PhoneGuide: Museum Guidance Supported by On-Device Object Recognition on Mobile Phones

Paul Föckler, Thomas Zeidler and Oliver Bimber
Bauhaus-University Weimar, Media Faculty, Dept. Augmented Reality
Bauhausstr. 11, 99423 Weimar, Germany
{foeckler, zeidler2, bimber}@uni-weimar.de

ABSTRACT

We present *PhoneGuide* – an enhanced museum guidance approach that uses camera-equipped mobile phones and on-device object recognition.

Our main technical achievement is a simple and light-weight object recognition approach that is realized with single-layer perceptron neuronal networks. In contrast to related systems which perform computational intensive image processing tasks on remote servers, our intention is to carry out all computations directly on the phone. This ensures little or even no network traffic and consequently decreases cost for online times. Our laboratory experiments and field surveys have shown that photographed museum exhibits can be recognized with a probability of over 90%.

We have evaluated different feature sets to optimize the recognition rate and performance. Our experiments revealed that normalized color features are most effective for our method. Choosing such a feature set allows recognizing an object below one second on up-to-date phones. The amount of data that is required for differentiating 50 objects from multiple perspectives is less than 6KBytes.

Categories and Subject Descriptors

H.5.2 [**Information Interfaces and Presentation**]: User Interfaces—*Input devices and strategies*; I.2.10 [**Artificial Intelligence**]: Vision and Scene Understanding.

General Terms

Design, Evaluation, Accuracy, Performance, Application

Figure 1. Application of PhoneGuide in a museum.

Keywords

Mobile phones, object recognition, neural networks, museum guidance

1. INTRODUCTION

More than 500 million mobile phones have been sold worldwide in the year 2004 [MIT04]. It has been estimated that by the end of the year 2005 over fifty percent of all cell phones will be equipped with digital cameras [Mac04].

Today, a large variation of communication protocols allows the transfer of data between individual units, or accessing larger networks – such as the Internet. Leading graphics board vendors are about to release new chips that will enable hardware-accelerated 3D graphics on mobile phones – including geometry processing and per-pixel rendering pipelines. Some exotic devices even support auto-stereoscopic viewing, GPS navigation, or scanning of RFID tags. Due to the rapid technological advances of cell phones, the distinction between PDAs and mobile phones might soon be history.

Obviously, mobile phones are becoming platforms that have the potential to bring various new applications to a mass market. This will influence areas, such as entertainment, edutainment, service, and many others.

Audio guides, for instance, are other mobile devices that are frequently being utilized by museums to communicate additional information about exhibits. Visitors select audio sequences of particular pieces by keying in individual identification numbers that have to be selected from a catalog or a map, or are presented on the spot. Such devices hold several drawbacks to both – the museum visitors and the museum operators: First, audio guides can present auditory information only – excluding other effective presentation forms, such as written text, images, computer graphics, videos, and interactive applications. Second, they require the visitors to look up and key in the exhibit individual identification number – which can be annoying after a while. Third, the expense for acquisition and especially maintenance of such devices has to be covered by the museum.

2. MOTIVATION

To overcome these drawbacks, we propose to apply camera-equipped mobile phones to support enhanced museum guidance. We call this concept *PhoneGuide*.

Today's mobile phones allow presenting advanced visual information (2D/3D graphics, videos, text, animations, etc.) on their displays in addition to the audio provided through the build-in speakers or head-sets. This potentially provides a more efficient communication of information than possible with auditory information only.

Camera phones in combination with computer vision techniques make it possible to automatically recognize exhibits in photographs, instead of forcing the visitors to look up and key in identification numbers. Essential for such an approach is that the automatic recognition of exhibits is robust, fast and scalable. Furthermore, it is desirable to perform the entire application directly on the phone to avoid the continuous transmission of data between the end device and a server during runtime. This would cause additional costs to the user for the utilization of a provided communication service.

Since the visitors bring along their own phones, the acquisition and maintenance cost required for special purpose devices will be reduced or even eliminated.

Visitors need to install the necessary information (application, presentation content and identification data) on their phones. This can be done through a memory card, by downloading them from a local base-station at the museum entrance or over a global communication service. Thereafter, the application is independent from global services.

This paper mainly focuses on indoor museum applications.

The remainder of this paper is organized as follows: Section 3 will discuss the related and previous work that has been carried out in areas such as image retrieval and object recognition. Section 4 will present our light-weight on-device object recognition approach that is suitable for mobile phones. Section 5 will discuss the recognition results and performance that can be achieved with our method. Implementation details on the software framework are described in section 6. Finally, section 7 concludes our work and gives an outlook on future extensions.

3. RELATED WORK

This section discusses related computer vision techniques for stationary and mobile devices first. Alternative guidance systems that utilize mobile devices are explained next. Finally, we differentiate our particular approach from the presented ones.

3.1 Computer Vision

Object recognition is a wide field of computer vision. Objects are normally analyzed by extracting *global* or *local features* from an image before being recognized.

Today's recognition systems mostly apply *local features* (e.g., local corner points or image fragments) which can be affine invariant and support the recognition of partly occluded objects.

One of the first recognition systems using local features was proposed by Schmid and Mohr [Sch97] who used local gray value feature points extracted with a Harris corner detector [Har88] for image retrieval. These features are rotational invariant and the system provides a robust recognition.

Lowe [Low99] presented an algorithm to detect local scale invariant features based on local extrema found in Gauss-filtered difference images. Later Lowe [Low04] demonstrated the possibility to extract highly distinctive features (SIFT – scale invariant feature transform) that could be matched in a large database with a high hit rate. Helmet et al. [Hel04] uses this method for object classification using a stochastic model.

Fritz et al. [Fri04a] use discriminative regions to analyze an object. These regions are calculated in two steps. First, small sub-images are mapped into a subspace by a principal component analysis (PCA) to receive a low dimensional vector. Second, the entropy is estimated for the components of this vector. After

this calculation a region is selected as a local feature if the entropy is above a threshold.

Many other examples for object recognition systems that are based on local features can be found today (e.g. Mikolajcyk et al. [Mik02]). To discuss them all is out of the scope of this paper.

Object recognition techniques that use *global features* are based on image properties such as color, texture, or gradient images.

Swain et al. [Swa91], for instance, presented a recognition system based on color histograms. Another system uses receptive field histograms [Sch96]. Beside object recognition global feature extraction is applied for context based image retrieval [Iqb02]. Lehman et al. [Leh00] uses global features to categorize medical images.

A combination of local and global features is shown in [And03] where a mobile robot recognizes indoor object by calculating local features for tracking and global features (histogram of pixel velocity) for pattern recognition.

3.2 Computer Vision with Mobile Devices

Object recognition performed directly on mobile devices, such as PDAs or mobile phones is nearly unexplored. This may be due to the hardware limitations for such devices. Therefore, most approaches use the mobile device for image capturing, simple pre-computation, and streaming tasks only, while a powerful remote server performs the computational intensive classification. Fritz et al. [Fri04b] proposed such a system for recognizing outdoor objects like buildings and statues using a PDA and a wireless connection (WLAN/GPRS/UMTS) to a server. The server classifies the objects using the discriminative regions described in [Fri04a] and sends back the results to the PDA.

Seifert et al. [Sei04] introduced a system to detect and classify road signs, using a PDA for image capturing and a server for carrying out all computations. Similar as in Fritz et al. [Fri04a], an image is analysed by local regions which are detected via pixel classification using trained color filters. Two simple extractors (ellipse fitter and Hough transform) are applied to retrieve the shape of the sign. The detected sign is classified by matching the object's pattern to reference patterns in a road sign database.

Various ongoing initiatives follow the same principle, but use mobile phones instead of PDAs. Lowe's distinctive image feature method [Low04] is sometimes being applied for recognition on the server side [Bon04].

Corrs et al. [Cor00] use pattern matching on a wearable tablet computer to determine position and line of sight of a user in an outdoor environment. In a first step reference images from different views and angles are taken with known external camera parameters determined via GPS. In the second step, a relative transformation between the two cameras is calculated by matching the new image pair-wise (with unknown extrinsic camera parameters) to the reference images. Using this transformation the new user position and line of sight is calculated.

Some computer vision methods can be performed locally on the mobile device itself. The Semacode Cooperation [Sem04], for instance, developed a system for mobile phones that recognizes a URL encoded in a printed barcode. Therefore, three steps are performed (node localization, correction for image distortion, and raw data retrieval of a node) before the information can be decoded.

Several groups perform marker detection and tracking to support augmented reality applications with mobile-phones or PDAs. Some carry out the main computations on remote servers [Wag03, Ass03] while others perform all tasks directly on the mobile device itself [Moe04, Wag03].

Finally mobile augmented reality games (live video-stream augmented with virtual objects e.g. Mosquito Hunt [Sie04] or Kickreal [Cla05]) run directly on mobile phones using simple computer vision algorithms like optical flow [Sie04] or simplified marker-less tracking.

3.3 Guidance with Mobile Devices

Guidance systems (e.g., in museums, tourist navigation in cities, etc.) are one potential application of object recognition with mobile devices. Beside computer vision, other technologies can be applied to provide spatial awareness:

Bombara et al. [Bom03], for example, use infra-red (IR) beamers for determining the position of a user to provide additional information in museums. If the user enters the small zone of an IR beamer and this beam can be detected by a PDA or cell phone, the location information is transferred to the mobile device. Infrared beamers, however, are active and require batteries to operate.

Passive radio frequency identification (RFID) tags are activated if a mobile device carrying an RFID reader gets in range. They receive power through the radio

signal from the RFID reader. With the provided ID number additional local information can be displayed on the mobile device, as demonstrated for PDAs in [Mnh04].

The above approaches are normally not well suited for large scale outdoor applications. The position of a mobile user in an urban environment can be estimated by the global positioning system (GPS) ([Fri04b], [Cor00] or [Fei97]). With GPS data only, however, it is not possible to determine the direction of the device. Computer Vision can be used for computing the line of sight in addition. In the approach of Feiner et al. [Fei97] different sensors and a back-pack computer are used to track the motion of the user´s head. Corrs et. al. [Cor00] calculated the direction by pattern matching. Fritz et al. [Fri04b] research the possibility of object recognition to compute the line of sight of a user. GPS is not suited for indoor applications.

A good overview of such approaches is given in [Aok00].

3.4 Our Approach

Our intention is to support indoor guidance applications for museum navigation using computer vision. We intend to avoid the integration of passive or active reference markers into the environment, as done for other approaches such as [Bom03] or [Mnh04]. Thus, in contrast to most related object recognition systems (such as Fri04b, [Bon04], [Cor00] or [Sei04]), we perform all computations directly on the local device – a mobile phone. This avoids a network connection to a remote server and consequently reduces network traffic and the cost for online-times.

A local recognition on mobile phones is hardly possible with the approaches described above. The performance of Coors' pattern matching method [Cor00], for instance, decreases quickly with an increasing number of images. This algorithm is not scalable because a pair-wise matching is used.

Although the recognition rate is about 90 percent for 20 objects and 36 views, the approach by Fritz et al. [Fri04a] is not suitable for direct computation due to the hardware limitations of today's mobile phones. The calculation of discriminative regions and recognition of one object takes about 2-3 seconds on a personal computer [Fri04b].

A calculation of local features with SIFT [Low04] on a mobile phone is possible if the number of features is limited. But up until now the recognition rate on a mobile phone is about 50 percent [Bon04]. This is too low for an application in museums. Another problem is the computation time required for classification which can take up to several minutes [Bon04].

The approaches of Schmid et al. [Sch97] and Lowe [Low99] provide, similar as in [Low04], a high recognition rate on a workstation computer (over 90 percent). However, the performance problems of mobile phones still remain. The recognition time of Lowe's method [Low99] for detecting three different objects in one image, for instance, is about 1.5 seconds on a workstation.

Yet another problem exists in relation with several recognition approaches: The feature extraction has to be limited to the object. Thus, there is additional computation cost for extracting an object out of an image (e.g., through segmentation of foreground and background).

Seifert et al. [Sei04] presented an approach for a very special problem. Road signs have limited shapes (e.g. ellipses, rectangles etc.), thus, it is possible to restrict the feature extraction for regions with such a shape. This is not useable in general because an arbitrary object can have any shape.

4. OBJECT RECOGNITION

To recognize an object, multiple images are taken from different perspectives. Each image is decomposed into a fixed set of normalized *features*. Such a *feature vector* ($\vec{f}$) is unique for one particular image. Consequently, multiple feature vectors are related to one object. These feature vectors can be used directly for recognition algorithms that apply a closest-neighbor match strategy. Such algorithms determine the image from the previously taken ones, which comes closest to a new image, based on their feature vectors. Since our goal is the recognition of the object, rather than a particular image of it, more efficient methods can be found. Thus, we follow a *linear separation* strategy implemented with a single-layer artificial neural network on the phone rather than a closest neighbor match. The training of such a network allows compressing all feature vectors that belong to the same object into a single set of normalized weights. This *weight vector* ($\vec{w}$) is assigned to a single object, rather than to a single image and serves as a fingerprint for recognizing the same object in other images. It has the same dimension as the feature vector.

Note that we do not differentiate between background and foreground elements within an image. Thus, we recognize the entire *image content*, rather than particular foreground objects.

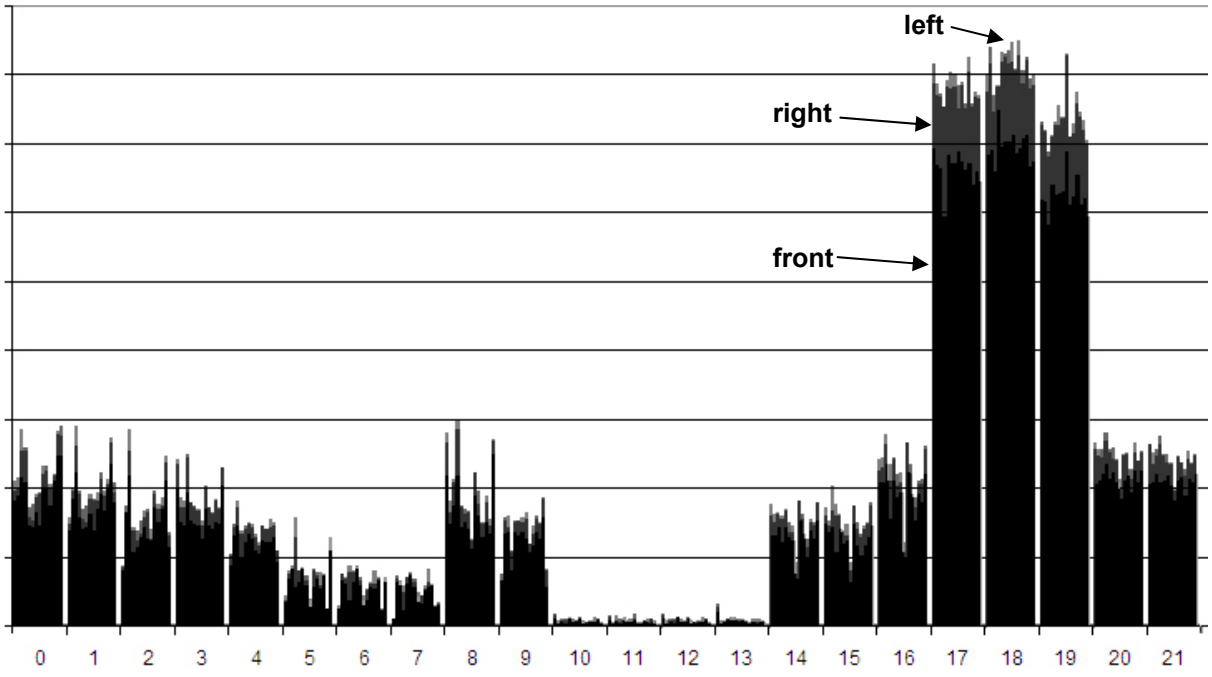

Figure 2. All 22 features computed for all 14 test objects (cf. figure 4) for three perspectives.

In the following, our recognition approach is described in detail.

4.1 Features

An optimal selection of features is essential for achieving a high recognition rate. We have identified and investigated several normalized features that are suitable for recognition with a linear separation strategy. These features describe different color and intensity relations as well as structural properties of the image content. We will describe these features and evaluate their efficiency for achieving an optimal recognition rate. Note that each feature is normalized to a range of 0..1.

4.1.1 Color and Intensity

The means of the red, green and blue color channels allow identifying the *absolute color portions* of the image:

$$f_{0,1,2} = \frac{\sum_{y=0}^{Y-1}\sum_{x=0}^{X-1} I_{[0,1,2],[x,y]}}{XY}$$

where X and Y is the image resolution and $I_{0,1,2}$ refers to the response in the red, green and blue color channels. Note that the mean value for a gray scale composition of the image is always 0.5. This is due to the automatic white balancing performed by the phone, which is beneficial for our approach. But the constant gray scale mean, however, cannot serve as a feature.

The absolute means can be set in relation as follows:

$$f_3 = \frac{f_0}{f_0 + f_1 + f_2}\ ,\ f_4 = \frac{f_2}{f_0 + f_1 + f_2}$$

These features describe the relation of the red and blue means relative to all others. Note that a *color ratio* for the remaining channel is given implicitly and is redundant.

The *color variance* of each channel is then computed with

$$f_{5,6,7} = \frac{\sum_{y=0}^{Y-1}\sum_{x=0}^{X-1} (I_{[0,1,2],[x,y]} - f_{0,1,2})^2}{XY}$$

and can be set into relation in a similar way as the mean values:

$$f_8 = \frac{f_5}{f_5 + f_6 + f_7}\ ,\ f_9 = \frac{f_7}{f_5 + f_6 + f_7}.$$

Again, the *variance ratio* of the remaining channel is redundant.

The *intensity distribution* within a single color channel, as well as within the grey scale image can be derived by evaluating their histograms:

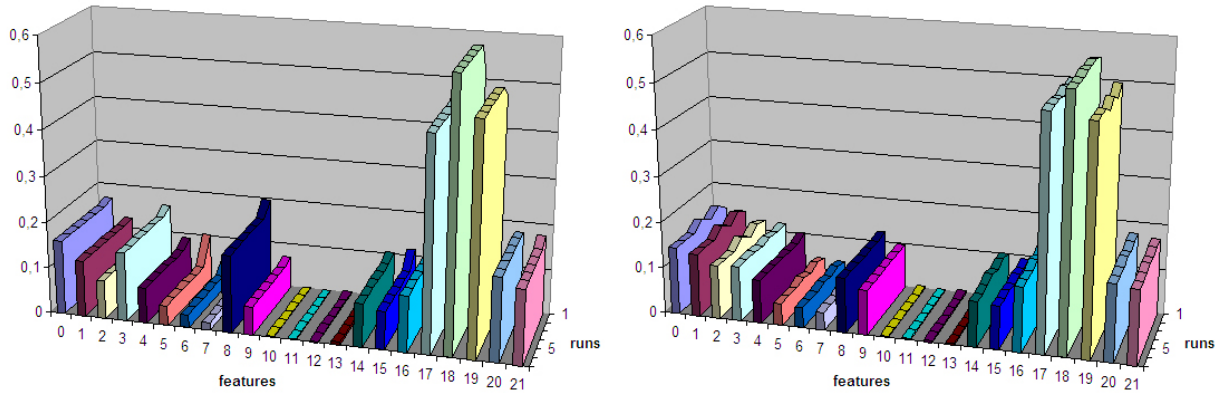

Figure 3. Convergence of weights over several runs: object A (left) and object B (right). Objects are highlighted in figure 4.

$$f_{10,11,12,13} = \frac{\max_{i=0}^{N-1}\left(H_{[gs,0,1,2],i}\right)}{XY}$$

Note that H_{gs} refers to the gray scale histogram, and $H_{0,1,2}$ to histograms of the red, green, and blue color channels. This ratio compares the maximum number of pixels with the same intensity in each color channel to all pixels. It is equivalent to comparing the peaks in the three histograms with their total areas. A value of N/XY indicates an equal shading distribution (where N is the number of different shades of each histogram), while a value of 1 indicates an absolute dominance of one shade.

4.1.2 Structure

To analyze the geometric structure of the image we compute the horizontal and vertical edges in the gray scale image with a Sobel operator.

We compute the average gradients in both directions with:

$$\bar{g}_{h,v} = \frac{\sum_{y=0}^{Y-1}\sum_{x=0}^{X-1} I_{[h,v],[x,y]}}{XY}$$

With the average gradients, we count the pixels that belong to edges with gradients that are above the corresponding average in horizontal only ($\bar{c}_h$), vertical only ($\bar{c}_v$), and horizontal *or* vertical ($\bar{c}_{hv}$) directions.

We defined another gradient threshold that allows identifying pixels which belong to relatively strong edges as:

$$\bar{g}'_{h,v} = \frac{\hat{g}_{h,v} + \bar{g}_{h,v}}{2}$$

where $\hat{g}_{h,v}$ is the maximum gradient of the image.

Using this threshold allows counting pixels that belong to relatively strong edges in horizontal only ($\bar{c}'_h$), vertical only ($\bar{c}'_v$), and horizontal *or* vertical ($\bar{c}'_{hv}$) directions.

With these parameters, we can define several structural relations.

The *proportion of structure information* compared to the entire image content can be defined as the ratios between edge pixels and all pixels:

$$f_{14} = \frac{\bar{c}_h}{XY},\ f_{15} = \frac{\bar{c}_v}{XY}, \text{ and } f_{16} = \frac{\bar{c}_{hv}}{XY}$$

These features, for example, allow differentiating noisy objects form smooth ones.

To differentiate between *strong edges and noise*, we can set the strong edges and all edges in relation with:

$$f_{17} = \frac{\bar{c}'_h}{\bar{c}_h},\ f_{18} = \frac{\bar{c}'_v}{\bar{c}_v}, \text{ and } f_{19} = \frac{\bar{c}'_{hv}}{\bar{c}_{hv}}$$

Finally, we compare the amount of *horizontal edges (moderate and strong) with the total amount of edges (moderate and strong)*:

$$f_{20} = \frac{\bar{c}_h}{\bar{c}_h + \bar{c}_v},\ f_{21} = \frac{\bar{c}'_h}{\bar{c}'_h + \bar{c}'_v}$$

Note that the corresponding ratios for vertical edges are redundant again, and do not have to be considered.

Figure 2 illustrates the computed features for the 14 test objects that are shown in figure 4. All features have been computed for three perspectives (front, left and right). Note that same features of different objects have been combined to point out their general behavior

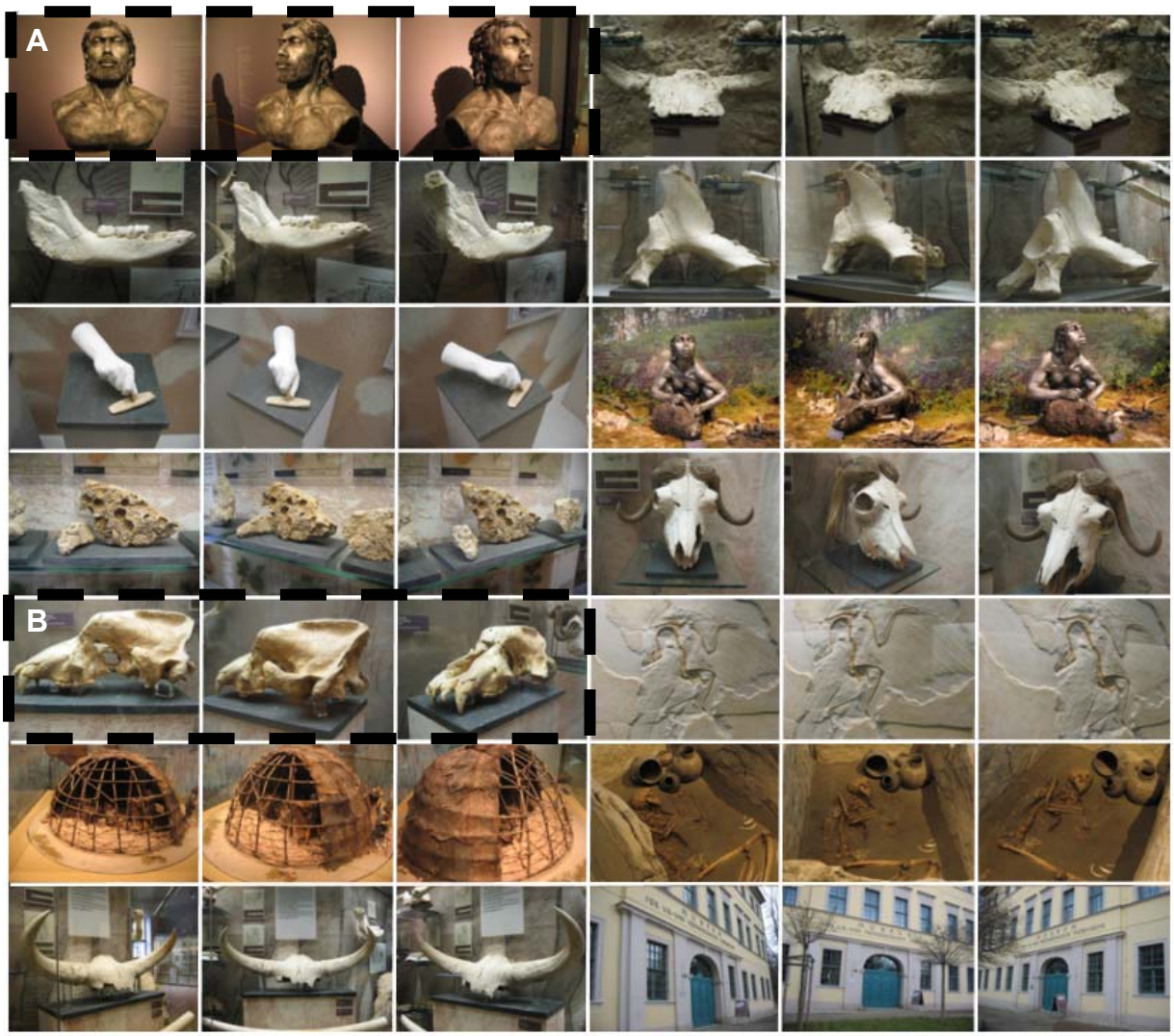

Figure 4. Fourteen representative test objects from three perspectives.

in relation to each other. The indicated different perspectives have been superimposed and appear in different gray scales in figure 2. The numbering equals the indexing used in this section. Their efficiency will be discussed in section 5.

4.2 Recognition and Training

Recognition of objects (image contents) is achieved through a linear separation strategy realized with a single-layer artificial neural network containing N *perceptrons*, where N is the number of recognizable objects. Thus, each object is associated with a single perceptron with $M = \dim(\vec{f}) = \dim(\vec{w})$ input channels and one output channel. Each input channel (j) of a perceptron (i) is assigned to a weight component of the corresponding weight vector ($w_{i,j}$). These weights are initialized with random numbers (normalized to a range of 0..1) before recognition and training.

For *recognition*, a new image of an object has to be taken and its feature vector is computed.

The object is recognized by finding the perceptron with the maximal output excitation over all perceptrons with the following activation function:

$$\max_{i=0}^{N-1}\left(\sum_{j=0}^{M-1} w_{i,j} f_j'\right), \text{ with } f_j' = \frac{f_j}{|f|}$$

If the recognition failed, the network has to be trained. The *training* is achieved by amplifying the weights of the perceptron that should be activated (but was not, because the object was not recognized). This is done by the following learning function:

$$w_{i,j}' = \frac{w_{i,j} + f_j'}{|w_{i,j} + f_j'|}, \text{ with } f_j' = \frac{f_j}{|f|}$$

Note that the training can be performed online or offline. The feature vectors that are computed for every image do not have to be stored. Only the weight vectors of all perceptrons have to be kept after training. Thus, the *data size* of the neural network does scale with the number of recognizable objects, but does not scale with the number of images taken!

Figure 3 shows how quickly the weight vectors for two different objects (A and B in figure 4) converge throughout the training (if all 14 objects are trained for 3 perspectives).

5. EVALUATION

We want to evaluate the impact of individual features with respect to the overall recognition rate and performance with a small set of test objects. All test objects have been captured with a resolution of 160x120 pixels from three different perspectives – front, left, and right. Our goal is to identify a feature set that leads to a maximum recognition rate.

With a selection of three out of seven best feature sets from this test series, we have carried out a more realistic experiment directly in a museum. The task was to recognize 50 different exhibits from arbitrary perspectives under realistic conditions.

5.1 Lab Experiment

To identify the combination of features that lead to the highest recognition rate, we use the 42 color photographs (resolution 160x120 pixels) shown in figure 4).

Based on the results illustrated in figure 2, we have composed seven different feature sets which have been tested for their total recognition rate and convergence behavior (see table 1).

An *automatic offline training* was performed for each feature set. The recognition rate was tracked over each run (one run is equivalent to training all 42 images with the corresponding feature set, as explained in section 4.2).

total recognition rate			**remarks**
1. run	*2. run*	*3. run*	
55%	100%	100%	color only (0-9)
55%	100%	100%	color and histogram only (0-13)
52%	95%	100%	all (0-21)
31%	81%	81%	structure only (14-21)
52%	95%	100%	no histogram (all except 10-13)
52%	95%	100%	all except 17-19
57%	100%	100%	all except 10-13 and 17-19

Table 1. Convergence and recognition rate for different feature sets using test images (cf. figure 4).

First, only the color features (with and without the histogram features) have been used for training. Then training was carried out with structure features only (without color features). Finally, all 22 features, as well as several combinations of color and structure features have been trained. Since features 10-13 (intensity distribution in histograms) and 17-19 (strong edges relative to all edges) show a strong divergence from the average range of the other features, we have tested their influence in particular. It can also be observed that features 17-19 show a stronger divergence for varying perspectives, which give an indication that they might not be well suited for our task.

As it can be seen in table 1, all feature sets (except structure features only) converge quickly to 100% recognition rate. After only 3 runs, all 14 objects could be recognized from the three perspectives without error.

5.2 Field Survey

From these seven feature sets, we selected three final sets that were of particular interest for further experiments: The *color features* (with histogram), *all 22 features*, and *all features except the diverging ones* (highlighted in table 1).

With these sets, we have carried out a field survey in a museum under realistic conditions: The sets were online trained on three different phones (Nokia 6600, Nokia 7610, and Nokia 6630) over 50 different objects using three general perspectives for each object. Thus 150 images (color, 160x120 pixels) were used for training per run. Due to the online training, these perspectives always varied slightly. This survey was carried out over four half working days (in the morning and in the afternoon/evening) to take differences of the environment illumination into account that occur throughout the hours and days. The recognition rate

and convergence was continuously recorded. The results are shown in table 2.

5.3 Size and Performance

In our implementation, each single weight has a size of 8 Bytes. Consequently, the size of the three weight vectors used for our survey described in section 5.3 is 112 Bytes, 176 Bytes and 120 Bytes respectively.

The total amount of data that is required for recognition depends only on the *number of objects (N)* that need to be classified - even if an n-fold of images is used to train the network. Thus, the size of the database which is stored on the phone (or needs to be transmitted to it) is no more than N-times the corresponding weight vector size. In our survey, this was 5.6 KBytes, 8.8 KBytes, and 6 KBytes respectively to recognize all 50 objects from multiple perspectives.

total recognition rate			**remarks**
1. run	*2. run*	*3. run*	
39%	83%	91%	color and histogram only (0-13)
41%	81%	83%	all (0-21)
35%	77%	78%	all except 10-13 and 17-19

Table 2. Convergence and recognition rate for different feature sets under realistic conditions.

This light data volume does not only ensure *low memory requirements* on the device, but also *little network bandwidth* (if frequent update of the data is required as explained in section 7), and *fast recognition rate*.

Performance (in s)			**remarks**
6600	*7610*	*6630*	
1,5	1,2	0,8	color and histogram only (0-13)
5,0	3,8	3,1	all (0-21)
4,5	3,5	3,1	all except 10-13 and 17-19

Table 3. Recognition performance for different feature sets on different phones.

Table 3 shows the recognition performance using the three different weight vectors on three different phones. Note that for computing the structure features a Sobel convolution operator has to be applied to the image twice – in horizontal and vertical direction. This causes the enormous increase of the run time for feature sets that contain structure information.

Note that about 95% of the recognition time is spent for the feature computation, while the remaining portion belongs to the actual recognition task. Since the feature computation is carried out only once for a new photograph, its running time is constant. Only the marginal recognition portion scales linear with the number of differentiable objects.

6. SOFTWARE FRAMEWORK

Our software framework was realized with Symbian C++ using the Nokia Series 60 developer platform, which is based on the Symbian operating systems. The final application can be executed on the phone in different modes:

The *presentation mode* returns an assigned identification number of the recognized object, after taking a photograph of it. The multi-media content that has been assigned to the ID is then played instantly. Each presentation implies visual and auditory information. A button on the keyboard allows pausing, continuing, fast-forwarding and rewinding the presentation. Head-sets can optionally be used for private audio presentations.

In the *training mode* the user has the possibility to add new perceptrons for each object that needs to be recognized and assign its individual ID. In addition, correctly recognized objects can be confirmed while incorrect recognitions can be trained. Beside this manual *online training* mode, on automatic *offline training* using a set of pre-taken images is possible.

7. CONCLUSION AND FUTURE WORK

We have presented *PhoneGuide* – an enhanced museum guidance approach that uses camera-equipped mobile phones and on-device object recognition.

The technical focus of this paper is on a light-weight object recognition method using single layer perceptron neural networks. We have shown that this method is capable of differentiating 50 museum objects from multiple perspectives and under realistic conditions with a recognition rate of more than 90%. The size of the data set that is needed for this task is less than 6KBytes, and the recognition performance is less than 1 second on up-to-date phones. Thereby the corpus of the computation time (the feature computation) is constant, while only a fraction (the actual recognition) scales linear with the number of objects.

These results are comparable to advanced local feature methods (e.g. [Low99]) – although they are carried out directly on the mobile device rather than on a powerful

remote server. This decreases online-times and consequently the cost for network traffic.

Our experiments have shown that simple global color features lead to a maximum recognition rate. The view-dependence of additional structure features have always led to impairment. Furthermore, using only the set of color features maximizes the recognition performance.

However, some limitations exist. Due to varying color responses of the cameras used in different phone types, our method is not compatible. The recognition rate drops significantly if a data set is trained on one phone type and is then used on another type. To implement and evaluate inter-camera color calibration methods [Pol03] belongs to our future work. Individual color calibration functions for each phone type will then allow achieving constant recognition rates.

Our method is not well suited for outdoor applications. Extremely varying lighting situations (especially shadows, highlights, gloaming, afterglow, etc) decrease the recognition rate even more. In addition, the possibility to photograph an outdoor object from very different distances results in a variable amount of background pixels, and also leads to a drop-down of the recognition rate. Fortunately, both problems are reduced to a minimum in a museum environment.

Finally, the recognition rate decreases with an increasing number of objects. If a massive amount of exhibits has to be differentiated with an acceptable rate, our method alone is not practical. In future we want to investigate the combination of our on-devices object recognition method with a grid of local (passive or active) emitters, such as infra-red, Bluetooh, or RFID tags. These emitters can provide a rough spatial awareness to phones within their signal ranges. Only the objects that are located within the signal range of one emitter would need to be recognized. Thus an individual (smaller) data set can be assigned to each emitter which is simply exchanged or interpolated on the phone if a new zone is entered. This ensures a high recognition rate for a large number of objects.

Another possibility for future research is to investigate the capabilities of the high recognition performance. We believe that we can accelerate our method to reach a frame rate of up to 2-3 frames-per-second. This opens interesting possibilities for continuous object recognition.

8. ACKNOWLEDGMENTS

We thank the Senckenberg Museum Frankfurt, the Museum for Pre- and Early History Weimar, Nokia and CellIQ for their support and cooperation. Special thanks go to Tim Golub and Sebastian Derkau for their help.

9. REFERENCES

[And03] Andreasson, H. and Duckett, T., "Object Recognition by a Mobile Robot using Omni-directional Vision", Proceedings of the Eighth Scandinavian Conference on Artificial Intelligence, Norway, 2003.

[Aok00] Aoki, P.M. and Woodruff A., "Improving Electronic Guidebook Interfaces Using a Task-Oriented Design Approach", Proc. 3rd ACM Conf. on Designing Interactive Systems, New York, NY, pp. 319-325, 2000.

[Ass03] Assad, M., Carmichael, D. J., Cutting, D., and Hudson, A., "AR phone: Accessible Augmented Reality in the Intelligent Environment", Australiasian Computer Human Interaction Conference OZCHI'03, pp. 232-235, 2003.

[Bom03] Bombara, M., Cali, D., and Santoro, C., "KORE: A Multi-Agent System to Assist Museum Visitors", Joint Workshop "From Objects to Agents": Intelligent Systems and Pervasive Computing, pp. 175-178, 2003.

[Bon04] University of Bonn, "Das Fotohandy als Fremdenführer", retrieved from WWW, http://www.ipb.uni-bonn.de/FotoNav/ and http://www.geoscience-nline.de/index.php?cmd=wissen_details&id=1713&datum=2004-10-12, 2004.

[Cla05] C-Lab, "Kickreal", retrieved from WWW, www.c-lab.de or www.kickreal.de, 2005.

[Cor00] Coors, V., Huch, T., and Kretschmar, U., "Matching Buildings: Pose Estimation in an Urban Enviroment", Proc. IEEE and ACM International Symposium on Augmented Reality, Munich, Germany, pp. 89-92, 2000.

[Fei97] Feiner, S., MacIntyre, B., Höllerer, T., and Webster, A., "A Touring Machine: Prototyping 3D Mobile Augmented Reality Systems for Exploring the Urban Environment", Proceedings 1st International Symposium on Wearable Computing, Cambridge, MA, pp. 74-81, 1997.

[Fri04a] Fritz, G., Seifert, C., Paletta, L., and Bischof H., "Rapid Object Recognition from Discriminative

Regions of Interest", Proc. National Conference on Artificial Intelligence, AAAI, San Rose, CA, 2004.

[Fri04b] Fritz, G., Seifert, C., Luley, P., Paletta, L., and Almer, A., "Mobile Vision for Ambient Learning in Urban Enviroments", Proc. International Conference on Mobile Learning, MLEARN 2004, Lake Bracciano, Rome, July 2004.

[Har88] Harris, C. and Stephens, M., "A combined corner and edge detector", In Alvey Vision Conference, pages 147–151, 1988.

[Hel04] Helmer, S. and Lowe, D. G., "Object Class Recognition with Many Local Features", Workshop on Generative Model Based Vision 2004 (GMBV), Washington, D.C., July 2004.

[Iqb02] Iqbal, Q. and Aggarwal, J. K., "CIRES: A System for content-based retrieval in digital image libraries", Seventh International Conference on Control, Automation, Robotics and Vision (ICARCV'02), 2002.

[Leh00] Lehmann, T.M., Wein, B.B., Dahmen, J., Bredno, J., Vogelsang, F., and Kohnen, M., "Content-Based Image Retrieval in Medical Applications": A Novel Multi-Step Approach, Proceedings SPIE'00, vol. 3972, pp. 312-320, 2002.

[Low99] Lowe, D.G., "Object Recognition from Local Scale-Invariant Features", International Conference on Computer Vision, Greece, pp. 1150-1157, 1999.

[Low04] Lowe, D.G., "Distinctive image features from scale-invariant keypoints", International Journal on Computer Vision, vol. 60, pp. 91-110, 2004.

[Mac04] Macedonia, M., "Small is Beautiful", IEEE Computer, vol. 37, no. 12, pp.122-123, 2004.

[Mik02] Mikolajczyk, K and Schmid, C., "An affine invariant interest point detector", European Conference on Computer Vision, Copenhagen, pp. 128-142, 2002.

[MIT04] MIT's Technology Review, "Markets and Trends", p. 16, February 2004.

[Mnh04] Museum of Natural History, "Museum Puts Tags on Stuffed Birds", retrieved from WWW, http://rfidjournal.com/article/articleview/1110/1/1/, 2004.

[Moe04] Moehring, M., Lessig, C., and Bimber, O., "Optical Tracking and Video See-Through AR on Consumer Cell Phones", In proceedings of Workshop on Virtual and Augmented Reality of the GI-Fachgruppe AR/VR, pp. 193-204, 2004.

[Pol03] Porikli, F.M., "Inter-Camera Color Calibration by Cross-Correlation Model Function", IEEE International Conference on Image Processing (ICIP), Vol. 2, pp. 133-136, September 2003.

[Sch96] Schiele, B. and Crowley, J.L., "Object recognition using multidimensional receptive field histograms", Fourth European Conference on Computer Vision, Cambridge, UK, pp. 610–619, 1996.

[Sch97] Schmid C. and Mohr, R., "Local grayvalue invariants for image retrieval", IEEE Transactions on Pattern Analysis and Machine Intelligence, 19(5):530–534, 1997.

[Sei04] Seifert, C., Paletta, L., Jeitler, A., Hoedl, E., Andreu, J.P., Luley, P., and Almer A., "Visual Object Detection for Mobile Road Sign Inventory", In: Brewster S. and Dunlop M., (Eds.): Mobile HCI, LNCS 3160, pp. 491-495, Springer Verlag Berlin, 2004.

[Sem04] Semacode Cooperation, "Semacode", retrieved from WWW, http://www.semacode.org, 2004.

[Swa91] Swain, M. and Ballard, D., "Color indexing", International Journal of Computer Vision, vol. 7, no. 1 , pp. 11–32, 1991.

[Wag03] Wagner, D. and Schmalstieg, D., "First steps towards handheld augmented reality", In proceedings of International Conference on Wearable Computers, pp. 127-136, 2003.