Subobject Detection through Spatial Relationships on Mobile Phones

Benjamin Brombach
Bauhaus-University Weimar
Weimar, Germany
brombach@uni-weimar.de

Erich Bruns
Bauhaus-University Weimar
Weimar, Germany
bruns@uni-weimar.de

Oliver Bimber
Bauhaus-University Weimar
Weimar, Germany
bimber@uni-weimar.de

ABSTRACT

We present a novel image classification technique for detecting multiple objects (called subobjects) in a single image. In addition to image classifiers, we apply spatial relationships among the subobjects to verify and to predict locations of detected and undetected subobjects, respectively. By continuously refining the spatial relationships throughout the detection process, even locations of completely occluded exhibits can be determined. This approach is applied in the context of *PhoneGuide*, an adaptive museum guidance system for camera-equipped mobile phones.
Laboratory tests as well as a field experiment reveal recognition rates and performance improvements when compared to related approaches.

Author Keywords

Mobile Computing, Museum Guidance Application, Spatial Relationships, Subobject detection

ACM Classification Keywords

I.4 Image processing and computer vision: Applications; H.1.2 Models and principles: User/Machine Systems—*Human factors*; C.3 Special-Purpose and Application-Based Systems: Microprocessor/Microcomputer Applications

INTRODUCTION AND MOTIVATION

Many museums are lacking in engaging and intuitive forms of information presentation. In general, text labels are placed close to exhibited objects for displaying related content, while audio guides can provide auditive complements. Modern museum guidance systems will enable further types of multimedia presentations in addition to text and audio, such as images, videos, 2D and 3D graphics. They will also make the identification of individual objects more intuitive. Instead of keying reference numbers, as it is the case for conventional audio guides, exhibits can be automatically detected through image classification techniques.

We developed an adaptive museum guidance system called *PhoneGuide* [11, 7, 4, 6, 5]. It utilizes the visitors' personal mobile phones for information retrieval and serves as basis for our subobject detection approach presented in this paper. The front-end application of PhoneGuide is executed on the camera-equipped mobile devices of the visitors, which allows identifying individual exhibits by simply taking a single photo of them. Image classification techniques are carried out locally on the phone that result in a probability-sorted objects list which is presented on the screen [4]. With a minimum number of clicks, the user can select the object of interest from this list to retrieve related multimedia information. No online server connection is required – neither for classification nor for retrieving the multimedia content, since all classification steps are executed directly on the phone and the entire data is kept on the device. This makes PhoneGuide scalable: Waiting times for classification results are independent of the number of simultaneous users and remain constant. No transmission costs for communication services are necessary.
So far, we combined different techniques, such as image classification with global features [11], pervasive tracking [7], dynamic classification adaptation [4, 6], and ad-hoc net-

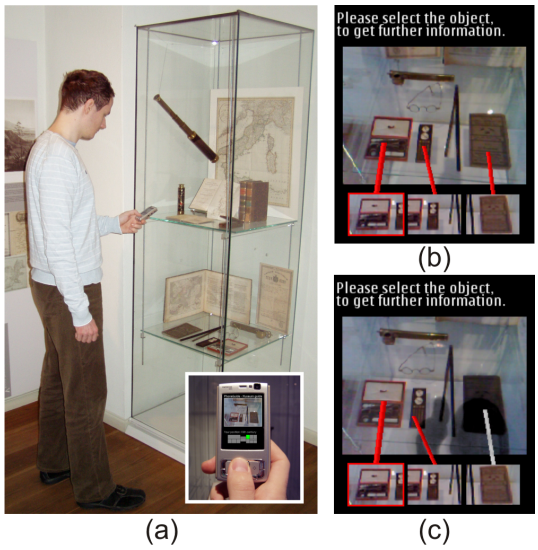

Figure 1. A visitor is taking a photo of a subobject group with his mobile phone (a). Three of the correctly detected subobjects are labeled (b). They could be identified through image classification in this case. If subobjects cannot be detected through image classification, such as in (c) where a shadow is cast onto the exhibit on the right-hand side, the known spatial relationships among the subobjects still allow a correct identification.

Submitted for review to IUI 2009.

work communication [5] for achieving recognition rates in the order of 82%-92% under realistic conditions (i.e., more than one hundred objects, in real museums, with real visitors). However, up to now, PhoneGuide is only able to detect single objects – by convention, the one that is centered in the camera image. In many cases, multiple objects are placed in showcases or behind other barriers to protect them against environmental influences and human curiosity. Thus, they are captured together in a single image.

In this paper, we present a new classification technique that is able to identify and to label all exhibits (called subobjects) that appear in one image. One approach for achieving this could be to apply sophisticated object recognition techniques based on local image features, such as SIFT [15]. This, however, would significantly increase the overall classification time and the amount of classification data required on each device compared to our approach. In addition, the complexity of such techniques scales with the number of objects to be identified.

To avoid such data overheads and to ensure scalability, we propose a new subobject detection technique that combines image classification based on global color features, artificial neural networks and spatial image relationships. Our method follows two basic steps: First, the global context of the captured photograph is identified via image classification (i.e., the regular object recognition technique based on global image features, as currently implemented for PhoneGuide [4]). With this context information, the context-related subobjects are detected in the image with a combination of image classification and spatial relationships in the second step. The spatial relationships become more and more reliable the more subobjects have been found. On the other hand, reliable spatial relationships will restrict the search regions for image classification. Thus, the entire classification becomes the more robust and faster, the more subobjects have been detected and their spatial relationships have been derived. Even partially or completely occluded subobjects (e.g., occluded by shadows or other exhibits) for which image classifiers fail, can be detected with our approach. Finally, all detected subobjects are labeled as shown in figure 1, and the user can select the object of interest for retrieving corresponding multimedia information.

The remaining sections of this paper will describe the different classification techniques in more detail. We will show that the recognition of subobjects using spatial relationships will be up to 68% faster than related approaches without spatial relationships. Results of a field experiment in a local museum will illustrate that unexperienced users reach an average recognition rate for subobjects of 85.6% under realistic conditions.

RELATED WORK

We divide the related work into two main categories: museum guidance systems that are similar to PhoneGuide and object detection approaches that are enhanced through spatial relationships.

Museum Guidance Systems

Fritz et al. [12] introduced a city guide for mobile phones: Datasets including photographs of buildings or monuments and the respective GPS information are captured by tourists and transferred to a remote server via UMTS or GPRS. On the server, the images are compared with a database of known sights via SIFT classification [15]. Finally, the corresponding multimedia data is sent back to the user's phone after the objects have been classified. Hare et al. [13] developed a museum guide for pocket PCs. Photographed images of paintings are transferred to a remote server to compute SIFT features. For classification, however, they apply an adapted text retrieval technique. Nonetheless, the recognition is comparable to that of [12].

Bay et al. [2] introduced a museum guide based on a tablet PC. In contrast to the previous two approaches, the identification is performed directly on the device, and no server communication is established. An enhancement of SIFT, called SURF [3], is applied for classification. In their previous work [1], they distributed Bluetooth emitters to determine the users' locations and consequently narrow the set of possible results. Takacs et al. [19] implemented a performance-improved version of SURF on today's mobile phones for outdoor Augmented Reality applications. To remove outliers of feature pairings, they perform a geometric consistency check based on an affine model.

Most of these approaches allow detecting multiple objects in one image. However, they rely exclusively on local image classification techniques or perform only basic transformation models [19] to verify detected image feature pairs. Instead, we take into account precise spatial relationships among the objects to narrow search areas as well as to verify and adapt results of the image classification during the recognition process. In addition, PhoneGuide supports a temporal adaptation to dynamic environmental changes and user behavior. It improves the recognition rate over time and adapts to preferred user locations [4].

Object Detection Enhanced by Spatial Relationships

Spatial relationships describe specific geometric dependencies between objects. They are applied in many different areas, such as geographic information systems or content-based image retrieval. Yet, their descriptions and definitions vary dependent on the application. For instance, topological relations [9] distinguish the relationships between two objects by analyzing the intersections of their boundaries and interiors (e.g. occluded, partly occluded, or disjunct). Directional relations [17], as another example, are described by directional attributes like north, west, south-east, etc.

Spatial relationships, however, are not only applied to separate individual objects but also to describe different parts within a single object. Pham et al. [18] introduced a detector that consists of several spatially distributed "part detectors" that are based on template matching. The spatial relations between the part detectors are defined by parameters of a Gaussian distribution which are extracted from the part detectors' locations. The object detection itself is carried out by maximizing a function based on the output of the part detectors and their locations. Such a detector configuration is able to achieve a higher recognition rate than a single fixed template based detector due to higher flexibility with respect to object distortion.

Spatial relationships are also utilized to generate a spatial

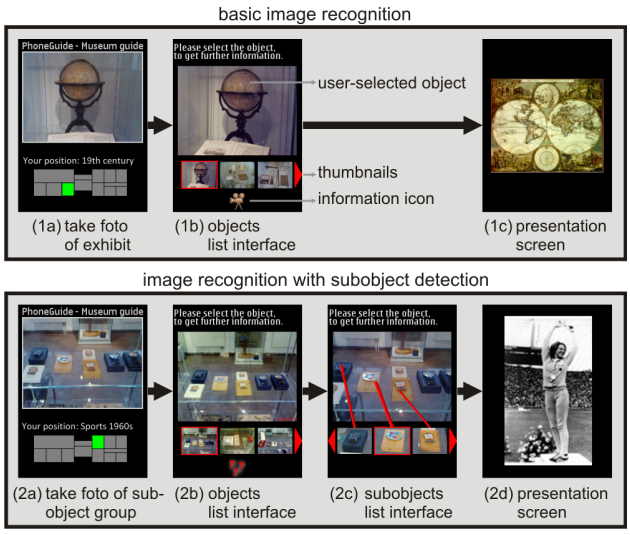

Figure 2. Flow chart of the user interface for single-object recognition and scene recognition with consecutive subobject classification. After the user has taken a photograph of an exhibit (1a), an image classification is carried out and the result is presented as a probability-sorted objects list (1b). The correct object can be selected with a minimum number of clicks for receiving multimedia information (1c). If a group of subobjects was captured rather than a single object (2a), the user has to acknowledge the scene classification first (2b), before a consecutive subobject classification is carried out. The detected exhibits are labeled in the photograph (2c). Finally, the user can select the desired subobject and the corresponding multimedia content is presented (2d).

orientation graph [10]. One node of a graph represents either a part of an object or a single object within a group of objects. The object detection is then realized by performing different graph matching algorithms. In [22], face recognition is carried out by elastic bunch graph matching. A face is defined by sets of wavelet components with different orientations and scales called ”jets”. They are connected with edges holding a distance and an angle. The initial location of the faces must be known. In [16], spatial relationships verify the classification of regions (e.g. sky, tree, street) after an image segmentation. In a post processing step, the consistency of all classified regions is checked and misclassifications (e.g. street located above the sky) are corrected. The spatial relationships are described by angle histograms, resulting from the slope of all possible point pairs of two regions.

All of these approaches utilize the spatial relationships in a post processing step only. Thus, the object locations have to be known before the spatial relationships can be applied. In our approach, the spatial relationships do support image classifiers during the actual classification process and predict subobjects' locations. This leads to a faster subobject detection and reduces misclassifications from the beginning. The classification becomes more robust, the more spatial relationships have been found.

OFFLINE REGISTRATION, TRAINING, AND EXTRACTION OF SPATIAL RELATIONSHIPS

As mentioned earlier, the classification process is separated into two steps (cf. figure 2): In the first step (1a, 2a), a scene, containing one or multiple exhibits, is photographed and identified as explained in [4]. It identifies the scene (and therefore provides the global context information) rather than individual subobjects in the image. Afterwards, a probability-sorted objects list is displayed (1b, 2b). It contains all possible candidates, beginning with the most likely candidate on the left-hand side. The user can now select the correct scene context with a minimum number of clicks (only one, if the scene has been classified correctly). Browsing through the list does not only show thumbnails but also icons indicating what kind of information is available. If, for instance, the image contains only one single object, these icons indicate the different types of multimedia content that are available (e.g., audio, video, text, images), which are played back after selecting the corresponding list entry (1c). Note, that the same technique was used in previous versions of PhoneGuide to detect objects which are centered in the image by definition. The information whether one or multiple objects are present in a captured photograph can simply be tagged to the classification result (i.e., together with the information about the recognized object or scene).

If the information icon indicates that the photographed scene contains multiple exhibits (2b), a consecutive classification step takes place that identifies all subobjects. The result is displayed in a subobjects list that labels the different exhibits (2c). After a final selection of the object of interest, the subobject's individual multimedia content is presented (2d).

The details on the individual classification steps will be described below. All classifiers (i.e., for scene context and for subobjects) are based on global color features and 3-layer artificial neural networks, as explained in [4]. For an initial training of the neural networks, videos are recorded for all exhibited scenes. The videos show the scene from different perspectives, orientations and scales. Keyframes are extracted from each video, clustered and features are computed for representative keyframes. These features are used for an initial training of the neural networks on a server during an one-time preprocessing step. The trained neural networks are then applied on the phones for the scene classification. After the initial training, the parameters of the neural networks can be updated through adaptation techniques – either when visitors enter or leave the museum [4, 6] or during runtime via ad-hoc phone-to-phone networks [5]. Describing details of these techniques is out of the scope of this paper. The interested reader is referred to the individual previous publications.

For supporting the identification of subobjects during the second classification step, however, each subobject has to be considered during the initial training phase. We achieve this by identifying the bounding box of each subobject manually in the first frame of the recorded training videos of each scene, and track them via a kernel-based mean shift algorithm automatically through the entire video sequence. For the bounding boxes of each subobject in each video frame, we compute the same global color features as described in [4] to train subobject-individual neural networks. In addition to this, the spatial relationships among the tracked subobjects throughout each scene video are computed, recorded and stored automatically. These two components (image classifiers and spatial relationships) are the basis of our sub-

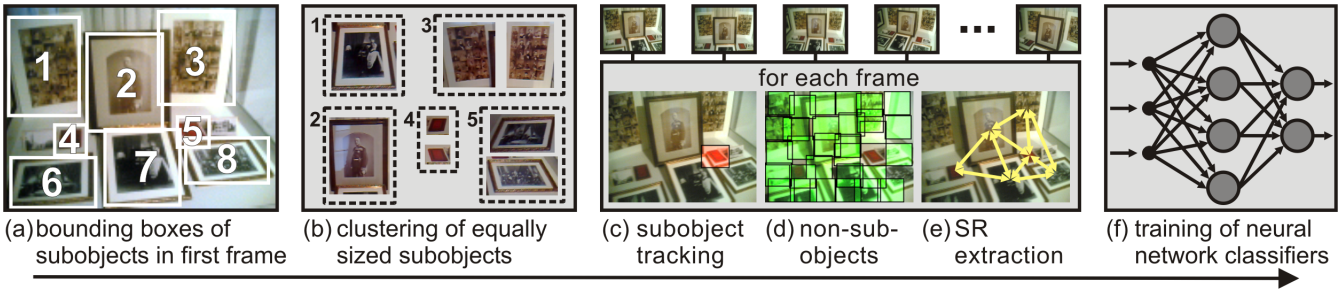

Figure 3. Flow chart of the offline preprocessing: After a video of a scene with subobjects was recorded, a bounding box for each exhibit is manually defined in the first frame (a). They are clustered automatically according to their size (b). For each frame in the video, all subobjects are tracked (c), subimages of subobjects and non-subobjects (d) are stored, and the spatial relationships are extracted (e). Finally, global color features of all subimages are computed to train the subobject-individual classifiers (f).

object detection algorithm. They are initially computed on the server as part of the one-time preprocessing step. Once computed, they are used on phones for subobject classification during runtime. The following sections will explain how these two components are computed in more detail.

Registration and Tracking of Subobjects

As indicated above, the bounding boxes of all subobjects are manually defined in the first frame of a scene video (cf. figure 3a). They have to be automatically tracked throughout the subsequent video frames to compute global features of the subimages framed by the axis-aligned bounding boxes and for deriving the spatial relationships among the detected subojects.

We evaluated three different tracking techniques for accomplishing this: Template matching with fast normalized cross-correlation [14], tracking based on SIFT features [15] and kernel based mean shift tracking [8]. We found that mean shift tracking is the most robust technique for our applied low-resolution video recordings (160x120 pixels). Local feature extraction techniques, such as SIFT, would perform similarly if the video resolution would be increased.

The tagged subobjects are clustered based on the size of their bounding boxes via a simple agglomerative clustering technique (3b). This is necessary to ensure that the correct subimage sizes (search masks) are selected for feature calculation on the phones during runtime. The subobjects are tracked throughout all frames via mean shift tracking (3c). The 2D pixel locations of each subobject's center on the image plane are then used for deriving the spatial relationships to other subobjects within each frame (3e). In addition to the subimages that actually contain exhibits, additional subimages of the same size are also automatically collected in each frame (3d). We refer to them as *non-subobject* subimages. They are used later as negative samples for training the neural networks.

Generation of Subobject Classifiers

After tracking all subobjects throughout the training videos, a certain number of subimages for each subobject is stored and available for training (figure 3f). The number of subimages can vary among the subobjects. Only subimages that contain a single subobject which is not occluded by others as well as subimages that are within the frame boundaries are considered. Global color features (three 10-bin color histograms, mean and variance in color channels [4]) are extracted from each subimage and combined to a feature vector that is applied for training two different 3-layer neural network classifiers: A general classifier C_{all} is trained by using the computed feature vectors of all detected subobjects. Consequently, for each subobject group, one C_{all} classifier is generated whose number of output neurons equals the number of exhibits. This classifier can identify which subobject of the subobject group has the highest probability of being located in a specified region.

The second type of classifiers C_{spec} are specialized to detect individual exhibits (i.e., one C_{spec} classifier per subobject). Thus, only one output neuron is necessary in this case. It is trained by applying the feature vectors of one particular subobject in combination with the features extracted from the non-subobject subimages which serve as negative training samples.

Applying the results of both classifiers ensures a more robust classification and improves the recognition results [21](cp. following chapter).

Extraction of Spatial Relationships

If the detection of subobjects would be exclusively performed through image classification, the entire image has to be scanned and tested against different subobject classifiers. This is both computational exhausting and unreliable. Spatial relationships describe how the subobjects are arranged in relation to one another (figure 3e). This has preliminary two advantages for the online classification during runtime: First, the spatial relationships localize specific search areas for undetected subobjects. Consequently, if at least one subobject is detected, the locations of the remaining subobjects can be approximated and the searching time decreases accordingly. The more exhibits are detected over time, the more precise the prediction of the remaining subobjects' locations becomes. The second advantage is that the spatial relationships serve as an additional classifier. If, for instance, classifiers C_{all} and C_{spec} detect a subobject at an impossible location (this can be derived from the spatial relationships), the result is discarded and a new search is initiated.

We use two geometric parameters for describing the spatial relationships among tracked subobjects: *distances* and *angles*. The distances describe the normalized range be-

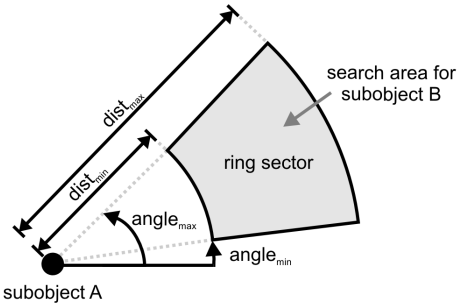

Figure 4. Predicted search area of an undetected subobject (B) relative to a detected exhibit (A). The corresponding ring sector is defined by the minimum and maximum distances and angles that were extracted during subobject tracking in the offline preprocessing.

tween two subobjects within the image. They are mutable against scaling (i.e., the distance of a visitor to the exhibits) but invariant against rotation (i.e., orientation of the mobile phone when a photo is taken). The angles between subobjects are defined by the slope of a straight line that connects two of them relative to the image's horizontal edge. They are rotation-variant, but invariant to scaling. Consequently, combining both parameters leads to a robust and precise geometric mapping of the spatial relationships – in contrast to e.g., a topological mapping.

Angle and distance parameters are usually different for each frame. Therefore a 4-tuple ($dist_{min}$, $dist_{max}$, $angle_{min}$, $angle_{max}$) of minimum and maximum distance and angle is defined by the individual distances and angles collected from each frame for each subobject pair. This 4-tuple defines a ring sector (cf. figure 4) that describes the location of one exhibit relative to another one. Each subobject is associated with all other exhibits by these 4-tuples. This leads to a total number of $\binom{N}{2}$ spatial relationship 4-tuples for N subobjects of one subobject group.

In summary, the result of the preprocessing as part of the initial one-time training procedure are the classifiers C_{all} (one per scene) and C_{spec} (one per subobject), the spatial relationships ($\binom{N}{2}$ 4-tuples for N subobjects per subobject group) and the clustered subobject sizes per subobject group. This data is transferred to the mobile phones and will be used for online classification during runtime.

ONLINE SUBOBJECT DETECTION

The online subobject detection algorithm can be separated into three main steps for identifying N subobjects: In the first step, it searches for M, $M < N$ subobjects that serve as anchors for determining reliably the current rotation and scale relationships among them. Then, the remaining $N-M$ subobjects can be detected faster while continuously refining the spatial relationships. Finally, subobjects that were not detected but are presumably in the image are located by prediction through the geometric dependencies. The following sections will explain this in more detail.

Detection of Anchor Subobjects

Since the correct scene context is given through the first classification step and the visitors' feedback, the corresponding classifiers (C_{all}, C_{spec}), spatial relationships (angles, distances) and cluster information (sizes of search masks) can be derived and selected accordingly.

For finding the first anchor subobject, no prior knowledge about geometric relationships or the actual number of subobjects in the image is available due to the unknown perspective of the user's location. Therefore, the algorithm starts searching for subobjects from the center of the image, since we assume that it is likely that visitors will center one of the subobjects to a certain degree. A search mask (cf. figure 5a) is moved spirally around the center with a step size that depends on the search mask's size. Empirically, the step size is chosen such that at least 80% of the previous search region is superimposed by the current one. In each step, the search mask's size is adjusted to all the clustered subobject sizes that were generated during the offline training. For each pixel region that is covered by a search mask, the global color features are computed from a precomputed integral image [20]. Integral images speed-up the computation of image features within subimage regions. These features serve as input for the classifiers to identify the first anchor subobject. It is detected if the following conditions are met (cf. figure 5b): (1) the maximum excitation of C_{all} is above a predefined threshold t_c, (2) the size of the identified subobject equals to the size of the current search mask, and (3) the specific classifier C_{spec} of the candidate confirms the result of the general classifier C_{all}. The final location of the detected subobject is refined afterwards (cf. figure 5c) by moving the search mask in a small step size within a pre-defined area around the initial position, and selecting the best match (i.e., the position with the highest classification excitation).

This first anchor subobject (figure 5d) provides basic information about the position of the remaining anchor subobjects. The region where the second anchor subobject is located is defined by the spatial relationships that were extracted during the offline preprocessing (figure 5e). The starting point for searching the second anchor subobject is the center of the derived ring sector.

After detecting the second and third subobject as explained above, reliable information about the scale and rotation of the phone and consequently of the captured image can be derived. This is important since the spatial relationships stored on the phone are absolute values and are either variant to scale or rotation. In addition, users align phones differently, which changes the geometric dependencies among different orientations and distances. Thus, correction factors have to be computed for both parameters (distance, angle) during the recognition process that compensate for different phone alignments: The required distance scaling factor is derived from the average ratio of the currently computed distance and expected (from the offline preprocessing) distance between all possible detected subobject pairs. The rotation correction angle is derived from the average quotient of the differences between the detected and expected angle as described in [10]. Newer phones have built-in accelerometers which can be used to determine the relative pose of the mobile phones. Such sensors can be applied to compute the rotation correction angle before the subobject detection starts. However, we also have to consider false positives (i.e., wrongly detected subobjects). False positives influence

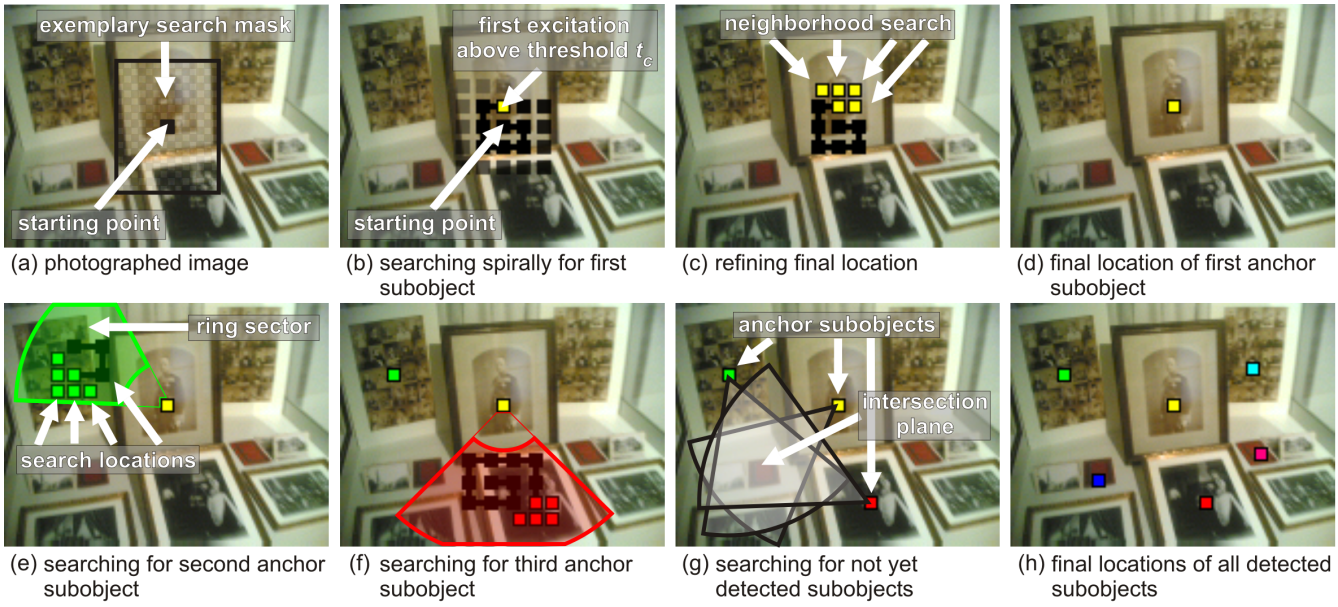

Figure 5. After an image was captured, the subobject detection searches for the first anchor subobject by varying the mask size (a). The search masks are spirally shifted around the center of the image until one subobject is identified through image classification (b). A neighborhood search (c) is performed next to refine the location of the anchor subobject (c) until the final position is found (d). Spatial relationships can be applied to find other exhibits (e-f). If enough anchor subobjects are detected, the spatial relationships span intersection planes which define reliable search areas of the remaining subobjects (g). This is repeated until all exhibits are detected (h).

the successive development of spatial relationships and lead therefore to wrong search areas and to misclassifications of subobjects. To overcome this, we apply the following function for expressing the classification quality of two related subobjects. It weights and combines the results of the image classification and of the spatial relationships:

$$SIM_{cda} = \omega_1 \cdot P_c + \omega_2 \cdot SIM_d + \omega_2 \cdot SIM_a \quad (1)$$

$$P_c = P(A) \cdot P(B) \quad (2)$$

$$SIM_d = \left[1 - \frac{|D_{AB} - d_{AB}|}{\sqrt{W^2 + H^2}}\right] \quad (3)$$

$$SIM_a = \left[1 - \frac{|\alpha_{AB} - \beta_{AB}|}{180}\right] \quad (4)$$

Equation 1 denotes the probability that two subobjects A and B are detected correctly. This can be derived from three components. The first component, P_c (equation 2), comprises the probability that both subobjects are detected correctly. It is the product of the output probabilities of the C_{all} classifier for A and B. The second component, SIM_d (equation 3), denotes the normalized similarity (W = width, H = height of image) of the currently computed distance d_{AB} between A and B, and the expected distance D_{AB} that was pre-computed offline. The last component, SIM_a (equation 4), defines the normalized similarity of the currently computed angle β_{AB} between A and B, and the expected (pre-computed) angle α_{AB}. All three components are weighted by ω_1, ω_2 and ω_3 (with $\omega_1 + \omega_2 + \omega_3 = 1$). The weights are empirical and define the classification reliability of the three components. We chose $\omega_1 = 0.2$, $\omega_2 = 0.4$ and $\omega_3 = 0.4$.
If new subobjects are found, the quality function SIM_{cda} is applied for each combination of detected subobject pairs. If the average quality is above a predefined threshold t_{cda}, the search for anchor subobjects is completed. In this case, enough exhibits are detected. We figured out that a minimum number of three anchor subobjects is necessary for reliably determining the scale and rotation of the phone relative to the real exhibits. From here, a faster detection technique that is mainly based on the spatial relationships can be applied to find the remaining subobjects. This is explained in the following section.
If SIM_{cda} of one subobject to multiple other subobjects is low while in comparison the quality among the others is high, then this indicates that this particular subobject was probably misclassified and its detection is discarded.

Detection of Remaining Subobjects

If a sufficient number of anchor subobjects are found, the remaining subobjects can be reliably detected by applying the spatial relationships. For each remaining subobject that was not yet detected, the spatial relationships (adjusted by the scaling factor and the rotation correction angle, as explained above) define different ring sectors (cf. figure 5g). The intersection planes that are spanned by the ring sectors of the identified anchor subobjects are the final search areas in which the remaining exhibits are located. In practice, these intersection planes are not computed since the computational costs would be too expensive. Instead, the search locations (cf. figure 5e) are tested against each ring sector individually. For detecting the remaining subobjects, only C_{spec} of the currently demanded subobject is applied. Remember, that we know which subobject is located in this search region based on the spatial relationships. Searching the exhibit within the constrained region is done as explained

above (i.e., spirally shifted search mask starting at the center of the search region, refining the initially found location through searches with smaller step sizes afterwards).
Consequently, finding the remaining subobjects is processed much faster than finding anchor subobjects, since the starting points in the search areas are more precise and reliable, and only one classifier is applied. Although the quality function is only used for the anchor subobjects, the scale factor and rotation correction angle are recomputed after each new detected exhibit for continuously refining the search areas.
However, if the output of C_{spec} for all tested locations is below the threshold t_c, no subobject will be detected, even though the spatial relationships might have indicated one. In these cases, the classifier is either not sufficiently trained to recognize the subobject correctly or the subobject is occluded by another one. Therefore, the locations of the missing subobjects are predicted exclusively from the spatial relationships. Its location is defined to be the center of gravity of the corresponding intersection planes. An example for such a case is illustrated in figure 1c: Although the user casts a shadow on the book which leads to an image-based misclassification, the exhibit is still detected from the spatial relationships.
Finding subobjects exclusively through their spatial relationships opens the opportunity to locate even exhibits that are always completely occluded by other objects, or ones that are so small that image classifiers can not detect them reliably. Such subobjects are tagged in the training video to extract the corresponding spatial relationships without training C_{spec} classifiers for them and without considering them for the C_{all} classifier.
After all subobjects have been detected (cf. figure 5h), the labeled subobjects list is presented to the user, as illustrated in figures 1b,c.

EVALUATION

We evaluated our approach with respect to two main questions: How high is its classification rate and performance compared to related approaches that do not apply spatial relationships? How well does it perform in the course of a field experiment under realistic conditions (i.e., in a museum, with unexperienced visitors)?
For the performance analysis, we have compared the subobject detection technique with a brute-force method that scans the whole image for subobjects, as well as with a brute-force method with early stopping (ES) that cancels the search if all subobjects have been found. This test was carried out in a laboratory with real image data that was captured in advance in the City Museum of Weimar, Germany. The field experiment was performed with 15 subjects in the same museum. For both experiments (laboratory and field test) 12 subobject groups were selected (6 of them are displayed in figure 6). The number of subobjects per group ranged from 3 to 8 subobjects (average: 5.4). Of each group, a video of 90 frames (160x120 pixels) was recorded from different perspectives and distances. Every third frame of each video was used for classification in the laboratory experiment such that in total 720 frames were applied for training and 360 frames were applied for simulating the recognition. The PhoneGuide application is developed in J2ME and the experiments were

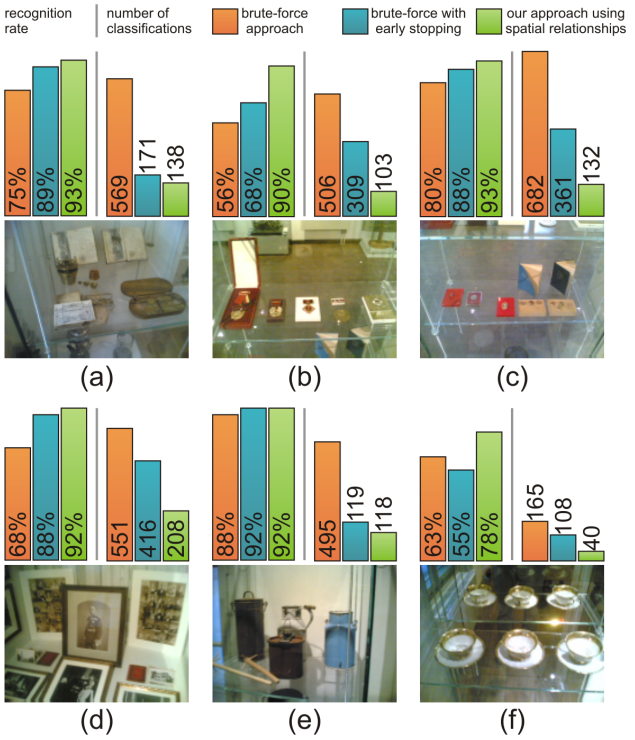

Figure 6. Average recognition rate and number of classifications for brute-force, brute-force with early stopping, and for our approach (6 out of 12 different subobject groups). Thirty images from different perspectives and distances were selected and classified for each group. The graphs show that our approach outperforms related approaches without spatial relationships, both in speed and recognition rate.

carried out on Nokia 6680 (CPU: 220 MHz) and Nokia N95 (330 MHz) mobile phones.

Performance Analysis

In general, a subobject detection that applies spatial relationships should perform faster than approaches that scan the entire image, since only predefined subregions are examined. In addition, they should even improve the overall recognition rate since the spatial relationships support the image classifiers (C_{all} and C_{spec}) by determining the rough location of a subobject. Thus, misclassifications at geometrically impossible locations should be avoided.
To prove that these two hypotheses (i.e., classification speed-up and improved recognition rate) are in fact true, we have compared our approach with a brute-force method that scans the whole image for subobjects: The search mask is spirally moved to each possible location until it has reached each part of the image, beginning from the center. At each location of the search mask, global color features are extracted to perform the classification with the C_{all} and C_{spec} classifiers. Parameters like search mask size and step size are the same as for our approach in order to compare both approaches properly. After the entire image has been scanned, the search areas with the highest sum of output excitations of both classifiers are selected as the final locations for the corresponding subobjects.
The brute-force method with early stopping is carried out in

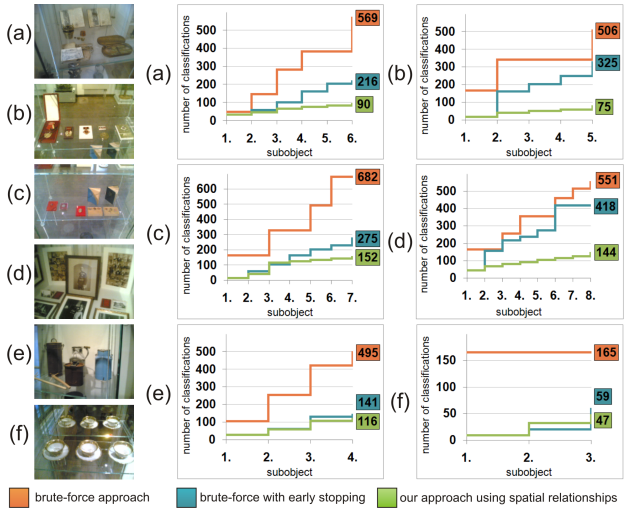

Figure 7. Number of classifications required for brute-force, brute-force with early stopping and for our approach. For each of the 6 subobject groups, one image was selected to determine the number of classifications for each subobject. It indicates that applying spatial relationships requires less classification steps.

a similar way as the prior brute-force method. The only difference is, that it stops searching for a specific subobject, if the output of C_{all} is above the threshold t_c and at the same time the excitation of C_{spec} is above t_c, too. Thus, compared to the brute-force method, the computational effort is reduced.

The recognition results of both methods in comparison to our approach are illustrated in figure 6. Six different subobject groups are displayed with their corresponding average recognition rates for each method. Furthermore, the number of classifications that were required to detect all subobjects are displayed.

For each subobject group, 30 randomly selected images from different perspectives and distances were used to determine the results. These images contained different numbers of subobjects, since they can be outside the images' boundaries or (partially) occluded. The brute-force method reaches an average classification rate of 83.2% (for 12 subobject groups) with 13.4% false positives. The brute-force method with ES achieves a similar average classification rate of 85.7% and 14.1% false positives. Our approach reaches an average classification rate of 94.4% with 3.0% false positives. Thereby, 11.6% of all correctly detected subobjects were found exclusively by applying the spatial relationships for situations in which the image classifier failed. The results prove that the classification rate of our method significantly outperforms brute-force and brute-force ES approaches.

Beside the improved recognition rate, figure 6 illustrates that the recognition process needs less classification steps on average, which correlates to lower classification times. Thus, our approach is much faster than brute-force methods and brute-force ES methods.

To determine the speed-up more precisely, we monitored the number of classification steps relative to the number of detected subobjects, as shown in figure 7. We have selected one image from each subobject group to show how the number of classification steps increases with the number of subobjects for each of the three approaches. For the first subobject group (cf. figure 7a), for instance, the brute-force method needs 49 classification steps to find one subobject, 148 for detecting two subobjects, and so on. Finally 569 classification steps are required. In some cases, the number of classification steps for the brute-force approach and brute-force ES approach does not increase for two consecutive subobjects. The reason for this is that these techniques can detect multiple subobjects within one image scan as long as they are equally sized. Thus, if all subobjects would have the same search mask size, the number of required classification steps is constant to the number of subobjects, as can be seen in figure 7f. However, even in such cases, the brute-force method's number of classification steps is still higher than in our approach.

If the overall computation times (including the necessary geometric computations) of the three approaches are compared rather than the number of classification steps, our approach is 68% faster than the brute-force method and approximately 50% faster than the brute-force ES method.

Field Experiment

Our field experiment was carried out over multiple days and different times of day in the City Museum of Weimar, Germany. Each of the 15 subjects (male: 12, female: 3, average age: 26.2 years) were asked to photograph all 12 subobject groups individually with the Nokia N95 mobile phone. The subobject groups, and consequently the spatial relationships and classifiers were identical to the ones that were applied for the performance analysis. The size of the necessary classification data for 12 subobject groups with 64 subobjects in total was 237 kb.

The recognition rate that was achieved by the subjects under realistic conditions was 85.9% on average (max: 100.0%, min: 52.4%, per subobject group). The recognition performance depended mainly on the visitors' perspectives and on the appearance of the subobjects. If subobjects could be visually separated easily, the classification performance was reliable. Thus, the worst recognition result (52.4%) occurred at a subobject set with three almost identical cups in front of a mirror (cf. figure 6f). The average recognition rate is lower compared to the laboratory results. This is mainly due to the individual behavior of subjects when approaching and photographing the exhibits. An adaptive classification technique, such as the one described in [4], would compensate for this. Combining subobject recognition and adaptive classification belongs to our future work. The time for subobject detection, including integral image computation, ranged between 1.25 seconds and 4.45 seconds, (average: 2.85 seconds), depending on the number of subobjects, the number of clusters and the number of necessary classifications. Since the first classification step (i.e., recognizing the scene context) takes less than 0.5 seconds [6] the computation of the integral image can be performed as part of the first classification step. This increases the classification time of the first recognition, but reduces the duration of the subobject detection in the second classification step by $\sim$0.6 seconds to 2.3 seconds.

We also asked each subject to fill out a questionnaire and rate

different aspect of our system with marks from 1 (worst) to 7 (best). With this, we wanted to receive feedback on the usability of the subobject detection as well as the users' acceptance on the required computation time and achieved classification rate. Basic questions concerning handling (e.g., How easy was it to take a photo?) were already evaluated in a previous field test [4] and led to satisfying results again. Additionally, the subjects were asked how comfortable they felt with the waiting time until the classification results of the first classification step (i.e., context) and of the second classification step (i.e., subobjects) are displayed. The duration of the first step took $\sim$0.95 seconds (including the computation of the integral image) and was voted with 6.5 ($\sigma = 0.5$). The second step needed on average 2.3 seconds and was evaluated with 5.0 ($\sigma = 1.1$). In general, 54% of the subjects would prefer a recognition duration of 2-4 seconds, and 46% would prefer a classification time of below 2 seconds (11% requested a classification time of below 1 second) for each of the two steps. One subject explained that she is not willing to accept long waiting times since she wants to concentrate on the exhibition itself rather than on her mobile phone. Consequently, the shorter the duration of the classification is, the better is the acceptance of such a guidance system. Since the subobject detection takes 2.3 seconds on the applied hardware, it suits the requirements of the majority of our subjects. The subobject detection rate of 85.9% was evaluated with 5.8 ($\sigma = 0.7$). The accuracy of the labels that indicate the exact location of the subobjects on the screen was judged with 5.6 ($\sigma = 0.6$). The readability of the detection result was ranked with 6.1 ($\sigma = 0.6$). This shows that most of the subjects were satisfied with the overall handling, the performance and the visualization of our system.

SUMMARY AND DISCUSSION

In this paper we have presented a new technique for the detection of subobjects in a single image. Our method combines light-weight image classification using global image features, artificial neural networks and spatial relationships. This has three advantages compared to related approaches that apply a brute-force search (with or without early stopping): First, the subobject detection is more reliable since the spatial relationships can be used to validate the locations of detected exhibits. Second, they speed-up the detection process by predicting the locations of undetected subobjects. This is continuously refined, the more subobjects are detected. Third, entirely occluded or similar subobjects can be located through spatial relationships, even if an image classification fails.

A field experiment revealed that the classification performance of 85.9%, the visualization of the results, as well as the recognition time of 2.3 seconds are acceptable for practical applications in a museum.

One drawback of our approach is the sensitivity to scaling (i.e., to the distance of the visitors to subobjects when taking a photograph). However, most people approach exhibits in a similar way and capture images from similar perspectives and distance, as found in [4].

Another problem arises if a large number of very small subobjects have to be detected simultaneously. The global features that are computed from their subimages would not be very representative, and their high variance would lead to an insufficient training and classification. Increasing the image resolution would solve this problem on the cost of classification performance. However, the continuously increasing processor speed of mobile phones will compensate this in future.

ACKNOWLEDGMENTS

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. H. Bay, B. Fasel, and L. V. Gool. Interactive museum guide. In *Proc. Ubiquitous Computing, Workshop on Smart Environments and Their Applications to Cultural Heritage*, 2005.

2. H. Bay, B. Fasel, and L. V. Gool. Interactive museum guide: Fast and robust recognition of museum objects. In *Workshop on Mobile Vision*, 2006.

3. H. Bay, T. Tuytelaars, , and L. V. Gool. Surf: Speeded up robust features. In *Proc. Conference on Computer Vision*, 2006.

4. E. Bruns and O. Bimber. Adaptive training of video sets for image recognition on mobile phones. *Journal of Personal and Ubiquitous Computing*, 2008.

5. E. Bruns and O. Bimber. Phone-to-phone communication for adaptive image classification. *Advances in Mobile Computing and Multimedia*, 2008.

6. E. Bruns, B. Brombach, and O. Bimber. Mobile phone enabled museum guidance with adaptive classification. *Journal of Computer Graphics and Applications*, 28(4):98–102, 2008.

7. E. Bruns, B. Brombach, T. Zeidler, and O. Bimber. Enabling mobile phones to support large-scale museum guidance. *Journal of MultiMedia*, 14(2):16–25, 2007.

8. D. Comaniciu and P. Meer. Mean shift: A robust approach toward feature space analysis. *Transactions Pattern Analysis Machine Intelligence*, 24(5):603–619, 2002.

9. M. Egenhofer and R. Franzosa. Point-set topological spatial relations. *Journal of Geographical Information Systems*, 5 (2):161–174, 1991.

10. E. A. El-Kwae and M. R. Kabuka. A robust framework for content-based retrieval by spatial similarity in image databases. *Transactions on Information Systems*, 17(2):174–198, 1999.

11. P. Föckler, T. Zeidler, B. Brombach, E. Bruns, and O. Bimber. Phoneguide: museum guidance supported by on-device object recognition on mobile phones. In *Proc. Mobile and ubiquitous multimedia*, pages 3–10, 2005.

12. G. Fritz, C. Seifert, and L. Paletta. A mobile vision system for urban detection with informative local descriptors. In *Proc. Computer Vision Systems*, 2006.

13. J. S. Hare and P. H. Lewis. Content-based image retrieval using a mobile device as a novel interface. In *Proc. Storage and Retrieval Methods and Applications for Multimedia*, pages 64–75, 2004.

14. J. Lewis. Fast normalized cross-correlation. In *Vision Interface*, pages 120–123, 1995.

15. D. G. Lowe. Distinctive image features from scale-invariant keypoints. *Journal of Computer Vision*, 60(2):91–110, 2004.

16. C. Millet, I. Bloch, P. Hede, and P.-A. Moellic. Using relative spatial relationships to improve individual region recognition. In *Proc. Integration of Knowledge, Semantic and Digital Media Technologies*, pages 119–126, 2005.

17. D. Papadias and Y. Theodoridis. Spatial relations, minimum bounding rectangles, and spatial data structures. *Journal of Geographical Information Science*, 11(2):111–138, 1997.

18. T. V. Pham and A. W. M. Smeulders. Learning spatial relations in object recognition. *Pattern Recognition Letters*, 27(14):1673–1684, 2006.

19. G. Takacs, V. Chandrasekhar, N. Gelfand, Y. Xiong, W.-C. Chen, T. Bismpigiannis, R. Grzeszczuk, K. Pulli, and B. Girod. Outdoor augmented reality on mobile phone using loxel-based visual feature organization. In *Multimedia Information Retrieval*, 2008.

20. P. Viola and M. Jones. Robust real-time object detection. Technical report, Compaq CRL, 2002.

21. N. Wanas and M. Kamel. Decision fusion in neural network ensembles. *Proc. Joint Conference on Neural Networks*, 4:2952–2957, 2001.

22. L. Wiskott, J.-M. Fellous, N. Krüger, and C. von der Malsburg. Face recognition by elastic bunch graph matching. *Journal of Pattern Analysis and Machine Intelligence*, 19(7):775–779, 1997.