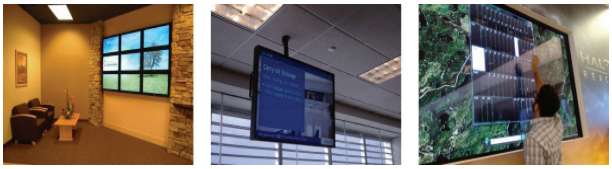

Figure 1. Digital signage displays are permeating public spaces and replacing static signs.

Modern studies of digital signage are actively exploring mechanisms for engaging users with interactive content [4,5]. Various interaction modalities have been proposed, including speech, facial expression, gaze, touch and hand gestures [6]. Interactive digital signage is appearing in urban life and architecture [7] as well as ubiquitously in computing [1]. Ojala *et al*. describe the problem of *Interaction blindness* [8], where observers do not interact with displays because the interaction process is too complex.

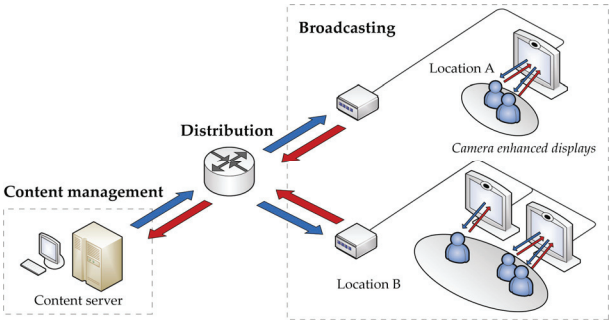

Figure 2. Schematic of the developed interactive and audience-adaptive digital signage system. Media content is managed at the central content server and then dispatched via a local or global network to each broadcasting location. A camera enhanced display tracks the observers and their characteristics, and broadcasts adaptive content in real-time, possibly at multiple locations A and B. Arrows denote a circular information flow which notably differs from the one-way information flow in common digital signage systems.

Digital signage displays are advantageous compared to static signs because they can display varying multimedia content such as images, animations, video and audio. Content can be changed in real-time, which, in principle, allows for full context and audience adaptation [9]. However, the high potential of digital signage displays has not yet been fully exploited, as the displayed content is most often generic and uninteresting for observers, causing the *Display Blindness* [10] effect. To make digital signage a more effective information interface, the displayed content should be informative, dynamic and attractive [11].

Such interactive audience adaptive digital signage systems have the potential to be applied to a large number of displays. Therefore, the hardware cost of a single display is important as well as the cost of software solutions, especially the use of algorithms without copyright limitations. These considerations were our main motivation for using a monocular camera in our system for the interaction of the system with the observers, as it is very frequently already built into the frame of flat digital displays. Alongside this, we applied reliable state-of-the-art computer vision solutions, centrally optimizing their advanced functions, real-time processing speed and ease of integration, as well as their license free implementation.

The outline of the paper is as follows: Section 2 presents the architecture of our interactive and audience adaptive digital signage system, Section 3 evaluates the performance of the system in a laboratory setting, Section 4 describes the use of the system in a real-world audience measurement study and Section 5 gives a discussion and conclusions.

2. Interactive and audience adaptive digital signage

To address the problems of display and interaction blindness, we designed our digital signage system to oversee and interact with the presence, activity and characteristics of the observers. A scheme of the proposed system is given in Figure 2.

We propose a novel and reliable spatial localization of observers using a single monocular camera and measuring interpupillary distance in human faces. Since interpupillary distance in human faces is fairly stable [12], it can serve as the measurement quantifier for determining the position of detected human faces in a 3D space, if the optical parameters of the camera are known.

Digital images captured by the camera are processed with the digital signage software in real-time, extracting temporal, spatial and demographic features of the observers. By comparing the determined features with the predefined content descriptors, the display software automatically selects and broadcasts content relevant to the specific detected observer, for example, this could be information that is targeted to young adult males in the age group 25-34 years.

The proposed system is implemented in C++, using the Qt application framework and OpenCV library [14]. A 24' Sony Vaio VPCL135FX/B computer display enhanced with a Logitech WebCam Pro 9000 camera is used as a broadcasting unit prototype.

Several computer vision methods are combined to achieve the optimal determination of the features of the observer. The advantages of this approach and a performance analysis are discussed separately in Section 3. The computer vision methods are designed into the system architecture, which includes a pipeline of three modules (see Figure 3). The first module detects the observers and determines their temporal (e.g. dwell time) and spatial parameters (e.g. distance from the digital signage system). The second module determines the observer's demographic information from the registered facial image, i.e. the gender and the age group of the observer. Finally, the combined features determined by Modules 1 and 2 are used by Module 3 to broadcast observer-profile-specific content on the digital signage display. Below, each module is separately presented and the privacy aspects of the system are explained.

2.1 MODULE 1: Temporal and spatial localization of observers

Pre-processing of captured images includes background segmentation. We use Mixture-of-Gaussians based background modelling [13] to extract foreground regions and define the possible presence of observers.

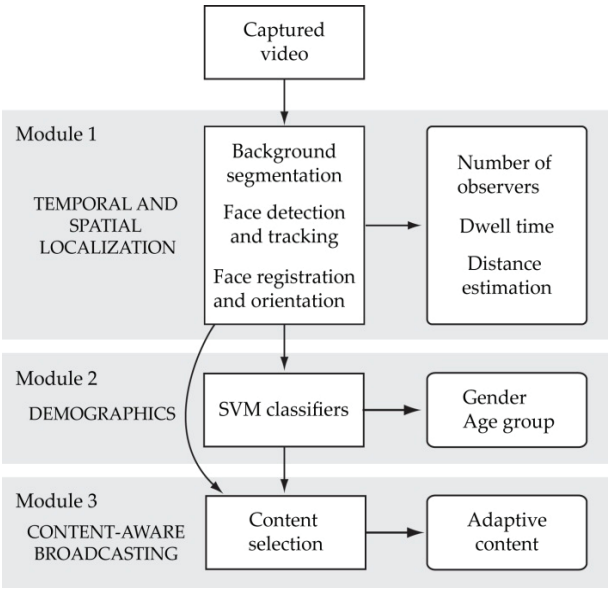

Figure 3. Modules of the developed interactive and audience adaptive digital signage system. Modules 1 and 2 provide localization and demographic information about currently present observers, which is then used by Module 3 for content-aware broadcasting.

The number of observers and their presence time are estimated by the face detection and tracking algorithms. The Viola & Jones face detection algorithm [14] is applied to the pre-processed foreground regions, to distinguish true observers facing the display from non-observer digital patterns. Using a frontal face detector, we can determine the location of the faces of the observers down to the size of 20x20 pixels per face in real-time, regardless of their actual position and physical scale. The positively identified segmented regions – i.e. the observers - are tracked using a Fast Match Template algorithm, specially adapted for real-time video processing, supplied in OpenCV library [15]. The observer's upper body is used as a template image.

Facial images are registered using the multi-view Active Appearance Model (AAM) method. The AAM simultaneously models the intrinsic variation in shape and texture of a deformable visual object as a linear combination of the basis modes of the variation [16]. Although linear in both shape and appearance, overall, the AAMs are nonlinear parametric models in terms of pixel intensity. Fitting an AAM to an image consists of minimizing the error between the input image and the closest model instance; i.e. solving a nonlinear optimization problem [17]. More specifically, the face registration determines the position of 66 facial feature points, where for example, each eye is described with 6 feature points that form a convex polygon around the eye orbit. The centroid of this polygon is calculated in order to determine the centre of the observer's eye. We denote the centroid points of the left eye and the right eye as *E1* and *E2*, respectively. Most adult human faces have an almost identical interpupillary distance (*IPD)* [12] and we use it as a constant parameter to estimate the observer's distance from the digital signage display. According to [12] more than 90% of adults have an *IPD* between 57mm and 69mm, with a mean of ~63mm. Alternatively, in the digital image analysis, the *IPD* corresponds to the Euclidean distance between *E1* and *E2*:

$$IPD = ||E1 - E2||$$

and is inversely proportional to the distance between the face and the camera. Therefore, we can use the following distance estimation function $\hat{F}_{dist}$:

$$\hat{F}_{dist} = Ax^{-1} + B$$

where x is the estimated *IPD* in pixels, and *A* and *B* are camera specific constants that depend on the field of view of the camera lens and on the resolution of the camera.
In this work, we use a Logitech WebCam Pro 9000 camera with a horizontal field of view of 63.1°, vertical field of view of 49.4° and an image resolution of 1600x1200 pixels. The parameters *A* and *B* were determined to be *A*=21528.8 and *B*=7.78 using standard numerical packages and benchmark-testing on exemplary images. A more detailed evaluation of the localization module is presented in Section 3.

2.2 MODULE 2: Obtaining demographic features

The demographic features of age and gender are determined according to 7 age groups: 1-14, 15-24, 25-34, 35-44, 45-54, 55-64 and over 65 years, which are all either male or female. We use the AAM facial registration method of Module 1 to register a face and warp it to the normalized frontal form of 50x50px in size. The FERET database [18] is used to train Support Vector Machines (SVM) as classifiers of gender and age.

2.3 MODULE 3: Content-aware broadcasting

In order to formalize the adaptive content selection procedure we defined groups according to (i) broadcasting features and (ii) audience features.

Broadcasting features BF describe various temporal, location-specific and other screening parameters of the digital signage system such as: hourly, daily or weekly time scheduling, screening frequency, time of last appearance and content's broadcasting history and selection of specific groups of displays based on their location.

Audience features AF include the audience's demographic parameters, such as gender and age group, and other temporal (dwell time) and spatial parameters (distance) associated with the observers.

Upon setting up the digital signage broadcasting, each item of broadcasting item is described with a content descriptor *CD*, which is chosen as a set of broadcasting and audience features, which are interlinked. Content descriptor *CD* is introduced in line with [19] as:

$$CD = \{\{\alpha_i * AF_i\} \cup \{\beta_j * BF_j\} : i, j \geq 0\}$$

where α_i and β_j are the weights of each audience and broadcasting feature. Alongside this, the state descriptor SD is a set of audience features that describe audience characteristics in front of a single display *at a given time* and can be introduced as:

$$SD = \{AF_k : k \geq 0\}$$

where k covers all considered audience features for the present observers.

Finally, to choose the actual content item to be broadcasted j_{next}, out of n available content items, we determine the minimum of the distance function *dist* between *SD* and CD_j:

$$j_{next} = \arg\min_{j \leq n} dist(SD, CD_j)$$

For *dist* various distance functions or machine learning algorithms can be used. We use the k=3 nearest neighbours method, as it proves to give efficient and reliable results.

2.4 Privacy aspects

The architecture of our interactive and audience adaptive digital signage system is designed according to the Privacy-by-design [20] principles, to ensure the fully appropriate handling of personal data. Image capturing and processing is performed by the display unit in real-time, therefore no visual records are stored or distributed over the network. Video images of actual observers are discarded immediately after processing, storing only the observers' audience *features* (e.g. demographics) which are sent to the central server for statistical analysis via an encrypted data transfer. At the actual location of the digital signage system, the observers are notified of the video-monitoring, in compliance with national privacy legislation.

3. Laboratory evaluation of the system

The performance analysis of the proposed digital signage system was first conducted in a laboratory study. The modules were monitored separately to obtain clear information about their performance.

Module 1 performance: The accuracy of the localization module was determined by performing a tracking experiment. As a widely available reference benchmark, we compared the localization results with the outputs of the commercially available Kinect RGB-D sensor. Kinect is seen as a possible platform for advanced user interfaces [5]. Since Kinect's official nominal depth range is only up to ~4m we used the OpenNI framework that returns depth ranges of up to 9m [21], which are common in our developed IPD system.

Five observers were asked to walk along a straight path, facing the camera of our IPD system and the Kinect (aligned and positioned one above the other). Markers were placed on the floor along the path at 10 cm intervals from 0.5m to 8m in distance. When the observer reached a marker, the estimated distance was recorded by both systems. Figure 4 shows the root mean square error (RMSE) of distances obtained by this experiment.

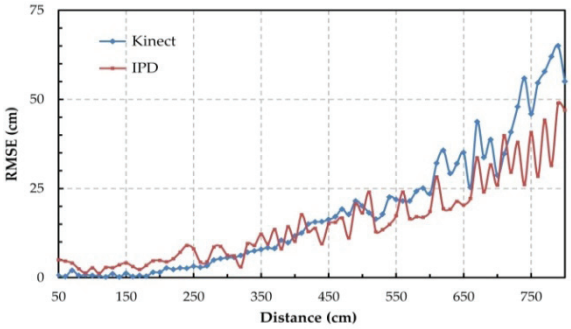

Figure 4. Root mean square error (RMSE) of the IPD based (red) and Kinect (blue) distance measurement systems increase with distance but stay in general agreement.

The comparison of the Kinect and the IPD estimator shows an 18.7cm and 19.3cm overall RMSE, respectively. Kinect gives stable results up to a distance of 4m with error increasing exponentially with distance in agreement with [21]. The relative standard error of our IPD estimator for a given distance range is 4.3%. On average, the IPD distance estimator gives results almost equivalent to the Kinect stereo system and at distances of over 4m even more stable results, which is a clear indication of the accurate performance of Module 1.

Module 2 performance: Module 2, which determines the demographic features of the observers from the aligned facial images, was evaluated on the FERET database [18]. The database includes annotated facial images with corresponding gender and year of birth data of 856 individuals. Feature vectors of aligned frontal faces were used as a classification input. Several machine learning algorithms were inter-compared using 10-fold cross validation, including Classification tree, Random forest, SVM and k-NN. The results are presented in Table 1.

	Classification accuracy	
Algorithm	Gender	Age group
Classification tree	0.924	0.893
Random forest	0.931	0.909
SVM	**0.952**	**0.917**
k-NN	0.875	0.823

Table 1. Comparison of evaluated machine learning methods for gender and age group classification within Module 2.

The highest classification accuracy was achieved using SVM for both gender and age group classification resulting in a classification accuracy of 95.2% and 91.7%, respectively. Following this accuracy study, the SVM method is used in our system as the main method for gender and age group classification.

Module 3 performance: Content-aware broadcasting was tested in a laboratory field study with a limited group of selected observers, which offered a more controlled environment, i.e. observers, locations and demographics were controlled in the testing of the developed modules, which are the main focus of this paper. Under laboratory conditions the performance measurements were fully reproduced and repeated, giving a good bench-marking system ready for full implementation in a real-world environment.

4. Field application of the system

We used the developed system in a real-world environment for an audience measurement study [22] which is part of a larger marketing study. With the described system we measured the attention time that shoppers were giving to a digital signage display in a clothing store. Normally, such studies rely on qualitative assessments based on interviews or questionnaires [9-11].

The described system, however, provided full quantitative data on the number, age and gender of the visitors, as well as their attention time based on the observations of the digital signage display. The accuracy and volume of the data collected with our system is almost incomparable with traditional methods of market analysis. This creates the possibility of analyzing the collected data with machine learning methods. Besides gathering data in relation to the display blindness hypothesis [10] in general, the goal of the audience measurement study was also to determine what type of information (for example, static or dynamic, such as video) attracts higher attention.

The digital signage system was installed in a clothing boutique in Ljubljana, Slovenia. The shop sells higher priced sports fashion and apparel, which can affect demographic and behavioural characteristics. The floor plan of the store consisted of a main area situated between the entrance and the cashier's desk, with additional room in the back used for changing. To make the most out of the floor plan, the display was mounted at eye-level on a special shelf next to the cashier's desk, directly facing the entrance (see Figure 5a).

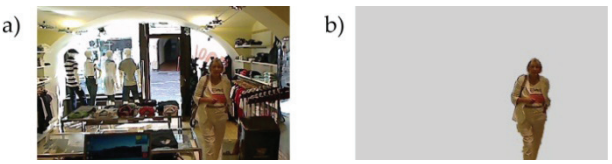

Figure 5. Real-life audience measurement study with the developed system. a) A typical image captured by the digital signage screen unit. b) Image after segmentation described in Section 2.1.

The audience measurement study was performed in 23 daily sessions, recording a total of 214 hours. The system detected 1294 people and determined their gender and age group as well as their attention time given to the display.

The audience measurement study reveals that 61% of all detected customers in the store were female and 39% were male. The age distribution of customers was as follows: 7% in the category 1-14 years, 10% in 15-24 years, 20% in 25-34 years, 25% in 35-44 years, 19% in 45-54 years, 12% in 55-64 years and 7% in the category 65+ years. Dwell time for each customer is the time they spend in the same room as the display. Attention time is the time that each customer spends looking at the digital signage display. The attention time quantifier reveals, that, on average, men pay attention to the digital signage display for 1.2s, whereas women only 0.4s. Age group comparison shows that attention time to digital signage is the highest (2.4s) in the age group for children (1-14 years) while the average attention time of all customers is 0.7s. Interestingly, the average attention time is lowest in the 35-44 years age group (0.42s). The content quantifier, dynamic or static, shows that broadcasting dynamic and not static digital signage content increased attention time by 43%.

We performed a correlation analysis of the audience measurement data acquired with the described digital signage system using Spearman's rank correlation coefficient ρ [23]. We selected this correlation measure since it can be used with ordinal and continuous variables, deals well with non-linear dependencies and does not require normally distributed data. The distance map between all pairs of attributes of our field study was calculated using the absolute Spearman's dissimilarity ρ_d measure:

$$\rho_d = \frac{1 - |\rho|}{2},$$

where ρ denotes Spearman's rank correlation coefficient. The results are presented in Figure 6.

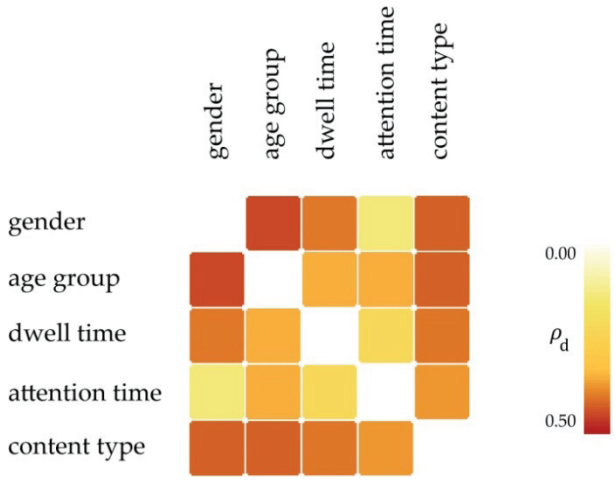

Figure 6. Distance map of audience measurement data using absolute Spearman's dissimilarity ρ_d measure. Higher values of ρ_d correspond to larger distances between attributes, whereas lower values imply similarity of attributes.

As expected, the distance map reveals large distances among gender, age group and content type since they are independent. Small distances between attention time and independent variables (gender, age group, content type) confirm statistical dependency and this validates trends observed in the audience measurement study.

The described algorithm-determined audience measurement parameters (gender, age group, dwell time and attention time) were, for the purpose of evaluating the precision in the marketing study, that required also some additional parameters describing customers that arrived in groups, evaluated also manually. The comparison of algorithm- and manually determined parameters from the described field study offers an exact means with which to determine the accuracy of the automatically determined parameters, i.e. the full performance of the digital signage system. Indeed, the comparison shows that the system performs with a high level of accuracy, giving gender classifiers 86.6% and age classifiers a classification accuracy of 77.1%.

Full system performance: Finally, the performance benchmark of our prototype broadcasting unit shows that the system is capable of video processing at 21 FPS using two cores of Intel Q8400 (2.66Ghz) processor making it suitable for broadcasting adaptive content in real-time.

5. Conclusion

In conclusion, we have described the design, implementation and use of an adaptive digital signage system. Based on functional requirements of an audience adaptive digital signage system which can perform in real time, we identified its components and overall architecture. Three main modules were designed and developed: (i) Module 1 for the spatial and temporal localization of the observers, (ii) Module 2 for the demographic features of the observers and (iii) Module 3 for content-aware broadcasting. The accuracy of the developed system was evaluated in a laboratory study achieving spatial accuracy in the tracking of observers with a relative standard error of 4.3%. With our monocular spatial localization module, we achieved results comparable to the distance estimates obtained by the stereo-based Kinect. In this way we demonstrated that it is possible to reliably determine the distances of people with a single monocular camera.

We assessed the performance of our system also in a real world environment in the context of a marketing study. The goal was to determine the attention time which visitors of a clothing store paid to the digital signage display. The digital signage system demonstrated accurate and superior quantitative collection of data regarding shoppers, otherwise typically obtained only qualitatively via marketing studies.

License-free computer vision and machine learning algorithms that operate in real-time on low-price hardware were selected. All of the algorithms and software components that we used are copyright free, which makes the proposed architecture even more suitable for practical implementation.

Clearly, more complex hardware enhancements of the standard display, e.g. infrared sensors or multiple cameras for stereo vision, could lead to even more accurate results, but an integral part of our design was also good price-to-performance and simple implementation. This choice is even more rational because displays with built-in monocular cameras are becoming widely accessible. The use of more complex algorithms could improve observer tracking and classification accuracy but would at the same time require more processing time or more processing power. The price-performance trade-off of future interactive signage systems will probably be determined only after more experience is gained with such systems.

We believe that the main contribution of this paper is an operational, cost-effective interactive signage system where the viewing statistics and interaction with the audience are achieved with efficient and real-time capable computer vision techniques. We hope that this will contribute to the future development and design of intelligent digital signage systems.

6. Acknowledgments

This work was supported by the Slovenian Research Agency, research program Computer Vision (P2-0214).

7. References

[1] Krumm J (2011) Ubiquitous advertising: The killer application for the 21st century. IEEE Pervasive Computing 10:66-73.

[2] Davies N, Langheinrich M, Jose R, Schmidt A (2012) Open Display Networks: A Communications Medium for the 21st Century. IEEE Computer 45(5): 58-64.

[3] Lundstrom LI (2008) Digital Signage Broadcasting: Content Management and Distribution Techniques. Focal Press.

[4] Michelis D, Müller J (2011) The Audience Funnel: Observations of Gesture Based Interaction With Multiple Large Displays in a City Center. Int. J. Hum. Comput. Interaction 27(6): 562-579.

[5] Müller J, Walter R, Bailly G, Nischt M, Alt F (2012) Looking glass: a field study on noticing interactivity of a shop window. In: Konstan JA, Chi EH, Höök K, editors. Conference on Human Factors in Computing Systems. Austin: ACM. pp. 297-306.

[6] Müller J, Alt F, Michelis D, Schmidt A (2010) Requirements and design space for interactive public displays. In: Bimbo AD, Chang SF, Smeulders AWM, editors. ACM Multimedia, ACM, pp. 1285-1294.

[7] Kuikkaniemi K, Jacucci G, Turpeinen M, Hoggan EE, Müller J (2011) From Space to Stage: How Interactive Screens Will Change Urban Life. IEEE Computer 44(6): 40-47.

[8] Ojala T, Kostakos V, Kukka H, Heikkinen T, Lindén T, Jurmu M, Hosio S, Kruger F, Zanni D (2012) Multipurpose Interactive Public Displays in the Wild: Three Years Later. IEEE Computer 45(5): 42-49.

[9] Bauer C, Spiekermann S (2011) Conceptualizing Context for Pervasive Advertising. In: Müller J, Alt F, Michelis D, editors. Pervasive Advertising, Springer. pp.159-183.

[10] Müller J, Wilmsmann D, Exeler J, Buzeck M, Schmidt A, Jay T, Krüger A (2009) Display Blindness: The Effect of Expectations on Attention towards Digital Signage. In: Tokuda H, Beigl M, Friday A, Brush AJB, Tobe Y, editors. Pervasive, LNCS 5538, Springer. pp. 1-8.

[11] Huang EM, Koster A, Borchers J (2008) Overcoming assumptions and uncovering practices: When does the public really look at public displays? In: Indulska J, Patterson DJ, Rodden T, Ott M, editors. Pervasive, LNCS 5013, Springer. pp.228-243.

[12] Dodgson NA (2004) Variation and extrema of human interpupillary distance. In: Woods AJ, Merritt JO, Benton SA, Bolas MT, editors. Proceedings of SPIE: Stereoscopic Displays and Virtual Reality Systems XI. Vol. 5291. San Jose. pp. 36-46.

[13] Bouwmans T, Baf FE, Vachon B (2008) Background Modeling using Mixture of Gaussians for Foreground Detection - A Survey. Recent Patents on Computer Science 1 (3):219-237.

[14] Viola PA, Jones MJ (2004) Robust real-time face detection. International Journal of Computer Vision 57 (2):137-154.

[15] Bradski G, Kaehler A (2008) Learning OpenCV: Computer Vision with the OpenCV Library. O'Reilly Media.

[16] Matthews I, Baker S (2004) Active appearance models revisited. International Journal of Computer Vision 60 (2): 135-164.

[17] Saragih J, Göcke R (2009) Learning AAM fitting through simulation. Pattern Recognition 42 (11): 2628-2636.

[18] Phillips J, Moon H, Rizvi SA, Rauss PJ (2000) The FERET evaluation methodology for face-recognition algorithms. IEEE Trans. Pattern Anal. Mach. Intell. 22 (10): 1090-1104.

[19] Adomavicius G, Tuzhilin A (2011) Context-Aware Recommender Systems. In: Ricci F, Rokach L, Shapira B, Kantor PB. Recommender Systems Handbook, Springer. pp.217-253.

[20] Brey P (2005) Freedom and privacy in ambient intelligence. Ethics and Inf. Technol. 7 (3): 157-166.

[21] Andersen MR, Jensen T, Lisouski P, Mortensen AK, Hansen MK, Gregersen T, Ahrendt P (2012) Kinect Depth Sensor Evaluation for Computer Vision Applications. Tech. report ECE-TR-6, Department of Engineering, Aarhus University. pp.37.

[22] Ravnik R, Solina F (2013). Audience measurement of digital signage: Quantitative study in real-world environment using computer vision. Interacting with Computers. *In press.*

[23] Myers JL, Well AD (2003) Research Design and Statistical Analysis. Lawrence Erlbaum.