



Mobile Phone-Enabled Museum Guidance with Adaptive Classification

**Erich Bruns,
Benjamin
Brombach,
and Oliver
Bimber**
*Bauhaus-
University
Weimar*

Over the past four years, we developed PhoneGuide, an adaptive museum guidance system that uses mobile phones for on-device object recognition (see Figure 1). PhoneGuide allows museum visitors to use their own mobile phones for retrieving information about exhibited objects. This is possible by taking a photograph of presented objects; the system then recognizes them automatically using image classification techniques. After identifying an object, PhoneGuide provides corresponding multimedia information, such as replayed audio or displayed text and video content. Compared to audio guides, we believe that this approach is more intuitive and more flexible for museum visitors, and more economic for museum owners.

In contrast to several related methods, our system performs classification directly on the local phones instead of sending images to a remote server—as such, classification requests don't have to be processed sequentially on a server (which can lead to unacceptably long response times). This makes our approach scalable with respect to the number of users.

But rather than focusing solely on the object recognition task from an image-processing point of view, we created a self-improving adaptive sensor network system that supports image recognition with user feedback, ad hoc communication, and location information.

This approach lets us apply a relatively simple but fast classification algorithm based on global color features and artificial neural networks. It reaches a maximal recognition rate of 92.6 percent for identifying 139 museum objects from different perspectives and distances and under realistic conditions. In contrast to emerging related approaches that apply local feature extraction techniques, such as Speeded Up Robust Features (SURF), in combination with a nearest-neighbor matching strategy,¹ our system is much faster (approximately by factor of 5) and more scalable than these approaches given that the classification performance doesn't

decrease when the number of training images per object increases.

Consequently, a continuous collection of captured image data, together with user feedback, is possible in our system and leads to specialized and robust image classifiers, such as neural networks, over time. This lets the system automatically adapt to different environmental situations within a museum—such as varying lighting conditions—that it couldn't initially capture. User studies (which we'll examine in more detail later) strengthen our approach and also revealed that visitors won't tolerate waiting times for recognition that are more than 2.1 seconds on average.

To achieve realistic classification rates in dynamic and complex public environments (indoors or outdoors) as well as an applicable performance rate, an intelligent system adaptation is an essential component. This article gives an overview of the PhoneGuide system's different components.

Adaptive classification

The PhoneGuide system's major challenge is to locate and recognize museum objects automatically in captured images. Often hundreds or even thousands of objects must be reliably classified under varying lighting conditions and from arbitrary perspectives and distances. Small objects located in showcases, for instance, can't be photographed separately and must be distinguished automatically from each other in a single image. The object recognition process becomes even more demanding if a mobile device's computational possibilities are restricted.

To overcome these limitations, we developed an adaptive classification infrastructure (see Figure 2) that continuously collects image data and user feedback to adapt and improve the local classification process over time. This is a crucial element because we can't reliably predict how visitors approach exhibits or how objects appear over time when influenced, for instance, by daylight. Additionally, our system carries out a two-step recognition process for identifying multiple objects in one

image and shares classification results with other users through ad hoc phone-to-phone networks to improve the recognition rate.

PhoneGuide's adaptive classification infrastructure consists of a stationary server and an arbitrary number of mobile phones and sensor boxes. The server continuously carries out two main tasks. First, it constantly collects and stores adaptation parameters (user feedback, image data, and lighting information) that individual mobile phones have gathered during runtime. As an outcome of the mobile object identification, a probability-sorted objects list is displayed on the phone after the user takes a photograph of an exhibit. The user selects the correct object from among others (represented as a photograph) from this list with a minimum number of clicks. The mobile application stores the correct object IDs as well as the captured images, and then transmits the data from the phones to the server when the user leaves the museum. Note that no on-line connection exists between the mobile devices



Figure 1. The basic concept of the PhoneGuide system in a museum. Our system uses adaptive classification in dynamic large-scale museum environments. Ad hoc sensor networks and phone-to-phone communication support PhoneGuide.

and the server during runtime. This sets our system apart from related approaches that transmit captured images to a remote server for the recognition process.² Second, the server applies the collected adaptation parameters for creating and improving the required

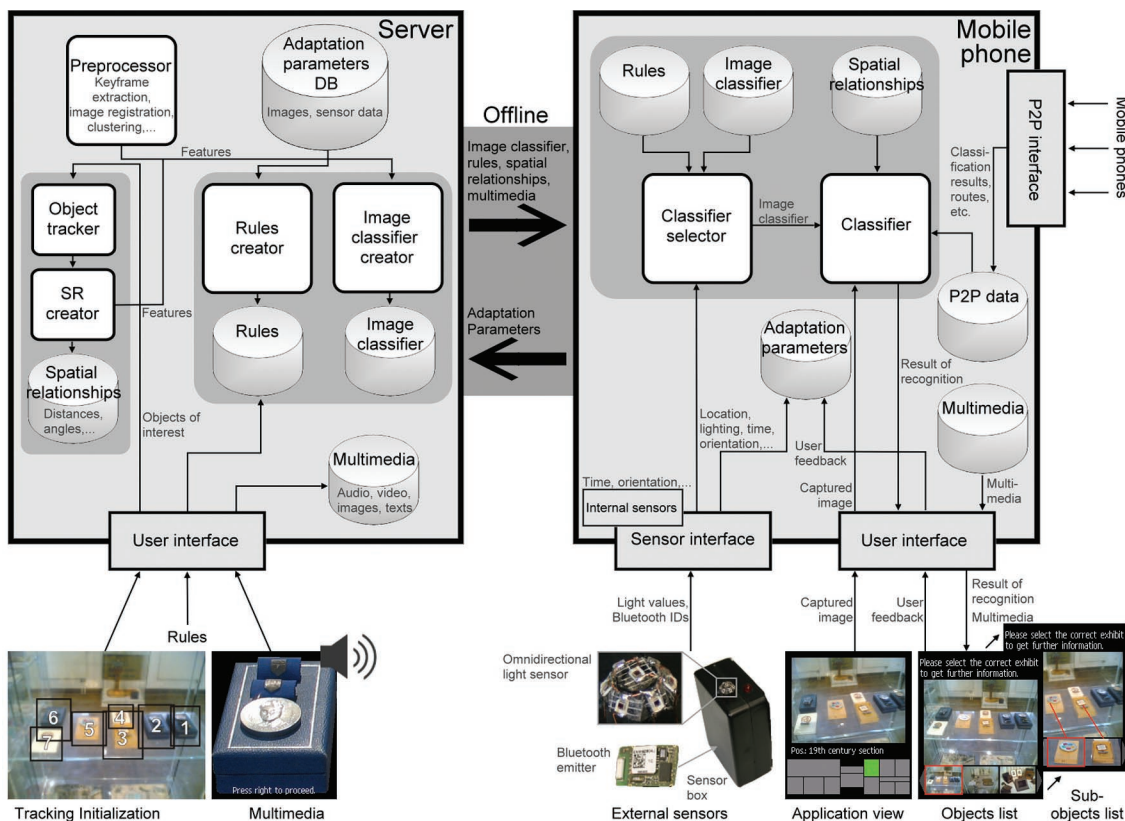


Figure 2. Overview of the adaptive classification infrastructure. During application, each phone collects position and lighting data from distributed sensor boxes, as well as user feedback and image data on the local devices. When leaving the museum, this information is transmitted to the server, which stores and applies the gathered information to generate and improve the required classification elements. The adapted elements are transmitted to new visitors' mobile phones upon entering the museum. Through the phone-to-phone interface, mobile devices broadcast evaluated classification data to improve the local recognition process. Sensor boxes provide information on the current user's location and lighting conditions.

classification elements, such as image classifiers, rules, and spatial relationships, offline. The server transmits the improved elements to the mobile phones of new visitors when entering the museum.

To accomplish these tasks, the server consists of three major components. One module handles the preprocessing steps (the object tracker and spatial-relationship creator) that classify multiple objects in a single image. Furthermore, the server possesses a second module called the preprocessor that prepares image data (keyframe extraction and clustering) for training image classifiers. The third module dynamically creates rules and image classifiers based on the adaptation parameters. The rules, which currently consist of a lookup table, determine which classifier must be selected for the current user's location.

The front-end application on the mobile device of-

The front-end application on the mobile device offers a user interface and tracks actual recognition results, and provides user feedback.

fers a user interface and tracks actual recognition results (unnoticed by the user), and as mentioned earlier, provides user feedback. Scene recognition is performed when users take a photograph; the mobile application then divides it into subpatches and extracts global color features from each patch, which serve as input for three-layer neural networks.³ Besides this function, classifiers can use spatial relationships to identify multiple objects within the captured image after the scene classification. We apply the phone-to-phone interface for exchanging local classification results dynamically during runtime (that is, without users having to check in or out at the server when entering or leaving the museum). These parameters won't be used for retraining the classifiers on the phone (as they would be used on the server) but for adapting pretrained classifiers to handle momentary situations in the museum. Sensor boxes in the museum provide information to determine the current user's location. In addition, we're currently working on using illumination sensors attached to the sensor boxes (see Figure 2). The sensors record the illumination state and transmit the data to the mobile phones. With only minor additional computation time, our recognition process will become invariant to extreme illumination situations, given that the phones automatically select

the optimal classifier that was trained for the corresponding illumination state.

Perspective invariance and large-scale classification

In practice, our system must be flexible enough to compensate for individual user behavior. The ways in which visitors approach and observe an object can vary greatly, leading to significantly different perspectives in photographs taken for classification. For ensuring an initially acceptable recognition rate, the classification process must be scale and perspective invariant for possible user locations.

To accomplish this, we apply mobile phones' video-capturing functionality to record videos containing multiple perspectives and distances of each museum object.⁴ The server preprocesses these videos by extracting and clustering keyframes, as indicated previously. The aim is to eliminate redundant frames and select those frames that contain descriptive perspective and scale information. The remaining frames are forwarded to the image classifier creator which—based on these frames—configures and trains an optimized classifier.

Although the system adapts to individual user behavior for each object, one classifier can't cope with hundreds or thousands of objects in a museum. Therefore, the sensor boxes provide information required to determine the users' rough locations through a simple pervasive tracking method.³ Each box is equipped with a Bluetooth chip that transmits a unique ID to all mobile phones located in its signal range (up to 10 meters). All sensor box IDs in the network together with their known positions and signal ranges span a coarse grid of possibly overlapping signal cells. Estimating the cells in which a phone is currently located by analyzing all detectable sensor boxes indicates to each device its own rough position within the museum. Consequently, for each cell, one classifier is trained and the correspondences are stored in a lookup table. Before the classification is carried out on the mobile phone, the application selects the correct image classifier on the basis of the current location. In our test environment, an accuracy of the location estimation of 5 to 7 meters was adequate to ensure a constant recognition rate independent of the number of objects.

Subobject recognition

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't take photographs of individual objects without capturing other ob-

jects simultaneously. To overcome this, our framework identifies multiple objects (subobjects) in one image. Classifying subobjects happens in a two-step recognition process; after deriving the correct scene context when recognizing a group of objects, the mobile application automatically classifies the individual subobjects in the photograph. It then labels the results and links the subobjects with a subobject list (see Figure 2). From this list, the user can finally select the object of interest, prompting the system to present multimedia content.

Our subobject classification technique is based on spatial relationships that give us the opportunity to recognize similar objects. They are precomputed offline on the server by automatically tracking individual subobjects through all video frames, and by computing relative geometric relationships, such as maximal search angles and distances between all subobjects. Additionally, the server determines and stores the size of each subobject's bounding box and its individual classification features. The system then uses these features to train individual classifiers. The trained classifiers—the scales as well as the spatial relationships—are transmitted to the mobile devices during updates with the server.

Classifying subobjects on the mobile device is performed by the application by identifying an anchor object first, which it assumes is located in the center of the photograph. To cope with different scales, we apply multiresolution classification to the captured image, thereby ensuring perspective invariance, as explained earlier. We then use the scale and position of the anchor object's bounding box to select the correct spatial relationships. Based on the maximal search angles and distances to neighboring subobjects, a search mask shifts spirally around the initial position until the classifier's excitation is above a predefined threshold for a certain position. This indicates that a new subobject is found if, in addition, the excitations for neighboring search points are lower than for this position. The application considers possible rotations by aligning the spatial relationships to the mobile phone's current orientation. We can detect rotations using either a built-in accelerometer or by testing for the most likely rotations (90, -90, or 180 degrees). Because the spatial relationships can be optimized continuously the more subobjects have been detected, the classification process will speed up with each detected subobject. The adjusted spatial relationships can be stored on the phone and transmitted to the server as part of the adaptation process.

Phone-to-phone communication

To improve the classification process during run-

time, the phone-to-phone communication lets the system broadcast current classification results of individual phones to all other phones that are currently in the proximity.

The implicit user feedback leads to a mapping between the object the user selects and the probability-sorted objects list suggested by the classification process. The application weights the candidates on the objects list on the basis of their rank, and they are then added to corresponding entries of a correlation matrix. These entries correlate selected and recognized objects and are concatenated with the probability lists of new classification results. This leads to a continuous and ad hoc adaptation of the local classification process, given that local user feedback as well as the correlation matrices broadcasted from other phones are frequently merged while visitors are moving through a museum. As for the location esti-

In cooperation with the City Museum of Weimar, we were able to test and evaluate our system during regular opening hours for the past two years.

mation, this process is carried out in the background and remains unnoticed by the user.

In cooperation with the City Museum of Weimar, we were able to test and evaluate our system during regular opening hours for the past two years. In our current implementation, we apply a well-selected set of global image features (mean and variance in RGB color channels, as well as a 10-bin color histogram) extracted from images that have a resolution of 160 to 120, and three-layer neural networks for image classification. In total, we applied 7,464 frames (min: 15, max: 80, average: 51.8 per object) for training the neural networks. During evaluation, nine sensor boxes spanned 16 different location cells. The entire size of all necessary classifiers was 350 Kbytes. This outperforms approaches that apply local image descriptors (because they have to store thousands of descriptors individually). On Nokia 6630 mobile phones, our local object recognition algorithm implemented in Java 2 Micro Edition requires on average 3.8 seconds (including a duration for capturing the image and presenting the objects list).

On newer phones (such as the Nokia N95), the feature extraction takes 370 milliseconds (on a

Nokia 6630, this takes 2250 ms) and the classification 100 ms (on a Nokia 6630, this take 200 ms) for 34 trained objects within the same location cell. For 139 objects, we achieved a recognition rate of 92.6 percent for users who were familiar with our system and 82 percent for totally inexperienced museum visitors. In the context of an additional user study we carried out,⁴ we achieved these results under realistic conditions; with arbitrary perspectives and scales and evaluated over four business days at different times and illumination situations. Twelve of these 139 exhibits consisted of three to eight sub-objects (on average: 5.4); we achieved a recognition rate of 92.3 percent (85.9 percent for inexperienced users) for identifying these sub-objects. Depending on the number of sub-objects, their recognition requires 5.2–5.8 seconds on a Nokia 6630 and approximately 2.0–3.0 seconds on a N95.

On the basis of additional tests, we can also show that a temporal adaptation through the server leads to a continuous improvement of the recognition rate over time.⁴ The phone-to-phone updates result in quick adaptations to spontaneous changes. In initial experiments with 34 objects, the classification rate increased from 92.0 percent to 97.1 percent after only three updates. Besides estimating the quantitative benchmark data, we were interested in the subjective impression of museum visitors after using our system. Therefore, we asked 15 subjects (average age 27.4, seven females and eight males) to fill out a questionnaire and to rate different characteristics of the system (1 = worst, 7 = best) during our user study. Handling (rated 5.9), subjective recognition rate (5.8), and overall performance (6.1) were highly

ranked. The subjects could also well imagine that PhoneGuide can be used instead of audio guides in museums (5.8). The most criticized aspect of our approach was the relatively long waiting time required for the device localization. In our current implementation, it takes approximately 13 seconds (depending on the number of Bluetooth devices) to scan nearby Bluetooth emitters. This waiting time occurs only during transitions between signal cells—the waiting time for the recognition process remains constant. In addition, another user study with 18 subjects revealed that they're not willing to concede much time for the overall recognition process: 11 percent of all subjects would prefer a recognition time of less than 1 second, 50 percent of 1 to 2 seconds, 33 percent of 2 to 4 seconds, and 1 subject would accept 4 to 6 seconds.

The most related approach to our system has implemented an improved version of SURF on current mobile phones for outdoor applications.¹ However, because it uses a nearest-neighbor matching strategy, adaptive approaches can't be performed without a loss of classification performance or runtime. The size of our classification data scales much less with an increasing number of objects and doesn't increase at all with a rising number of training images per object. This benefits a fast wireless transmission, if necessary. ❏

Acknowledgments

The Stiftung für Technologie, Innovation und Forschung Thüringen (STIFT) supported the PhoneGuide project. Further information is available at www.uni-weimar.de/medien/AR.


References

1. W.-C. Chen et al., "Efficient Extraction of Robust Image Features on Mobile Devices," *Proc. Int'l Symp. Mixed and Augmented Reality*, 2007, IEEE Press, pp. 287–288.
2. G. Fritz, C. Seifert, and L. Paletta, "A Mobile Vision System for Urban Detection with Informative Local Descriptors," *Proc. 4th IEEE Int'l Conf. Computer Vision Systems*, IEEE Press, 2006, p. 30.
3. E. Bruns et al., "Enabling Mobile Phones to Support Large-Scale Museum Guidance," *IEEE MultiMedia*, vol. 14, no. 2, 2007, pp. 16–25.
4. E. Bruns and O. Bimber, "Adaptive Training of Video Sets for Image Recognition on Mobile Phones," to appear in *J. Personal and Ubiquitous Computing*, 2008; www.springerlink.com/content/x166871351767265/.

Contact author Oliver Bimber at bimber@uni-weimar.de

Contact the department editors at cga-vr@computer.org.

WHAT'S AHEAD



Mesh networking: July/Aug. 2008
Service mashups: Sept/Oct. 2008
Data stream management:
Nov./Dec. 2008

Visit us at
www.computer.org/internet/