CONTEXTUAL INFORMATION ACCESS WITH AUGMENTED REALITY

A. Ajanki[1], *M. Billinghurst*[2], *T. Järvenpää*[4], *M. Kandemir*[1], *S. Kaski*[1], *M. Koskela*[1],
M. Kurimo[1], *J. Laaksonen*[1], *K. Puolamäki*[3], *T. Ruokolainen*[1], *T. Tossavainen*[3]

[1]Aalto University School of Science and Technology
Adaptive Informatics Research Centre
Helsinki, Finland

[2]Human Interface Technology Laboratory, NZ
University of Canterbury
Christchurch, New Zealand

[3]Aalto University School of Science and Technology
Department of Media Technology
Helsinki, Finland

[4]Nokia Research Center
Tampere, Finland

ABSTRACT

We have developed a prototype platform for contextual information access in mobile settings. Objects, people, and the environment are considered as contextual channels or cues to more information. The system infers, based on gaze, speech and other implicit feedback signals, which of the contextual cues are relevant, retrieves more information relevant to the cues, and presents the information with Augmented Reality (AR) techniques on a handheld or head-mounted display. The augmented information becomes potential contextual cues as well, and its relevance is assessed to provide more information. In essence, the platform turns the real world into an information browser which focuses proactively on the information inferred to be the most relevant for the user. We present the first pilot application, a Virtual Laboratory Guide, and its early evaluation results.

1. INTRODUCTION

In pervasive computing systems, there is often a need to provide users with a way to access and search through ubiquitous information associated with real world objects and locations. Technology such as Augmented Reality (AR) allows virtual information to be overlaid on the users' environment [1], and can be used as a way to view contextual information. However, there are interesting research questions that need to be addressed: how to know when to present information to the user, how to decide what to present given the plenitude of information, and what is the best way for users to interact with the information. As pointed out in [2], pervasive computing applications need to place few demands on the user's attention and be sensitive to context.

The authors appear in alphabetical order.

We are interested in the problem of how to efficiently retrieve and present contextual information in real-world environments where (i) it is hard to formulate explicit search queries and (ii) the temporal and spatial context provides potentially useful search cues. In other words, the user may not have an explicit query in mind or may not even be searching, and the information relevant to him or her is likely to be related to objects in the surrounding environment or other current context cues.

The scenario is that the user wears data glasses and sensors measuring his or her actions, or alternatively a handheld display with virtual see-through capability. The measured actions include gaze patterns and the visual focus of attention. Using the implicit measurements about the user's interactions with the environment, we can infer which of the potential search cues (objects, people) are relevant for the user at the current point of time, retrieve and place augmented retrieved information into the user's view. This new augmented information forms part of the user's visual context, and once the user's interaction with the new context is measured, more fine-grained inferences about relevance can be made, and the search refined (Figure 1).

To realize this scenario, several elements are needed. First, objects and people should be recognized as potential cues. People are detected using face recognition techniques. Objects are recognized from the fiducial markers attached to them. The relevance of these cues needs to be inferred from gaze patterns and other implicit feedback using machine learning methods. Second, context-sensitive information retrieval needs to operate proactively given the relevant cues. Finally, the retrieved information needs to be overlaid on the user's view with AR techniques and modern display devices.

In this paper, we present a hardware and software platform we have developed which meets these needs, and a

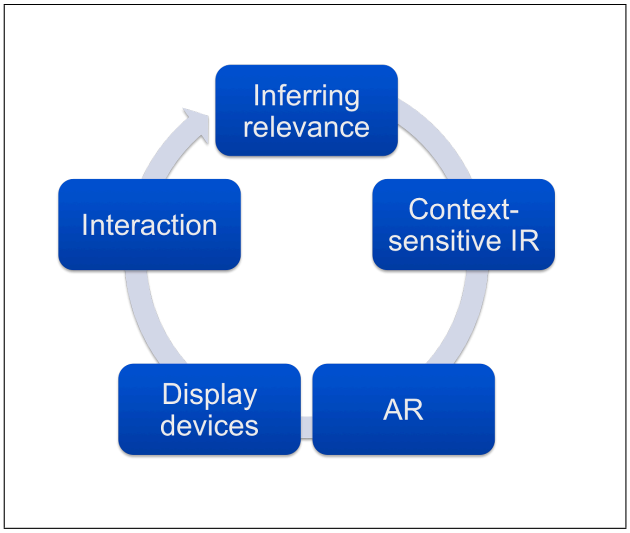

Fig. 1. Information retrieval (IR) is done in a loop where relevance of already augmented information (Augmented Reality, AR) and of the objects in the scene is inferred from user's observed interaction with them. Then more information is retrieved given any contextual cues and the inferred relevance, and new information is augmented.

demonstration prototype application created using the platform. This application is a *Virtual Laboratory Guide*, which will help a visitor to a university department by searching and presenting context-sensitive virtual information about the people, offices, and artifacts that appear as needed and disappear when not attended to.

In the remainder of this paper, we first review earlier related work, and describe the lessons learned which our research builds on. Then we discuss the hardware and software platform we developed, and finally present the Virtual Laboratory Guide application and a user evaluation of the technology.

2. RELATED WORK

In previous research, wearable and mobile AR systems have been used to display virtual contextual cues in the surrounding real environment. For example, the Touring Machine [3] added virtual tags to real university buildings showing which departments were in the buildings. A similar effect is created using the commercially available Layar[1] or Wikitude[2] applications for mobile phones, both of which provide virtual information tags on the real world. These interfaces highlight the need to filter information according to the user's interest, and present it in an uncluttered way so that it is easy to interact with. In most cases, mobile AR interfaces require explicit user input specifying the topics of interest to the user. In our research we want to develop a system that uses unobtrusive implicit input from the user to select and present relevant contextual information.

Contextual information retrieval takes into account the task the user is currently involved in, such as shopping or medical diagnosis, or the context expressed within a given domain, such as locations of the restaurants — see [4] for a recent overview. We use face and speech recognition and gaze location to measure context. Gaze and face recognition provide important implicit cues about where the user is looking and whom he is interested in. For example, [5] describes on a conceptual level how wearable computers could be used as an ideal platform for mobile augmented reality, and how they could enable many applications, including face recognition.

Studies of *eye movements* during natural behavior, such as driving a car or playing ball games, have shown that eye movements are highly task-dependent and that the gaze is mostly directed towards objects that are relevant for the task [6]. Gaze has been used as a measure of interest in virtual environments [7]. Eye tracking has been used as implicit feedback for inferring relevance in text [8, 9], and image [10, 11] retrieval applications on a conventional desktop computer. We have shown that there is information in the gaze signal useful for constructing implicit queries for textual search from reading patterns [12]. These results indicate that gaze direction is a useful information source for inferring the focus of attention, and that relevance information can be extracted even without any conscious effort from the user.

Gaze-controlled augmented reality user interfaces are an emerging research field. So far, the research has been concentrated on the problem of explicitly selecting objects with gaze [13]. The conceptual idea of using gaze to monitor the user's interest implicitly in a mobile AR system has been presented previously [13, 14], but until now a working system has not been demonstrated or evaluated.

Speech recognition in human-computer interfaces has been a subject of extensive study (for a review, see [15]). For example, the observed sound information has been augmented using a model of attention based on measuring the head posture [16], and [17] by measuring gaze on a computer display. However, as far as we are aware, the idea of combining gaze-based and speech-based implicit input about the interests and context in interaction with persons and objects in the real world is novel.

Julier et al. [18] have presented the idea of using real-world context as a search cue in information retrieval, and implemented a system which filters information based on physical location, for selecting what is displayed to the user in the AR interface. The main purpose of information filtering is to prioritize and reduce the amount of information presented in order to show only what is most relevant to the

[1] http://layar.com/ [2] http://www.wikitude.org/

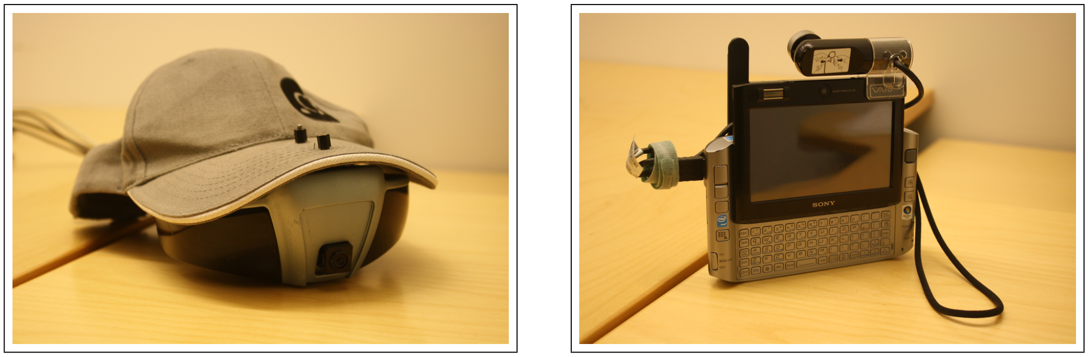

Fig. 2. The display devices. On the left, the head-mounted display (HMD), a wearable virtual see-through near-to-eye display with an integrated gaze tracker. On the right, an ultra-mobile PC (UMPC), a handheld computer with a virtual see-through display.

user. Already in the Touring Machine [3] more information and menu choices were shown for objects that had remained in the center of the user's view for a long enough time. This kind of contextual user feedback is, however, more explicit than implicit by nature. With our gaze tracking hardware we have been able to detect the implicit targets of the user's attention and to use that data in information filtering. As described in the previous section, there have been studies on using gaze as a form of relevance feedback, but to the best of our knowledge, the current work is the first one to use implicit gaze data for contextual information filtering in an AR setup and to evaluate its usefulness with a user study.

3. CONTEXTUAL INFORMATION ACCESS SYSTEM

We have implemented a pilot software system that can be used in on-line studies of contextual information access. It recognizes potential targets of interest, infers their relevance by observing user's behavior, retrieves related information, and augments it onto a display.

Our system can use two alternative output devices to display the retrieved information; (1) a head-mounted near-to-eye display (HMD) with an integrated gaze tracker and a camera (Figure 2 left), and (2) a handheld Sony Vaio ultra-mobile PC (UMPC) or a standard laptop with a camera (Figure 2 right). The near-to-eye display device is a research prototype provided by Nokia Research Center [19].

The system incorporates a *face recognition* system that detects and recognizes human faces. The detection is done by the Viola & Jones face detector [20] as implemented in OpenCV library. The detected faces are transmitted wirelessly to an image database server for recognition using the MPEG-7 Face Recognition descriptor [21] and a k-nn classifier. If a face is missed (*e.g.*, due to occlusion, changes in lighting, excessive rotation or camera movement), we initiate an optical flow tracking algorithm [22], which continues to track the approximate location of the face until either the face detector is again able to detect the face or the tracked keypoints become too dispersed and the tracking is lost.

The system can also detect two-dimensional *AR markers* which are used to indicate objects and indoor locations. We use the ALVAR augmented reality library[3], developed by the VTT Technical Research Centre of Finland, for detecting the fiducial markers and determining camera pose relative to them.

The system uses *speech recognition* to gather contextual cues from the user's speech. The speech transcript of a recognized utterance is used to determine the underlying topic of discussion. In the current pilot system, we have a fixed set of potential topics, and the decision between topics is made according to keywords detected from the transcripts. Later, it will be possible to generate multiple topics with corresponding keyword lists automatically using basic information retrieval techniques. We use an online large-vocabulary speech recognizer developed at Aalto University [23]. The system utilizes triphone Hidden Markov models as context-dependent and gender- and speaker-independent phoneme models. As a statistical language model, the system has a large 6-gram model of data-driven morpheme-like units trained on a 150 million words text corpus.

The system needs to decide which of the currently visible objects or people should receive augmented information at any given time. It does this by *inferring the relevance* from gaze (with the HMD) or pointing (with the UMPC) patterns. In this first pilot system, the relevance of an object is estimated by the proportion of the total time an object, or related augmented annotations, have been under visual attention or have been pointed towards within a time window.

We use *contextual information retrieval* to select the most useful information to display. The information database of the current pilot contains short textual snippets about people and objects. Each database entry is constrained to ap-

[3] virtual.vtt.fi/virtual/proj2/multimedia/alvar.html

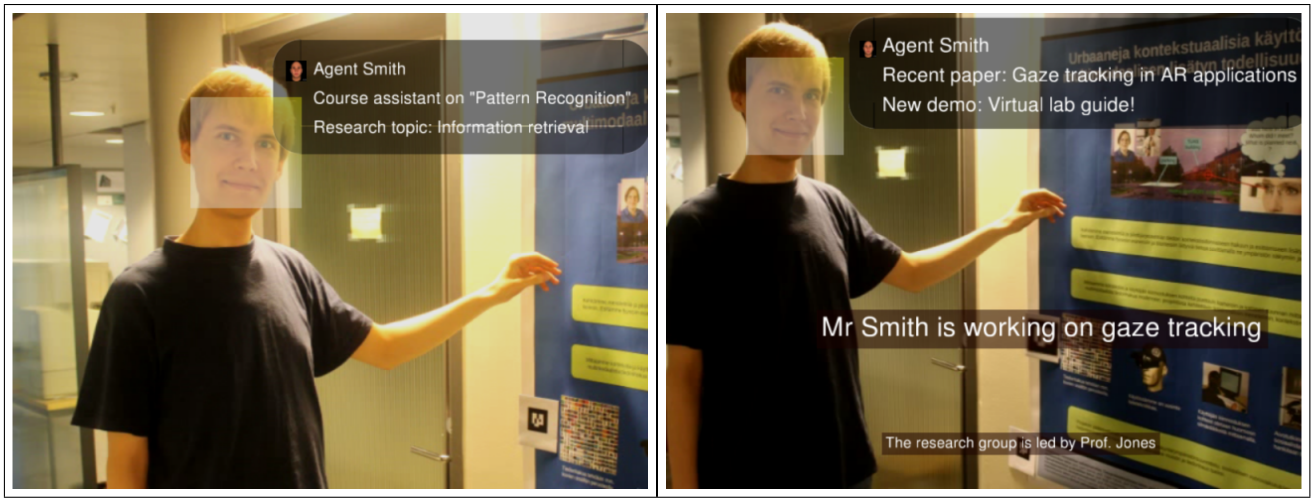

Fig. 3. Screenshots from the Virtual Laboratory Guide. Left: The guide shows augmented information about a recognized person. Right: The guide has recognized that the user is interested in research, and shows in the pop-ups mainly research-related information.

pear only in some contexts. The entry that best matches the current measured context is shown.

In our application the context, from which the query is inferred, is the set of real world objects and people around the user, and the currently visible augmented information. The role of the relevance inference is to indicate which of the many potential real or virtual objects should contribute to further information retrieval. The context is encoded as a feature vector that contains information about which objects or persons have been seen recently, and which topics the user is currently interested in. The potential topics are weighted based on the inferred the amount of attention different kinds of annotations have received previously, and the discussion topic recognized by the speech recognizer. Other context features, such as time and location, can be added in the future. For clarity, it is worthwhile to note that the topic of the context inferred from speech overrides the topic inferred from gaze, based on the prior belief that speech includes more accurate information than gaze.

The retrieved information is rendered on an AR overlay on top of video of the real world. The annotations are kept close to the objects or persons they are related to. The augmented content becomes part of the visible environment, and the user's interaction with the augmentations affects what kind of information is displayed in the future.

The objective of the application is to discover relevant contextual information and provide it to the user non-disruptively. This is accomplished by minimizing the amount of information shown, and by displaying only information that is assumed to be relevant for the user in the given context.

4. SYSTEM EVALUATION

As a pilot application for testing the framework we have implemented an AR guide for a visitor at a university department. The *Virtual Laboratory Guide* shows relevant information and helps the visitor to navigate the department. The guide is able to recognize people, offices and research posters, and complements them with information related to research or teaching.

Figure 3 shows sample screenshots from the pilot application; in the first screenshot two topics are displayed for the user to choose from, and in the second the user has been inferred to be more interested in research-related annotations.

A small-scale pilot study was conducted to provide an informal user evaluation of the Virtual Laboratory Guide application, and to test the usefulness of our AR platform. The main goal of the study was to collect feedback on how useful people felt the ability to access virtual information was and to find potential areas of further improvement in our system. We additionally compared the usability of alternative user interfaces: a handheld ultra-mobile PC (UMPC) and a head-mounted near-to-eye display (HMD). The head-mounted display is a very early prototype which will naturally affect the results. Figure 2 shows the display devices.

The 8 subjects (all male) were university students or researchers aged from 23 to 32 years old. None of them had prior experience with the Virtual Laboratory Guide application. Each subject used the display configurations (HMD or UMPC) in a counterbalanced in order to remove order effects. Before the experiments started each subject was trained on how to use the devices.

The subjects were asked to find answers to two questions about research or teaching. The answers were avail-

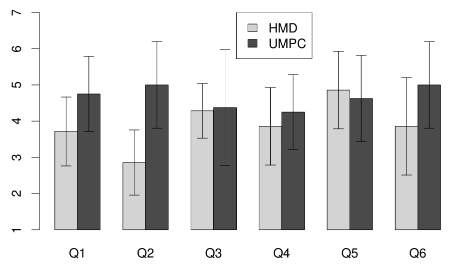

Fig. 4. Average scores with standard deviations for the usability questionary on 7-point Likert scale in ultra-mobile PC (UMPC) and head-mounted display (HMD) conditions.

able through the information augmented on the real objects (posters or course material) or on the people. When the subject had completed one task, the experiment supervisor changed the topic of the augmented information by speaking an utterance containing some keywords for the desired topic to the speech recognizer.

After each display condition the subjects were asked what they liked best and least about the conditions. We also collected answers to the following questions on a 7-point Likert scale: Q1: How easy was it to use the application? Q2: How easy was it to see the AR information? Q3 How useful was the application in helping you learn new information? Q4 How well do you think you performed in the task? Q5 How easy was it to remember the information presented? Q6 How much did you enjoy using the application?

4.1. Results

In general, users were able to complete the task with either the head-mounted display or the handheld display, and found the system a useful tool for presenting contextual information. Figure 4 shows a graphical depiction of the questionnaire results.

Although the test subjects generally liked the ability to see information about real-world objects in AR, there were some problems in both of the devices we tested that affected the ease of use (question 1). In further interview questions the users clarified that the most severe weakness in the head-mounted display was the quality of the image which made it difficult to read augmented texts. The subjects also felt the handheld display was too heavy and had too small a screen. Two persons said that they found the eye tracking and hands-free nature of the head-mounted display to be beneficial. There were no statistically significant differences between the two display conditions, except in the second question (how easy was it to see the AR information), where the subjects rated the UMPC as easier ($p = 0.016$, paired Wilcoxon signed rank test).

These preliminary results indicate that the test subjects were generally favorable towards the idea of contextual information access. However, both device interfaces we tested here still have weaknesses. A more formal user study will be completed in the future with a second-generation head-mounted display that is more comfortable and has better screen quality, and a lighter hand-held display.

5. DISCUSSION

We have proposed a novel AR application which infers the interests of the user based on his or her behavior, most notably gaze and speech. The application uses implicit feedback and contextual information to predict which pieces of information are relevant and when should they be shown over the video display of the real world. We have performed a usability study to validate our proposed approach. With the working prototype platform we are able to study related research questions more thoroughly in the future. Especially performance of the individual components, like face and speech recognition and the more open-ended IR implementation, should be studied.

Acknowledgements

Antti Ajanki, Melih Kandemir, Samuel Kaski and Kai Puolamäki belong to Helsinki Institute for Information Technology HIIT, and Kai Puolamäki to the Finnish Centre of Excellence in Algorithmic Data Analysis. This work has been funded by Aalto MIDE programme (project UI-ART) and in part by Finnish Funding Agency for Technology and Innovation (TEKES) under the project DIEM/MMR and by the PASCAL2 Network of Excellence, ICT 216886. We wish to thank H. Gamper for contributing to the manuscript.

6. REFERENCES

[1] R. Azuma, “A survey of augmented reality,” *Presence*, vol. 6, no. 4, pp. 355–385, 1997.

[2] Karen Hendricksen, Jadwiga Indulska, and Andry Rakotonirainy, “Modeling context information in pervasive computing systems,” in *Proceedings of the First International Conference on Pervasive Computing*, 2002, pp. 167–180.

[3] Steven Feiner, Blair MacIntyre, Tobias Höllerer, and Anthony Webster, “A touring machine: Prototyping 3D mobile augmented reality systems for exploring the urban environment,” *Personal and Ubiquitous Computing*, vol. 1, no. 4, pp. 208–217, 1997.

[4] Fabio Crestani and Ian Ruthven, “Introduction to special issue on contextual information retrieval systems,” *Information Retrieval*, vol. 10, no. 2, pp. 111–113, 2007.

[5] Thad Starner, Steve Mann, Bradley Rhodes, Jeffrey Levine, Jennifer Healey, Dana Kirsch, Rosalind W. Picard, and Alex Pentland, "Augmented reality through wearable computing," *Presence: Teleoperators and Virtual Environments*, vol. 6, no. 4, pp. 452–460, 1997.

[6] Michael F. Land, "Eye movements and the control of actions in everyday life," *Progress in Retinal and Eye Research*, vol. 25, no. 3, pp. 296–324, 2006.

[7] V. Tanriverdi and R.J.K. Jacob, "Interacting with eye movements in virtual environments," in *Proceedings of the SIGCHI conference on Human factors in computing systems*. ACM, 2000, p. 272.

[8] Kai Puolamäki, Jarkko Salojärvi, Eerika Savia, Jaana Simola, and Samuel Kaski, "Combining eye movements and collaborative filtering for proactive information retrieval," in *Proceedings of the 28th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, Salvador, Brazil, 2005, pp. 146–153, ACM.

[9] Thorsten Joachims, Laura Granka, Bing Pan, Helene Hembrooke, and Geri Gay, "Accurately interpreting clickthrough data as implicit feedback," in *Proceedings of the 28th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, Salvador, Brazil, 2005, pp. 154–161, ACM.

[10] László Kozma, Arto Klami, and Samuel Kaski, "GaZIR: Gaze-based zooming interface for image retrieval," in *Proceedings of 11th Conference on Multimodal Interfaces and The Sixth Workshop on Machine Learning for Multimodal Interaction (ICMI-MLMI)*, New York, NY, USA, 2009, pp. 305–312, ACM.

[11] Oyewole Oyekoya and Fred Stentiford, "Perceptual image retrieval using eye movements," in *International Workshop on Intelligent Computing in Pattern Analysis/Synthesis*, Xi'an, China, 2006, Advances in Machine Vision, Image Processing, and Pattern Analysis, pp. 281–289, Springer.

[12] Antti Ajanki, David R. Hardoon, Samuel Kaski, Kai Puolamäki, and John Shawe-Taylor, "Can eyes reveal interest? – Implicit queries from gaze patterns," *User Modeling and User-Adapted Interaction*, vol. 19, no. 4, pp. 307–339, 2009.

[13] Hyung Min Park, Seok Han Lee, and Jong Soo Choi, "Wearable augmented reality system using gaze interaction," in *IEEE International Symposium on Mixed and Augmented Reality*. 2008, pp. 175–176, IEEE.

[14] S. Nilsson, T. Gustafsson, and P. Carleberg, "Hands free interaction with virtual information in a real environment," in *Proceedings of COGAIN*, 2007, pp. 53–57.

[15] Carl M. Rebman, Jr., Milam W. Aiken, and Casey G. Cegielski, "Speech recognition in the human-computer interface," *Information & Management*, vol. 40, no. 6, pp. 509–519, 2003.

[16] D. Hahn, F. Beutler, and U.D. Hanebeck, "Visual scene augmentation for enhanced human perception," in *International Conference on Informatics in Control, Automation & Robotics (ICINCO 2005)*, Barcelona, Spain, September 2005, pp. 146–153, INSTICC Pres.

[17] Yong Sun, Helmut Prendinger, Yu Shi, Fang Chen, Vera Chung, and Mitsuru Ishizuka, "The hinge between input and output: Understanding the multimodal input fusion results in an agent-based multimodal presentation system," in *Conference on Human Factors in Computing Systems (CHI '08)*, Florence, Italy, April 2008, pp. 3483–3488.

[18] Simon Julier, Yohan Baillot, Dennis Brown, and Marco Lanzagorta, "Information filtering for mobile augmented reality," *IEEE Computer Graphics and Applications*, vol. 22, pp. 12–15, 2002.

[19] Toni Järvenpää and Viljakaisa Aaltonen, *Photonics in Multimedia II*, vol. 7001 of *Proceedings of SPIE*, chapter Compact near-to-eye display with integrated gaze tracker, pp. 700106–1–700106–8, SPIE, Bellingham, WA, 2008.

[20] Paul Viola and Michael Jones, "Rapid object detection using a boosted cascade of simple features," in *IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'01)*, Kauai, Hawaii, 2001, pp. 511–518.

[21] ISO/IEC, "Information technology - Multimedia content description interface - Part 3: Visual," 2002, 15938-3:2002(E).

[22] Carlo Tomasi and Takeo Kanade, "Detection and tracking of point features," Tech. Rep. CMU-CS-91-132, Carnegie Mellon University, April 1991.

[23] Teemu Hirsimäki, Janne Pylkkönen, and Mikko Kurimo, "Importance of high-order n-gram models in morph-based speech recognition," *IEEE Transactions on Audio, Speech and Language Processing*, vol. 17, no. 4, pp. 724–732, May 2009.