Mobile Phone Enabled Museum Guidance with Adaptive Classification

Erich Bruns, Benjamin Brombach and Oliver Bimber

***Abstract*—We present an adaptive museum guidance system called PhoneGuide. It uses camera-equipped mobile phones for on-device object recognition in ad-hoc sensor networks and provides location and object aware multimedia content to museum visitors.**

I. Introduction and Motivation

Although audio guides are widely established in many museums, they suffer from several drawbacks compared to state-of-the-art multimedia technologies: First, they provide only audible information to museum visitors, while other forms of media presentation, such as reading text or video could be beneficial for museum guidance tasks. Second, they are not very intuitive. Reference numbers have to be manually keyed in by the visitor before information about the exhibit is provided. These numbers are either displayed on visible tags that are located near the exhibited objects, or are printed in brochures that have to be carried. Third, offering mobile guidance equipment to visitors leads to acquisition and maintenance costs that have to be covered by the museum.

With our project *PhoneGuide* we aim at solving these problems by enabling the application of conventional camera-equipped mobile phones for museum guidance purposes. The advantages are obvious: First, todays off–the–shelf mobile phones offer a rich pallet of multimedia functionalities — ranging from audio (over speaker or head-set) and video (graphics, images, movies) to simple tactile feedback (vibration). Second, integrated cameras, improvements in processor performance and more memory space enable supporting advanced computer vision algorithms. Instead of keying in reference numbers, objects can be recognized automatically by taking non-persistent photographs of them. This is more intuitive and saves museum curators from distributing and maintaining a large number of physical (visible or invisible) tags. Together with a few sensor-equipped reference tags only, computer vision based object recognition allows for the classification of single objects; whereas overlapping signal ranges of object-distinct active tags (such as RFID) would prevent the identification of individuals that are grouped closely together. Third, since we assume that museum visitors will be able to use their own devices, the acquisition and maintenance cost for museum–owned devices decreases.

Yet, this approach holds several challenges. Museums are complex public environments that are —from a computer vision perspective— not very well controlled. Many hundreds, up to thousands of objects have to be classified from arbitrary perspectives, distances and under changing lighting conditions. In cooperation with local museums, we have tackled some of these problems over the past three years. Ideas, solutions and results are summarized in this article.

Fig. 1. Basic concept (a) and application (b,c) of the PhoneGuide system in a museum: Adaptive classification in dynamic large-scale museum environments supported by ad-hoc sensor networks and phone-to-phone communication.

II. Adaptive Classification

The major challenge of the PhoneGuide system is to locate and to recognize museum objects automatically in captured images. Although much research has been carried out in areas such as image retrieval and object recognition, it remains difficult to achieve high recognition rates in dynamic and uncontrolled large-scale environments such as museums. Often hundreds or even thousands of objects have to be reliably classified under varying lighting conditions and from arbitrary perspectives and distances. Small objects located in showcases, for instance, can not be photographed separately and have to be distinguished automatically from each other in a single image. The object recognition process becomes even more demanding if the computational possibilities of a mobile device are

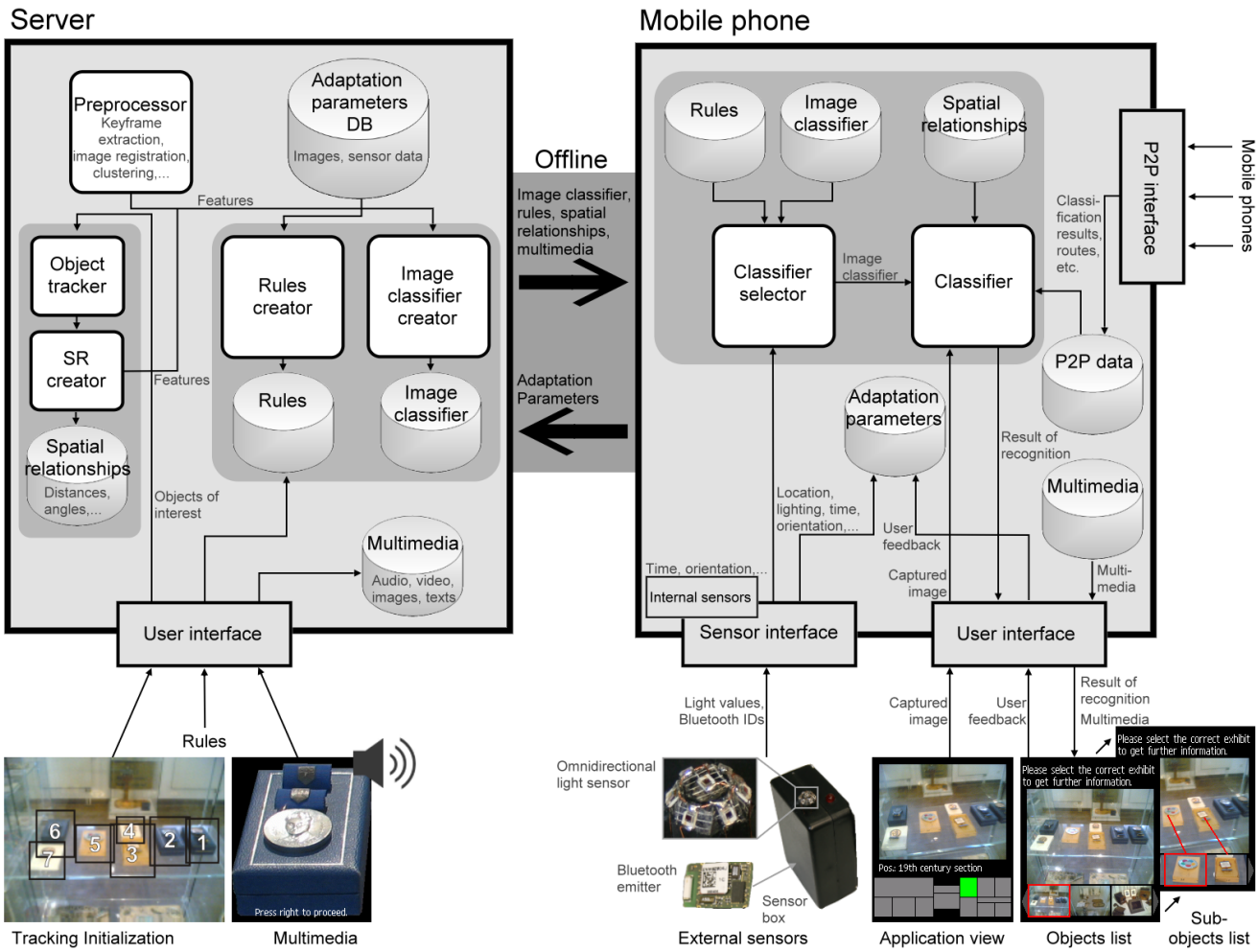

Fig. 2. Overview of adaptive classification infrastructure: During application, each phone collects environmental parameters and user feedback, sensor and phone-to-phone interfaces. When leaving the museum, these parameters are transmitted to the server, that stores and applies the gathered information to generate and improve the required classification elements (rules, image classifiers, spatial relationships). The adapted classification elements are transmitted to the mobile phones of new visitors upon entering the museum.

limited.

To overcome these limitations, we have developed an adaptive classification infrastructure (cf. figure 2). It continuously collects sensor data and user feedback to adapt and improve the local classification process over time. This is supported by utilizing a coarse sensor network that provides local information (such as rough position and environmental illumination data) to the mobile devices. Together with the user feedback gathered during the application of the system, these parameters are applied to adjust and optimize classifiers, and they can be shared with other users through ad-hoc phone-to-phone networks.

Many related approaches exist that adapt classifiers to specific technical circumstances, user behavior or environmental conditions. Most of them, however, perform an adaptation without future influence. Thus, they do not advance over time. Relevance feedback methods, for instance, are a common technique for information retrieval systems that evaluate the user's feedback on query results. MacArthur et al. [1] apply a decision tree for an image retrieval application that adapts to the subjective relevance (indicated by the user) for each query result. Based on this, a new query is performed. In contrast to our approach, this method does not apply the information gathered from one user to improve queries of other users. Draper et al. [2] introduced ADORE, an adaptive object recognition system that selects the optimal classification technique for an arbitrary recognition task automatically. Yet, ADORE does not adapt or improve over time since no feedback or other data is collected. In contrast our system applies only one static set of features.

Machine-learning techniques with adaptive learning behavior can be found in robotics. Hagras et al. [3], for example, present an autonomous mobile robot that continuously adapts to a changing environment and utilizes a continuous learning technique in order to accomplish tasks in agricultural domains. Our system collects and applies visitor feedback in addition to environmental parameters to adapt primarily to the users' application behavior as well as to environmental conditions.

The adaptive classification infrastructure of PhoneGuide consists of a stationary server and an arbitrary number of mobile phones and sensor boxes. Their general functionalities are briefly summarized below. More details are provided in the

following sections.

The server continuously carries out two main tasks: First, it constantly collects and stores adaptation parameters, such as environmental information and user feedback, that have been gathered by each individual mobile phone during run-time (i.e., during the museum visit). These parameters are transmitted from the phones to the server when leaving the museum. Note that there is no on–line connection between the mobile devices and the server during run–time. Second, the server applies the adaptation parameters for creating and improving the required classification elements, such as rules, image classifiers, and spatial relationships, off–line. These improved elements are transmitted to the mobile phones of new visitors when entering the museum. Note that our off-line adaptation allows for the application of computationally expensive training processes that would overload current mobile devices. This decentralized attempt makes the system highly scalable to an arbitrarily large number of users since the heavy-weight training process is carried out off–line on the server while the lower–weight classification task is performed individually and in parallel by each mobile phone. This distinguishes our approach from all centralized mobile classification systems that perform multiple recognition requests sequentially on a remote server [4], [5], or utilize high–performance (Tablet) PCs as mobile devices [6], [7], [8].

The server consists of three major components: One module is responsible for the preprocessing steps (object tracker and spatial relationship (SR) creator) that are required for classifying multiple objects in a single image (explained in section *recognition of sub–objects*). Furthermore, the server contains a preprocessor that prepares image data for training existing image classifiers. Another module dynamically creates rules and image classifiers based on adaptation parameters. The rules (e.g., defined by a naive bayes classifier) determine which classifier has to be selected for a specific environmental state. The states are defined by temporally collected environmental parameters, such as local position data and illumination information. Thus, for each state, the optimal classifier can be selected and trained.

The front–end application on the mobile phone provides a user interface and tracks (unnoticed by the user) actual recognition results as well as provided user feedback: As an outcome of the object identification, a probability-sorted objects list is displayed after taking a photograph. The user selects the correct object from this list with a minimum number of clicks (only one click if the object has been classified correctly, two clicks if the correct object has second highest probability, etc.). This provides essential feedback that is used later for adaptation on the server. Sensor boxes that are located in the proximity provide the necessary information to determine the users' rough locations through a simple pervasive tracking method [9], as well as the approximated local illumination state of the environment.

Before the classification is carried out, however, the correct image classifier (pre-trained by and transmitted from the server earlier) has to be selected based on the incoming environmental data of the nearby sensor boxes. This data serves as input for those rules which the classifier selector applies for determining an (for the given conditions) optimal classifier. Optionally, the selected classifier can utilize spatial relationships to identify multiple objects within a single image. The phone-to-phone interface can be applied for exchanging adaptation parameters dynamically during run-time (i.e., without a check-in/check-out at the server when entering or leaving the museum). These parameters will not be used for retraining the classifiers on the phone (as they would be used for on the server), but for adapting pre-trained classifiers to momentary situations in the museum. Note that this component is still under development and has not been formally evaluated. Consequently, it will be discussed in the outlook section.

The remainder of this article will discuss these components in more detail.

III. Perspective invariance

In practice, our system has to be flexible enough to compensate for individual user behavior. The ways in which visitors approach and observe an object can vary to a great degree. This leads to significant perspective differences in photographs that are taken for classification. For ensuring an acceptable recognition rate, the classification process must be scale and perspective invariant.

To solve this problem, we apply the video capturing functionality of mobile phones to record videos containing multiple perspectives and distances of each object in the museum. These videos are preprocessed by the server (as indicated in the previous section): keyframes are extracted and clustered. The aim is to eliminate redundant frames and select frames that contain descriptive perspective and scale information. The remaining frames are forwarded to the image classifier generator that —based on these frames— configures and trains an optimized classifier. Consequently, these videos are applied for an efficient initial training of the system. They are recorded for all objects only once by the museum operator when installing the system. Each subsequent modification of the exhibit requires the recording of a video for the changes only (e.g., one video of a new object or an existing object at a new location). However, the classifiers will be continuously updated and temporally improved over time. All images that are captured by the visitors, in combination with their individual recognition results and sensor values, are adaptation parameters and are consequently part of the adaptation process. Thus, step-by-step, the system will adapt to the most common photographed perspectives and distances that were chosen by the museum visitors —and consequently to the visitors' behavior and to periodic environmental (lighting) conditions. To prevent misapplications from adaptation drifts due to incorrect user feedback, the server eliminates outliers through clustering automatically. Details on the adaptive training method and results from a user study that validates a common visitor behavior (and consequently justifies this approach) can be found in [10]. Figures 3a and 3b illustrate two examples from a user study which show that visitors follow a similar behavior pattern and approach the same objects in very similar ways. They take photographs within small distinct areas, rather than from all possible perspectives and distances. Our system will

be able to adapt exactly towards these perspectives and scales after a period.

IV. Classification in large–scale and dynamic environments

Perspective and scale invariance are requirements that have to be met for all objects individually. In public environments, such as museums, two –more global– challenges have to be addressed: scale and dynamics. The large number of objects to be recognized under varying lighting conditions (mainly due to the changing daylight) represent other major barriers for achieving high recognition rates. To overcome these problems, we apply a coarse network of custom-built sensor boxes (cf. figure 2) that provides the additional information to phones located in their proximity.

A. Large-scale classification

In general, it holds that the more objects a classifier has to separate, the lower its recognition rate will be. Each sensor box is equipped with a Bluetooth chip for communicating with the mobile phones that are located in its signal range. Other than sensor data, the chip also transmits its unique ID. The IDs of all sensor boxes in the network together with their known positions and signal ranges span a coarse grid of (possibly overlapping) signal cells. Estimating the cell in which a phone is currently located by analyzing all detectable sensor boxes (and possibly by evaluating the strength of each signal — which is currently not supported due to an implementation in J2ME MIDP 2.0, CLDC 1.1) indicates to each device its own rough position within the museum. Based on this position data, the classifier selector chooses a classifier that is optimized for recognizing only the objects which are located in the proximity of the user (i.e., objects that are located within the same signal cell as the user). In practice, only a small number of objects need to be distinguished from each other while an arbitrary number of objects can be recognized with a suitable number of signal cells. More information about how our classification approach is guided by pervasive tracking techniques can be found in [9]. Figure 3 illustrates a floor plan of the City Museum of Weimar that was equipped with eight sensor boxes for experiments. The spanned grid of signal cells are color coded. The number of objects located in each signal cell ranged between 2 and 28.

B. Illumination invariance

Most image-based classification techniques become unstable with significant changes in illumination. In museums, for example, the lighting changes frequently next to windows or due to the fact that the room lights are sometimes turned off and sometimes turned on. Each of our sensor boxes uses seven hemispherically aligned photo diodes (cf. figure 3) that measure the incoming radiance at a solid angle of 180 degrees for the position where they have been placed. This is shown in figures 3c and 3d for the eight sensor boxes. The gray scale intensity values can be compared with omnidirectional photographs that have been taken from roughly the same positions. These seven values are broadcasted (together with the sensor ID) to each phone located in the proximity of the sensor box. Consequently, the local illumination information is available and is stored together with a time-stamp on the phone. As mentioned earlier, they are part of the adaptation parameters that can be used on the server for adapting classifiers to different local illumination conditions. If, for instance, a recognition fails, the captured image together with the corresponding illumination data will be transmitted to the server as part of the adaptation process. On the server, it has then to be decided whether the misclassification was due to varying lighting conditions or due to an invalid perspective or scaling. This can be done by geometrically registering the failed image to all existing images of the same object that are already stored on the server. If an appropriate match is found, it is selected and the brightnesses of the common image areas are compared. If, however, the registration process failed either due to changes in illumination or in perspective/scale a new classifier is created to cope with the new lighting conditions. Furthermore, the rules are updated. The illumination data is also used locally on the phone to select the correct pre–trained classifier.

V. Classification of sub-objects

Many exhibits in museums are protected against environmental influences or human curiosity by placing them into showcases or behind other barriers. In these cases, visitors can not take photographs of individual objects without capturing others simultaneously. This section explains how our adaptive classification framework is extended to support the recognition of multiple objects in a single photograph. Practically, the classification of sub-objects happens in two steps: After taking a photograph, a regular image classification is carried out first as described above. In this step, we do not differentiate between photographs that contain single or multiple objects. Since image classification techniques rather than object recognition methods are applied, a scene with multiple objects can be identified just like scenes with individual objects. As mentioned above, the classification result is presented to the users as a probability-sorted objects list. Just like for individual objects, the user selects —with a minimal number of clicks— the correct scene (if the scene is recognized, it is displayed as first entry in the list and one click is sufficient). After defining the correct scene context, the individual sub-objects in the photograph are classified automatically. The results are labeled and the sub-objects are linked with a sub-object list as shown in figure 4f. From this list, the user can finally select the object of interest and multimedia content is presented.

Our classification technique for sub–objects is based on spatial relationships [11]. For scenes that contain inseparable sub–objects, the video material that is applied for initial off-line training on the server (see section *perspective invariance*) is treated in a special way: In the first frame of these training videos, the operator manually identifies all sub-objects (cf. figure 4a). These objects are tracked by the object tracker (cf. figure 2) via SIFT throughout all subsequent frames (cf. 4b). If new objects appear, they have to be manually identified to

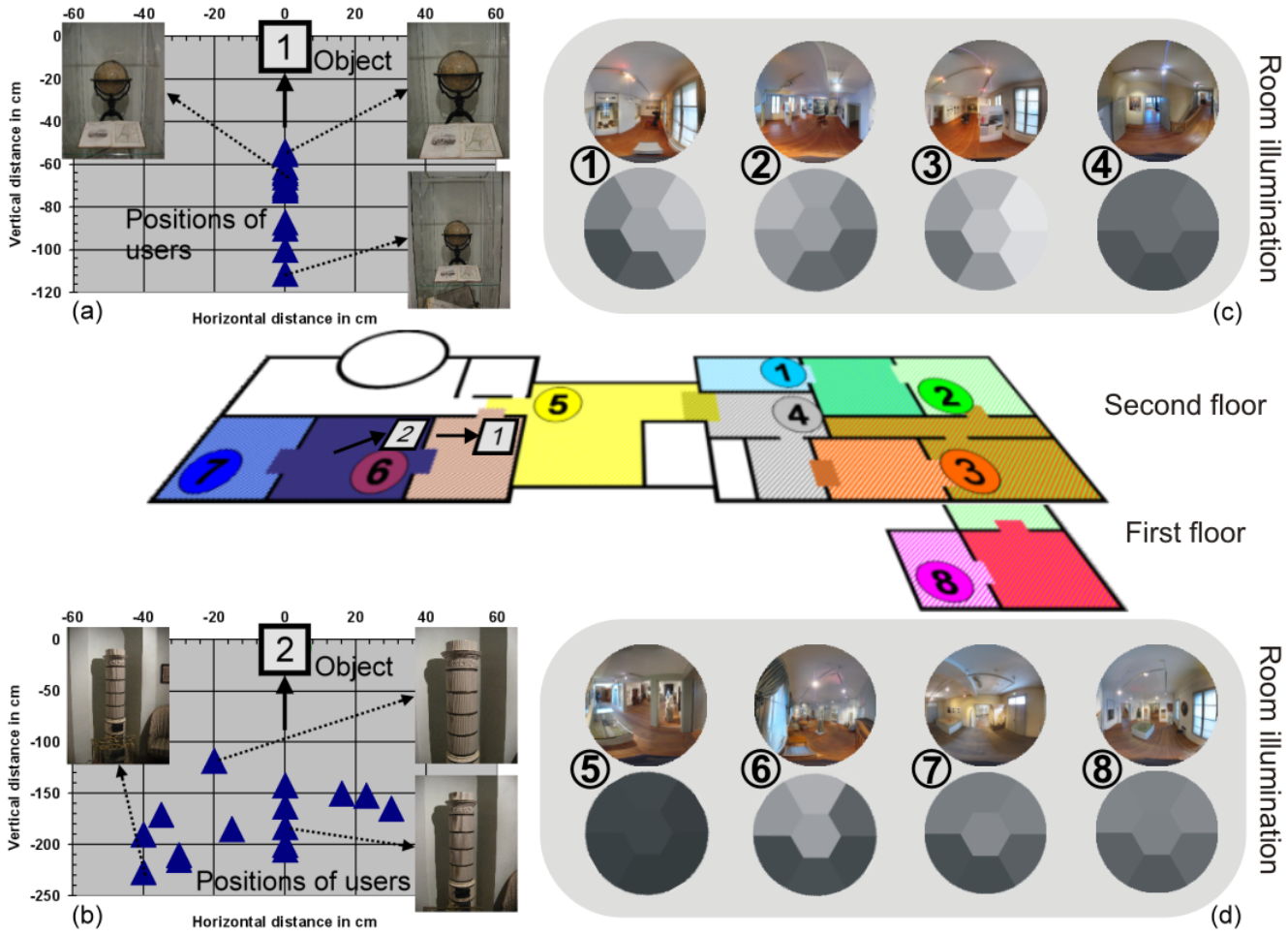

Fig. 3. Eight sensor boxes were distributed in the City Museum of Weimar during a formal user study. The floorplan in the center illustrates their location and the spanned signal cells (color coded). The measured incoming radiances of each sensor can be compared with omnidirectional photographs taken from the same positions (c, d). The location of 15 subjects when taking photographs of two different sample objects are outlined in (a, b). They indicate that visitors independently approach the same objects within the same small region —even though much more space was available (the areas shown in (a, b) are only fractions of the actual rooms).

be tracked. By doing so, the spatial relationship tracker (cf. figure 2) continuously computes spatial relationships (such as maximal search angles and distances) between all sub-objects. Additionally, the size of each sub-object's bounding box and it's individual classification features are computed and stored. The features are used to train individual classifiers on the server that are specialized to detect the sub-objects on the phone.

The trained classifiers and computed spatial relationships are transferred to the phones along with the additional data when entering the museum. After taking a photograph on the phone, the scene context has to be classified first, as explained above (cf. figure 4c). After this step, the corresponding set of classifiers and spatial relationships are selected automatically. An anchor object is classified that is assumed to be located in the center of the photograph. We apply multi-resolution classification to cope with different scales. Perspective invariance is hereby ensured, as explained above.

The scale and position of the anchor object's bounding box are then used for selecting the correct spatial relationships (cf. figure 4d). The maximal search angles and distances to neighboring sub-objects guide the following search process: The closest neighbor is searched at the mean distance and angle (green dot in figure 4e) which is defined by the spatial relationships. If a sub-object could not be classified at this position, a search mask is spirally shifted around the initial position (yellow dots in figure 4e) until a sub-object is classified: If the excitation of the classifier is above a predefined threshold (blue dot in figure 4e), a sub-object is found only if, in addition, the excitations for neighboring search points are lower (orange dots in figure 4e). Otherwise, the final position of the subobject is gradually moved further until the highest excitation is discovered (red dot in figure 4e). To ensure a rapid classification directly on the phone, integral images are initially computed that can be used for a fast feature computation within the search masks.

Additional sub–objects are found by repeating this process from already detected sub–objects. Note that only undetected sub–objects have to be traced. Since the spatial relationships can be optimized continuously the more sub-objects have been

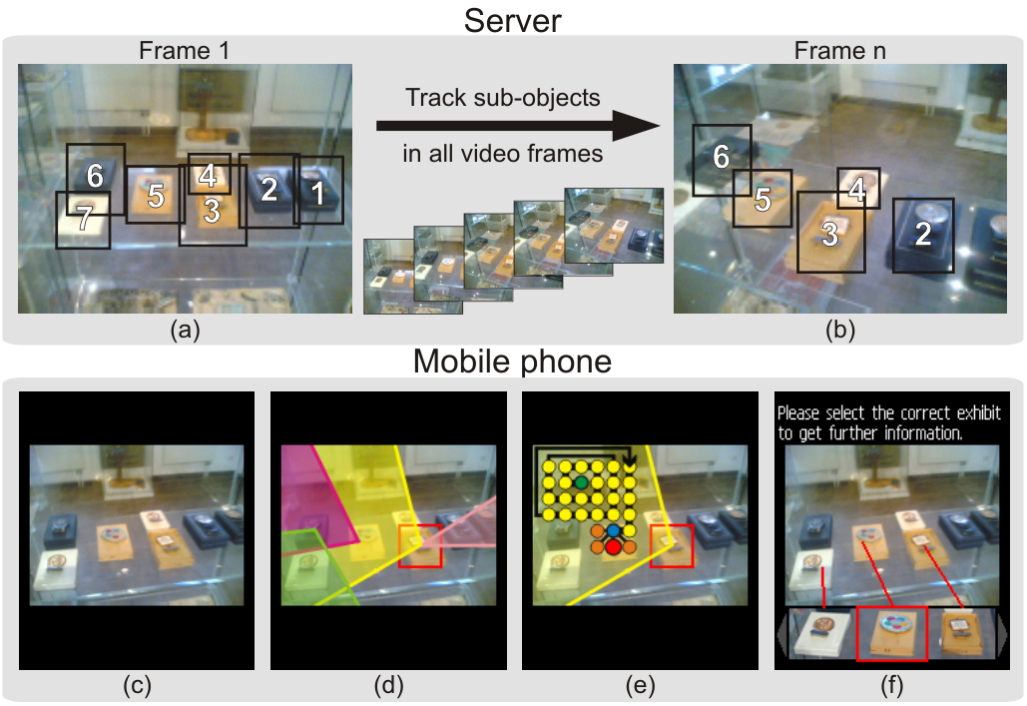

Fig. 4. Classification of sub-objects: On the server, individual sub-objects have to be manually identified in the first frame of each video (a). They are then automatically tracked throughout all frames (b) to compute spatial relationships. On the phone, the scene context has to be detected first (c). The corresponding spatial relationships are then used (d) to search for all sub-objects (e). The identified sub-objects are labeled, and the one of interest can be selected from a sub-objects list (f).

detected, the classification process will speed up with each detected sub-object. For instance, if the distance between two sub-objects has been determined for one particular perspective and scale, it can be used as a scaling factor to adjust the spatial relationships of the remaining sub–objects.

The adjusted spatial relationships can be stored on the phone and transmitted to the server as part of the adaptation process. This way, the approximation of spatial relationships for individual perspectives and scales will also be continuously optimized over time. Note that the adaptation of our sub-object classification is currently being implemented.

After all sub-objects have been found, they are labeled and linked to close–up pictures in the sub-object list (cf. figure 4f). The visitor can browse through this list to select the object of interest. Multimedia content of the selected object is then presented.

VI. Results and Outlook

In cooperation with the City Museum of Weimar, we were able to test and to evaluate our system during regular opening hours. In our current implementation, we applied a well selected set of global image features [12], [10] and three-layer neural networks for image classification [10]. Since the classification is widely independent from the adaptation framework, it can easily be replaced by enhanced algorithms, such as SIFT, as soon as their computation time on the mobile devices becomes acceptable. On Nokia 6630 mobile phones, our local object recognition algorithm implemented in J2ME requires on average 3.8 seconds. For 139 objects,

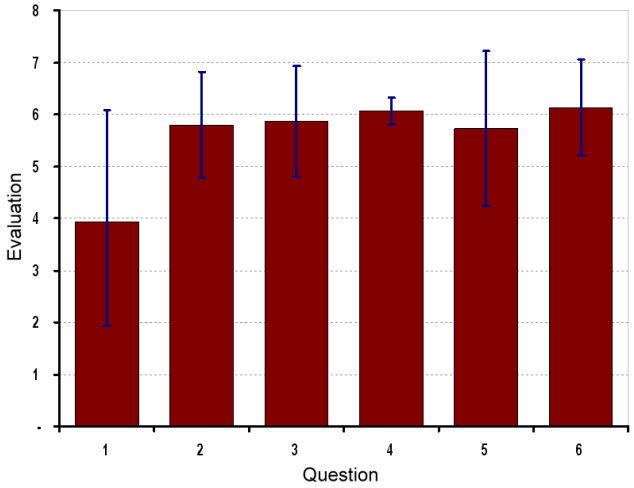

Fig. 5. Result of a questionaire gathered in the course of a user study from 15 museum visitors. Each question was answered through a ranking between 1 (worst) and 7 (best).

we achieve a recognition rate of 92.6% for experienced users and 82% for totally inexperienced museum visitors. In the context of a user study [10], we achieved these results under realistic conditions (i.e., arbitrary perspectives and scales, evaluated over a duration of 4 business days at different day times and illumination situations). We could also show that a temporal adaptation does lead to a continuous improvement

System	Location information	Recognition on server	Temporal adaptation	Recognition rates
Hare et al. [4]	No	Yes	No	850 images of 850 2D paintings: **80.0%** for 200 trials
Bay et al. [7]	No	No	No	205 images (splitted into two sets) of 20 objects: **91.5%** for 116 trials
Fasel et al. [8]	No	No	No	207 images (splitted into two sets) of 20 objects: **94.6%**[a] for 119 trials
Bay et al. [6]	Yes	No	No	130 images of 22 objects: **80.0%** for 200 trials rec. rate $\sim$ independent of #objects
Fritz et al. [5]	Yes	Yes	No	40 images of 20 objects: **97.5%** for 40 trials 1005 images of 201 objects: **91%** for 115 trials rec. rate $\sim$ independent of #objects
PhoneGuide	Yes	No	Yes	7200 images of 139 objects: **92.6%** for (6 perspectives x 139 objects =) 834 trials (expert user); **82%** for 139 trials (museum visitors) rec. rate $\sim$ independent of #objects

[a]97.5% for an unsplitted image set. Matching time is not applicable for mobile devices.

TABLE I
COMPARISON OF PHONEGUIDE WITH RELATED APPROACHES.

of the recognition rate over time [10].

We compare our system with the most related approaches in table I: Fritz et al. [5] introduced a city guide for mobile phones that identifies buildings or monuments. Photographs that are coupled with GPS location data are transferred to a remote server via UMTS or GPRS for classification using a variation of SIFT called i-SIFT. Hare et al. [4] developed a museum guide for Pocket PCs that recognized paintings. Images are sent to a server that computes SIFT features but applies image retrieval techniques for classification. Bay et al. [7] also demonstrated a museum guide based on a Tablet PC. However, in contrast to [5], [4] the classification is carried out locally using SURF. In their previous work [6], they also apply Bluetooth emitters for pervasive tracking. In their latest approach [8], they suppress multiple feature–point matches between test and model images by removing all matches above a minimal distance. None of these approaches is either adaptive or able to improve over time while being used. Consequently, we believe that realistic recognition rates and invariance against perspective, scaling, and illumination for large-scale and uncontrolled situations (hundreds or thousands of objects, and changing lighting) are difficult to achieve with such techniques.

Besides estimating the quantitative benchmark data, we were interested in the subjective impression of museum visitors after using our system. Therefore, we asked fifteen subjects to fill out a questionnaire and answer (inter alia) the following questions (1=worst, 7=best)[10]:

1) How convenient was the duration of waiting for the location estimation (pervasive tracking)?
2) How do you judge the recognition performance of PhoneGuide?
3) How simple was the handling of the application?
4) How satisfied were you with the integrated concept of PhoneGuide?
5) Can you imagine that PhoneGuide would become an adequate alternative to today's museum guidance systems (e.g., audio guides)?
6) Do you believe that PhoneGuide can be applied in different contexts (e.g., for city guidance)?

The results are presented in figure 5. The relatively long waiting time required for the device localization was the most criticized aspect of our approach. In our implementation, it takes approximately 13 seconds for the scanning of nearby Bluetooth emitters. This waiting time occurs only during transitions between signal cells. The waiting time for the recognition process remains constant. However, it is clear that such high waiting times can easily occur for individual recognition tasks if a centralized classification framework would be used (such as in [4] and [5]). They will not scale well with an increasing number of users and simultaneous classification requests. Our decentralized classification architecture addresses all recognition requests in parallel and directly on the local devices. Thus, no additional waiting time that is due to network communication, and sequential request–handling on the server is introduced. Newer and faster phones will even decrease the waiting time for individual classifications.

Note that the results described above do not yet consider the illumination sensor data (see section *large-scale classification*). A formal long-term evaluation of our system that incorporates this information is one of our future tasks. They also do not include the recognition rate and performance for sub-object classification (see section *recognition of sub-objects*). Although this must also be evaluated formally and under realistic conditions, we carried out an initial benchmark test: Our current implementation requires 2 to 3.5 seconds for classifying 6-8 sub-objects with a recognition rate of 93% (6% of all sub-objects were not found and 1% of all sub-objects were found at wrong places, 90 trials have been carried out with three different object sets). In some cases, reflections or shadows of visitors on (sub–)objects might lead to miss–recognitions and can prevent sub–objects from being located correctly. For instance, image portions of small sub–objects can be occluded by lens flare effects that modify their true appearance completely.

Currently, adaptations to user behavior and environmental changes do not become immediately effective. Adaptation parameters have to be uploaded to the server first to lead to improvements later. For enabling quicker adaptations to

changes that take place during the actual museum visit, we are investigating and implementing additional ad-hoc network techniques —as they might be used in the future for car-to-car communication and other areas. As illustrated in figure 1a, a short direct point-to-point connection is established between all phones that are within signal range (for compatibility reasons we use Bluetooth at the moment, but WiFi is also imaginable). Since each phone stores information about each individual succeeded or failed classification trial together with a time stamp, this data can be provided constantly to other phones. By doing this, we ensure that each phone stores statistical information about current object–individual classification rates as well as about confusions with other objects. Since this is continuously being repeated while the visitors move through the museum, the data that is stored on each phone is always as up-to-date as possible (likely to be different for each visitor — depending on the movements and actions of all visitors). This data allows for influencing the classification process directly without re-training. As for transmitting information from the sensor boxes, this process is carried out in the background and remains unnoticed by the user.

We believe that mobile-phone enabled guidance systems have a substantial potential in future —for indoor (such as for museum guidance) as well as for outdoor (such as for city guidance) applications, and that computer vision support is complementary to other sensory information (such as provided by GPS, RFID, etc.) and manual user input. For achieving realistic classification rates in dynamic and complex public environments, however, we see an intelligent system adaptation as an essential component. For large-scale and dynamic outdoor environments, established web services such as google earth indexing or Flickr can be applied to organize (geographically and temporally) the data accumulated by adaptive classification systems.

VII. Acknowledgments

We thank the Senckenberg Museum of Natural History Frankfurt, the Museum for Pre- and Early History Weimar, the Museum of the City of Weimar, and CellIQ for their support. The PhoneGuide project is supported by the Stiftung für Technologie, Innovation und Forschung Thüringen (STIFT). Further information is available at http://www.uni-weimar.de/medien/AR.

References

[1] S. MacArthur, C. Brodley, and C. Shyu, "Relevance feedback decision trees in content-based image retrieval," *IEEE Workshop on Content-based Access of lmage and Video Libraries*, pp. 68–72, 2000. [Online]. Available: citeseer.ist.psu.edu/macarthur00relevance.html

[2] B. A. Draper, J. Bins, and K. Baek, "ADORE: Adaptive object recognition," *International Conference on Vision Systems*, pp. 522–537, 1999. [Online]. Available: citeseer.ist.psu.edu/draper99adore.html

[3] H. Hagras, M. Colley, V. Callaghan, and M. Carr-West, "Online learning and adaptation of autonomous mobile robots for sustainable agriculture," *Autonomous Robots*, vol. 13, no. 1, pp. 37–52, 2002.

[4] J. S. Hare and P. H. Lewis, "Content-based image retrieval using a mobile device as a novel interface," *Proceedings of the first international workshop on mobile vision*, pp. 64–75, Dec. 2004.

[5] G. Fritz, C. Seifert, and L. Paletta, "A mobile vision system for urban detection with informative local descriptors," *ICVS '06: Proceedings of the Fourth IEEE International Conference on Computer Vision Systems*, p. 30, 2006.

[6] H. Bay, B. Fasel, and L. V. Gool, "Interactive museum guide," *The Seventh International Conference on Ubiquitous Computing UBICOMP, Workshop on Smart Environments and Their Applications to Cultural Heritage*, September 2005.

[7] ——, "Interactive museum guide: Fast and robust recognition of museum objects," *Proceedings of the first international workshop on mobile vision*, May 2006.

[8] B. Fasel and L. V. Gool, "Interactive museum guide: Accurate retrieval of object descriptions," *Adaptive Multimedia Retrieval*, pp. 179–191, 2006.

[9] E. Bruns, B. Brombach, T. Zeidler, and O. Bimber, "Enabling mobile phones to support large-scale museum guidance," *IEEE Multimedia*, vol. 14, no. 2, pp. 16–25, 2007.

[10] E. Bruns and O. Bimber, "Adaptive training of video sets for image recognition on mobile phones," *submitted to: Springer Personal and Ubiquitous Computing*, 2007.

[11] D. Papadias and Y. Theodoridis, "Spatial relations, minimum bounding rectangles, and spatial data structures," *International Journal of Geographical Information Science*, vol. 11, no. 2, pp. 111–138, 1997. [Online]. Available: citeseer.ist.psu.edu/article/papadias97spatial.html

[12] P. Föckler, T. Zeidler, B. Brombach, E. Bruns, and O. Bimber, "Phoneguide: Museum guidance supported by on-device object recognition on mobile phones," *MUM '05: Proceedings of the 4th international conference on Mobile and ubiquitous multimedia*, pp. 3–10, 2005.

I. Related work (optional sidebar)

For designing digital mobile guidance systems, a multitude of different technologies exist. They can be categorized in three major groups (cf. figure 1): User feedback approaches allow visitors to identify objects manually in order to retrieve further context information. They comprise traditional museum guidance techniques such as audio guides (object IDs are provided on human-readable lables or in handouts that have to be keyed in by users), as well as electronic lists (of images or text) provided on the mobile device that visitors browse through in terms of making a selection.

Sensor systems apply small sensor devices for identification. They are either attached to exhibits for identifying them directly, or they are used to determine the location and/or orientation of the mobile device within the environment (and consequently determine objects in its proximity). Wireless connection technologies such as RFID[1], [2], Bluetooth [1] or Infrared [3], [2], [4] are usually utilized for this. Precise object identification with sensors only could be accomplished through narrow signal ranges of the emitters to ensure unambiguousness. Välkkynen et al. [2], for instance, use sensor devices called SoapBoxes [5], which host several integrated sensors. Besides scanning for RFID chips in the proximity (which are are embedded in each SoapBox), a SoapBox can be triggered by a mobile device through infrared or laser light (via an integrated light sensor). A wireless connection between the corresponding SoapBox and the mobile device is established and context data, such as URLs are transmitted.

Several similar approaches exist that evaluate the user's location for providing context information [3], [6], [7]. Wide-range emitters (Bluetooth, WiFi, or GPS) are used frequently to determine the approximated location of users. With this information an assortment of close objects can be presented in a selection list. One of the first location–based mobile guidance systems was called Cyberguide [3]. Simple maps with outlines of buildings and context information are displayed on a hand-held device equipped with a GPS receiver. For indoor applications, infrared beacons are evaluated to estimate the device's position. Cheverst et al. [7] have introduced a city guidance system on a Tablet PC that combines user feedback with location information. For recognizing objects, users have to manually provide a rough indication of how far they are away. The location information is then estimated by detecting nearby WiFi hotspots. An overview of different location–based guidance systems can be found in [8].

In [9] a technique was presented that captures each room of a museum with a fisheye–camera. The resulting panorama images are presented on a PDA —registered to the real world via a digital compass attached to the device. Depending on the orientation, objects can be selected by simply clicking on the corresponding region of the panorama image. A remote server then delivers the appropriate multimedia content.

Computer vision approaches utilize integrated or attached cameras for classification —either by recognizing the object directly or by identifying machine–readable barcodes (e.g., QR-Codes, Datamatrix) located next to them (e.g., [10]). Especially direct recognition techniques seem to be very promising for object identification. This is also confirmed by user experiments, such as the ones performed by Davies et al. [11]. They investigated the acceptance of pure location–based versus computer vision based (classification was only simulated in their experiments for controlling the recognition rates) techniques for guidance systems. They found that both approaches are equally preferred —despite the imperfect classifications. The various direct object classification techniques that perform a recognition directly on the local device [12], [13], [14] or on a remote server [15], [16] have been summarized and compared with PhoneGuide in the main text.

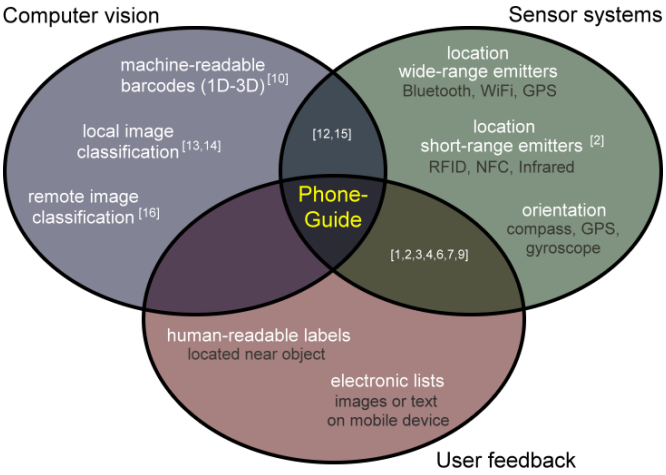

Fig. 1. Overview of different techniques that are applied for object identification in digital mobile guidance systems.

References

[1] F. Siegemund and C. Flörkemeier, "Interaction in pervasive computing settings using bluetooth-enabled active tags and passive rfid technology together with mobile phones," *PERCOM '03: Proceedings of the First IEEE International Conference on Pervasive Computing and Communications*, p. 378, 2003.

[2] P. Välkkynen and T. Tuomisto, "Physical browsing research," 2005.

[3] G. D. Abowd, C. G. Atkeson, J. Hong, S. Long, R. Kooper, and M. Pinkerton, "Cyberguide: a mobile context-aware tour guide," *Wirel. Netw.*, vol. 3, no. 5, pp. 421–433, 1997.

[4] T. Kindberg, J. Barton, J. Morgan, G. Becker, D. Caswell, P. Debaty, G. Gopal, M. Frid, V. Krishnan, H. Morris, J. Schettino, B. Serra, and M. Spasojevic, "People, places, things: web presence for the real world," *Mob. Netw. Appl.*, vol. 7, no. 5, pp. 365–376, 2002.

[5] E. Tuulari and A. Ylisaukko-oja, "Soapbox: A platform for ubiquitous computing research and applications," *Pervasive '02: Proceedings of the First International Conference on Pervasive Computing*, pp. 125–138, 2002.

[6] G. Pospischil, M. Umlauft, and E. Michlmayr, "Designing lol@, a mobile tourist guide for umts." *Mobile HCI*, pp. 140–154, 2002.

[7] K. Cheverst, N. Davies, K. Mitchell, A. Friday, and C. Efstratiou, "Developing a context-aware electronic tourist guide: some issues and experiences," *CHI*, pp. 17–24, 2000.

[8] J. Baus, K. Cheverst, and C. Kray, *A Survey of Map-based Mobile Guides Map-based mobile services - Theories, Methods and Implementations*. Meng/Zipf, Springer, 2005, ch. 13.

[9] L.-W. Chan, Y.-Y. H. Hsu, Y.-P. Hung, and J. Y.-j. Hsu, "Orientation-aware handhelds for panorama-based museum guiding system," *Ubi-Comp 2005 Workshop: Smart Environments and Their Applications to Cultural Heritage*, September 2005.

[10] J. Rekimoto and Y. Ayatsuka, "Cybercode: designing augmented reality environments with visual tags," *DARE '00: Proceedings of DARE 2000 on Designing augmented reality environments*, pp. 1–10, 2000.

[11] N. Davies, K. Cheverst, A. Dix, and A. Hesse, "Understanding the role of image recognition in mobile tour guides," *Proceedings of the 7th international conference on Human computer interaction with mobile devices & services*, pp. 191–198, 2005.

[12] H. Bay, B. Fasel, and L. V. Gool, "Interactive museum guide," *The Seventh International Conference on Ubiquitous Computing UBICOMP, Workshop on Smart Environments and Their Applications to Cultural Heritage*, September 2005.

[13] ——, "Interactive museum guide: Fast and robust recognition of museum objects," *Proceedings of the first international workshop on mobile vision*, May 2006.

[14] B. Fasel and L. V. Gool, "Interactive museum guide: Accurate retrieval of object descriptions," *Adaptive Multimedia Retrieval*, pp. 179–191, 2006.

[15] G. Fritz, C. Seifert, and L. Paletta, "A mobile vision system for urban detection with informative local descriptors," *ICVS '06: Proceedings of the Fourth IEEE International Conference on Computer Vision Systems*, p. 30, 2006.

[16] J. S. Hare and P. H. Lewis, "Content-based image retrieval using a mobile device as a novel interface," *Proceedings of the first international workshop on mobile vision*, pp. 64–75, Dec. 2004.